Descriptive and Inferential Statistics

Descriptive Statistics deals with main features of existing onhand dataset without making inference / prediction about the large population

- Population vs Sample
  - Population – Complete set of observations that we want to make predictions about
  - Sample: Subset of Pop
- Parameters Vs Statistics
  - A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. Parameters are generally unknown and are estimated using statistics. The goal of statistical inference is to use the information obtained from the sample to make inferences about the population parameters
- Sampling criteria (Ramdom Sampling)
  - Sample Size
  - Randomness
  - Representativeness
- Types of Data
  - Categorical / Qualitative data
    - Nominal
    - Ordinal
  - Numerical / Quantitative Data
    - Discrete Data
    - Continuous Data
- Measure of central tendency
  - Mean
    - Prone to outliers
  - Median (First sort the data)
  - Mode
  - Trimmed mean
    - Removing certain % of smallest and largest value and take mean of remaining value
  - Weighted mean (Ex)
    - (Sum of product of each value and weight) / (Sum of weights)
- Measure of dispersion (Spread in data)
  - Range - (Largest value – Smallest Value)
  - Variance - Average of the squared differences between each data point and the mean
  - SD (Ex) (Why can't we use SD to compare the variablity of two columns – Different units of measure + we square those measure while calculating SD)
  - Coefficent of variation (Application)
    - (SD / Mean) * 100%

- Compares the variability between two or more datasets / Columns, even if the datasets have different units or means
- Different type of Plots
  - Categorical Data vs Categorical Data
    - Build Contingency Table / CrossTab
  - Categorical vs Numerical
    - Bar Chart (Example)
  - Numeric vs Numeric
    - Scatter Plot

- 5 number theory
  - Box plot
    - First sorting of data
    - IQR (Q3 – Q1)
    - Min, Max
    - 25th %, 50th %, 75th %
    - Median – 50 percentiles
    - Minimum = Q1 – 1.5 IQR
    - Maximum = Q3 + 1.5 IQR
    - Q1, Q2, Minimum and Maximum are not the actual values in our data. We calculate these values
  - Percentile
    - PL = P/100 (N + 1)
    - Ex -> X Value at 75th percentile represent that 75% of values fall below X
  - Formulas to get the location
- Covarience
  - Only gives the direction if the two variables are positively or negatively correlated
- Correlation
  - Gives Direction as well as Strength (Magnitude)
    - Correlation between Two variables = [Cov(x,y) / SD(x) * SD(y)]
- Causation
  - Ex – More Ice-crème -> More Murder
    - Actual – More Ice crème -> More people traveling outside of their houses -> More Murder
- Probability Distribution
  - Random Variable
    - Set of Possible value from random experiment
  - Type of Random variable
    - Continous Varibale and Random Variable
  - Type of probability distribution (Discrete and Continous)
    - Probability distribuiton of **Discrete Random Variable** is Called Probability Mass Function (PMF)

      - The PMF of a discrete random variable assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two conditions:
        - The probability assigned to each value must be non-negative (i.e., greater than or equal to zero).
        - The sum of the probabilities assigned to all possible values must equal 1.
    - Probability distribuiton of **Continous Random Variable** is Called Probability Density Function (PDF)
      - Y axis is Density and area under the graph is probability
    - Both PMF and PDF has its own CDF
  - Why do we use PDF ?
    - Calculating Prob of a single value from a large value is not possible
  - What is on Y axis of PDF ?
    - Y axis is Density
  - What does area under the graph represent
    - Probability
  - PDF Estimation (Two Method)
    - Parametrics
      - Initially assuming certain type of distribution of our data using which the PDF is calculated
    - Non-parametric
      - How is KDE constructed ?
      - Computationally intensive and may require more data to achieve accurate estimates
- Normal Distribution (Gausian distribution)
  - X ~ N(Mean, SD)
  - What is standard normal variate (mean = 0 and SD = 1)
  - Emperical rule (68-95-99.7)
  - Basically when you have all the pop parameter, you calculate z score and find prob
  - Properties of normal distribution
    - (mean, meadian, mode) are equal
  - Skewness (Right and left skewed)
  - CDF of normal distribution – integration
- Non- Gausian distribution
  - Continous Non-gausian distribution
    - Uniform (continous)
    - Comman Types of Distribution
      - **Log normal (Heavy tailed righ skewed)**
        - In probability theory and statistics, a lognormal distribution is a heavy tailed continuous probability distribution of a random variable whose logarithm is normally distributed
        - All the right skewed dist cannot be called log normal

            - Pareto (80 % area lies in interval of 20% of x)
    - Discrete Non-gausian distribution
        - Uniform (discrete) distribution
        - Bernollies distribution
        - Binomial distribution
- Kurtosis Concept
    - Excess kurtosis of Normal Distribution is 0 (Mesokurtic)
    - Leptokurtic (more excess kurtosis then 0) – More flater curve (More outliers)
    - Platykutrtic (less excess kurtosis in 0) – Thinner tails (Less Outliers)
- How to determine if the plot are normal or not?
    - By plotting Histogram
    - Q-Q plot
    - Test the Skweeness
- Q-Q plot not only identifies if distribution is normal, but it identifies distribution of two curvers
- Transformation of non gausian to gausian distribution Method
    - Sk learn –
        - log transform (Only applied on right skweed data)
        - Resiprocal transformation
        - Square transform
        - Sq Root transform
    - Power transform
        - Box cox
        - Yeo Johnson
    - Quantile transformation
- For left skwed data -> sq transform -> Normal Distribution
- Log normal distribution -> log transform -> Normal distribution
- General distribution apart from normal-> Box Cox method -> Normal distribution

- ## Confidence Interval
  - Point estimate – It is not reliable i.e for example we cannot claim that my population parameter is A just by determining sample mean as A. Therefore you create a range around point estimate called confidence interval
  - **CI is calculated for parameter and not statistics. Statistics help us get confidence Interval for a parameter**
  - CI = point estimate +- margin of error
  - Two ways to calculate CI
    - Z procedure and T procedure
    - Z procedure is used when you have **SD of Population**
    - To infer something about the population based upon the sample statisitcs, we take the help of z and t Procedure
  - Z procedure
    - Assumptions
      - Sampling should be normal - The data must be collected using a random sampling method to ensure that the sample is representative of the population.
      - Underlying population distribution is normal (Exception Central Limit Therom (CLT))
      - Known SD of population
    - 2nd assumption will kick in when n < 30 i.e populations distribution needs to be normal if sample size is less than 30
    - What does a simple confidence interval [18,25] implies with 95% confidence interval – if we have 1 million people and if we draw sample of 100 people 100 time, out of 100 times, 95 times, the avg age of population will be between 18 and 25
    - As confidence interval expands, range of estimate will also increase
  - T procedure
    - When we don't know anything about population mean and SD, we use t procedure
    - Here Distribution which we got from CLT which was normal distribution, it will be no longer a normal distribution. It will now convert into t distribution.
    - t distribution almost behaves like normal distribution
    - As sample size increase, t distrbution will move towards normal distribution
    - In t distribution, tails are flatter compared to normal distribution

- **Hypothesis testing**
  - Hypothesis testing allows us to make probabilistic statements about population parameters.
  - Null Hypo (No change) and Alternate Hypo (Some Change)
  - Failing to reject null hypo does not mean null hypo is true. It just means that there is'nt enough evidence to support alternate hypo
  - Hypo can be conducted using two approches (**Rejection region approch** and **P value approach**)
  - **Steps in performing hypo testing (Rejection region approach)**
    - Formulate the hypo (Null and Alternate)
    - Select significance level
    - Check the assumptions
    - Decide which test is appropriate
    - State the relevant test statistics
    - Conduct the test
    - Reject or not reject the null hypo
    - Interpret the output
  - Significance **value** can be calculate using significance level and z table
  - What are problems with rejection region approach
    - In case of Rejection region, we need to define Significance level and calculate its Z **value VS** we get single p value for decision making
    - Rejection region approach dosent give you the strength **VS** you get a P value which showcases the strength of the test in magnitude terms
  - Type 1 error: rejecting null hypo when it is true
  - Type 2 error: Not reject alternate when it is false
  - There is trade of between null and alternate hypo, therefore we select 0.05 as significance level
  - One side test Vs Two side test
    - Mean > value
    - Mean < Value
    - Mean =! Value
  - **Hypo testing (P-Value approch)**
    - P-value – It is prob of getting a sample as or more extreme than our own sample given null hypo is true
    - Here instead of seeing where the Z stats is falling i.e beyond significance value or before significance value and then deciding to reject or not to reject null hypo, we calculate the p value based on calculate z statistics and compare it with significance level to decide either to reject or not to reject the null hypothesis

- T test
    - Types of t test
        - One sample t test
        - Independent two sample t test
        - Paired t test (dependent two sample t test)
    - One sample t test is almost like z test. (Difference – Z distribution and t distribution)
    - In two sample t test, we are working with two population and we draw two sample out of these two population
    - In paired t test, value pair should be same in both the sample. It is more of before and after senario test therefore object should remain same
    - **One sample t test**
    - Below are the assumptions
        - Normality
            - If number of data point in sample is < 30, then check if that sample is following normal distribution
            - To check normality, you can use qq plot or shepiro test or any other method used to check normality
            - Specifically for shepiro, if pvalue of shepiro test is less than 0.05 then sample is not following normal distribution
            - If number of data point is sample is >= 30, then even if the pop is not following normal distribution, CLT will come into play and this condition gets satisfied
        - Independence – one data point should be different then other (one packet of lays should be different then other)
        - Random sampling
        - Pop SD is not known
    - **Independent 2 sample t test**
    - Below are the assumptions
        - Independence of observation
        - Random sampling
        - Normality (Shepiro)
        - Pop SD is not known
        - Equal variance of two sample (Look for explantion in pdf)
            - Use either F test or laviance test
            - If p value statitics in laviance test is < 0.05 the variance is not equal

  - **Paired 2 sample t test (Before-After type test)**
    - Statistical test used to compare the means of two related or dependent groups.
    - Below are Assumptions
      - Pop SD is not known
      - Paired Observation - In paired t test, value pair should be same in both the sample. It is more of before and after senario test therefore object should remain same
      - Normality – Difference between the value pair should be appox normal
      - Independence of pair : One pair should be independent of other pair
- **Chi Square test (Non parametric test)**
  - Two types of test
    - Chi square goodness of fit test
    - Chi square test for independence
  - In both the test it is assumed that the Statistics calculated will follow chi square distribution
  - Chi sqaure distribution is derived from normal distribution only
  - With increase in DOF, the curve move toward normal distribution
  - Z test and t test was performed on continous varibale where as chi square test is performed on categorical column
  - **Chi square goodness of fit test**
    - This test is particularly useful when you want to assess if the sample data is consistent with an assumed distribution or if there are significant deviations from the expected pattern.
    - This test only needs one categorical variable
    - **Assumption :**
      - Observation should be independent from each other
      - It should be categorical data and not continuous or ordinal data
      - Each category should have expected frequency of at least 5
      - Theorotical distribution should be specified before the test is conducted.
    - **Procedure :**
    - Define Null and alternate hypo
      - H0 : observed data follow the expected theorotical distribution

        - H1 : observed data does not follow the expected theorotical distribution
    - We take the observed values of sample
    - Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
    - Calculate the chi square test statitics by comparing the observed and expected frequencies
    - Chi square statistics – summation of ((Expected - observed)^2)/Expected)
    - Calculate degree of freedom
    - Plot chi square distribution based on particular degree of freedom. [df = n -1]
    - Plot your statistic on that graph and get corresponding p value and compare that p value with significance level
    - Decide if you want to reject or not reject the null hypothesis
- **Chi square test for independence**
    - This test needs two categorical varibale and it checks if there's any significant relationship between those two categories
    - This test helps in feature selection as well
    - Assumptions :
        - Independence of observation – Random sampling
        - Category can be either continous or ordinal
        - Thumb rule is that frequency in each cell should be at least 5
        - The marginal totals (the row and column sums of the contingency table) should be fixed before the data is collected.
        -
    - Test Procedure :
    - Define null and alternate hypo
    - Create contingency table with observed data
    - Calculate contingency table of expected frequency for each cell
    - Compute the chi square statistics (same as before)
    - Calculate degree of freedom df = (number of row – 1) * (number of column -1)
    - Plot chi sqaure distribution for the given df
    - Plot your chi square statitics over the distribution and get the corresponsing p value and compare that p value with the significance level
- **ANOVA Test of Variance**

- Used to test Hypothesis about the equality of two variance in different samples or populations
- It needs two parameter : The degrees of freedom for the numerator (df1) and the degrees of freedom for the denominator (df2).
- Two test – One way ANOVA and Two way ANOVA
- **One Way ANOVA :**
    - One-way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if there are any significant differences between them. It is an extension of the t-test, which is used for comparing the means of two independent groups. The term "one-way" refers to the fact that there is only one independent variable (factor) with multiple levels (groups) in this analysis.
    - The primary purpose of one-way ANOVA is to test the null hypothesis that all the group means are equal. The alternative hypothesis is that at least one group mean is significantly different from the others.
    - It's important to note that one-way ANOVA only determines if there is a significant difference between the group means; it does not identify which specific groups have significant differences. To determine which pairs of groups are significantly different, post-hoc tests, such as Tukey's HSD or Bonferroni, are conducted after a significant ANOVA result.

Rejection Region Approch and P - Value Approch

The two main approaches to hypothesis testing are the rejection region approach and the p-value approach. The p-value approach was developed to address some of the limitations and downsides of the rejection region approach. Here are the key downsides of the rejection region approach that the p-value approach helps to mitigate:

1. Lack of Flexibility:

- **Rejection Region Approach** : This approach requires specifying a critical value (or rejection region) before conducting the test. The critical value is based on a pre-determined significance level (α), such as 0.05. If the test statistic falls into the rejection region, the null hypothesis is rejected; otherwise, it is not.

- **P-Value Approach** : Provides more flexibility because it calculates the exact probability of obtaining the observed test statistic under the null hypothesis. Researchers can then compare this p-value to their chosen significance level to make a decision, allowing for easier interpretation and comparison across different studies.

2. Binary Decision Making:

- **Rejection Region Approach**: This approach leads to a binary decision (reject or do not reject the null hypothesis) without providing information about the strength of the evidence against the null hypothesis. It does not indicate how extreme the test statistic is.

- **P-Value Approach**: Offers a continuous measure of evidence against the null hypothesis. A smaller p-value indicates stronger evidence against the null hypothesis, providing more nuanced information about the result.

3. Dependence on Pre-Set Significance Levels:

- **Rejection Region Approach**: Requires the significance level to be chosen beforehand, which can sometimes seem arbitrary (e.g., α = 0.05). This fixed threshold can lead to different conclusions for similar test statistics that fall on either side of the critical value.

- **P-Value Approach**: Allows researchers to see the exact probability and decide on the significance level after seeing the data, which can help in understanding the strength of the evidence and making more informed decisions.

4. Comparison Across Studies:

- **Rejection Region Approach**: Makes it difficult to compare results across studies with different significance levels, as the critical values may differ.

- **P-Value Approach**: Facilitates comparison across studies because p-values are standardized probabilities. Different researchers can use the same p-value to assess the strength of the evidence, regardless of their specific significance levels.

5. Interpretation:

- **Rejection Region Approach:** Can be less intuitive for those who are not statistically trained, as it involves understanding critical values and rejection regions.

- **P-Value Approach** : Provides a single number (the p-value) that is easier for many people to interpret. It directly tells how likely the observed data (or more extreme) would be under the null hypothesis.

Summary:

The p-value approach addresses the rigidity and limitations of the rejection region approach by providing a more flexible, informative, and standardized measure of evidence against the null hypothesis. It allows for more nuanced decision-making and easier comparison across different studies.

Assumptions of Single Sample t test

1. **Normality** - Population from which the sample is drawn is normally distributed (Exception - CLT)

2. **Independence** - The observations in the sample must be independent, which means that the value of one observation should not influence the value of another observation.

3. **Random Sampling** - The sample must be a random and representative subset of the population.

4. **Unknown population std** - The population std is not known.

Assumption of Two independent Sample t test :

Independence of observations: The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in the other group. The subjects in the two groups should be selected randomly and independently.

Normality: The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large (typically n ≥ 30) and the sample sizes of the two groups are **similar**. If the data is highly skewed or has substantial outliers, consider using a non-parametric test, such as the Mann-Whitney U test.

Equal variances (Homoscedasticity): The variances of the two populations should be approximately equal. This assumption can be checked using F-test for equality of variances. If this assumption is not met, you can use Welch's t-test, which does not require equal variances.

Random sampling: The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of the population and reduces the risk of selection bias.

P – Value Strength

1. **Very small p-values** (e.g., $p < 0.01$) indicate strong evidence against the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.

2. **Small p-values** (e.g., $0.01 \leq p < 0.05$) indicate moderate evidence against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.

3. **Large p-values** (e.g., $0.05 \leq p < 0.1$) indicate weak evidence against the null hypothesis, suggesting that the observed effect or difference might have occurred by chance alone, but there is still some level of uncertainty

4. **Very large p-values** (e.g., $p \geq 0.1$) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

Difference between One sample and Two sample t - test

One-Sample t-Test:

- Compares the mean of one sample to a known or hypothesized population mean.
- Used when there is a benchmark or standard value to compare against.
- Suitable for quality control, health studies, and educational research.

Two Sample t-Test:

- Compares the means of two Dependent / independent samples.
- Used to assess differences between two groups or conditions.
- Suitable for clinical trials, educational studies comparing different teaching methods, and market research comparing different regions or groups.

Paired two-sample t-test

Paired observations: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.

Normality: The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large (> 30).

Independence of pairs: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

Chi Square Goodness of Fit

The Chi-Square Goodness-of-Fit test is a **non-parametric test**. Non-parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation. In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The test doesn't rely on any assumptions about the underlying distribution's parameters. Instead, it focuses on comparing observed counts to expected counts, making it a non-parametric test.

Applications of Chi Square test in Machine Learning

Feature selection: Chi-Square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association between each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.

Evaluation of classification models: For multi-class classification problems, the Chi-Square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's predictions align with the actual class distributions.

Analysing relationships between categorical features: In exploratory data analysis, the Chi-Square test for independence can be applied to identify relationships between pairs of categorical features. Understanding these relationships can help inform feature engineering and provide insights into the underlying structure of the data.

Discretization of continuous variables: When converting continuous variables into categorical variables (binning), the Chi-Square test can be used to determine the optimal number of bins or intervals that best represent the relationship between the continuous variable and the target variable.

Variable selection in decision trees: Some decision tree algorithms, such as the CHAID (Chi-squared Automatic Interaction Detection) algorithm, use the Chi-Square test to determine the most significant splitting variables at each node in the tree. This helps construct more effective and interpretable decision trees.