DAV-Fundamentals Cheat Sheet

Hypothesis Testing

- This is a method of statistical inference to decide whether the data at hand sufficiently supports a particular hypothesis.
- A test statistic directs us to either reject or fail to reject the null hypothesis.

Before conducting a hypothesis test, we need to define:
- **Null hypothesis (H_0)** represents the assumption that is made about the data sample
- **Alternative hypothesis (H_a)** represents a counterpoint.

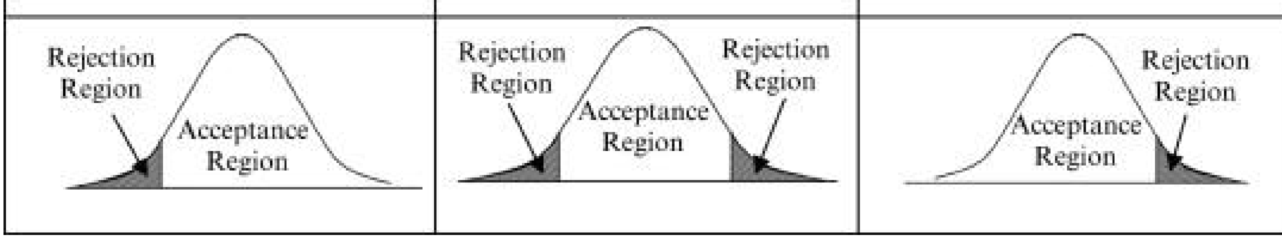
P-value: Probability of observing the Test statistic as extreme or more than $T_{observed}$ considering the null hypothesis as true.
- If P-value < Significance level; reject the null hypothesis,
- Otherwise, fail to reject the null hypothesis.

Significance level:
- The significance level, often denoted as **α**, is the threshold probability for rejecting the null hypothesis in a hypothesis test.
- Commonly set at 0.05, it represents the maximum acceptable probability of making a Type I error (incorrectly rejecting a true null hypothesis).

Confidence level:
- The confidence level is the complement of the significance level (1 - α) and represents the degree of certainty associated with a confidence interval.
- For example, a 95% confidence level implies a 95% probability that the interval contains the true population parameter.

Types of Hypothesis Testing:

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$



Types of Errors:

- **Type I error (α) -** When we Reject a null hypothesis that is actually true.
- **Type II error (β) -** When we fail to reject a null hypothesis that is actually false.

Power of a test: The probability that a statistical test will correctly reject a false null hypothesis (i.e., control Type II error): $Power\ = 1 - \beta$

Factors Affecting Power:

- Effect size: Larger effect sizes lead to higher power.
- Sample size: Larger sample sizes lead to higher power.
- Significance level: Lower significance levels (e.g., $\alpha = 0.01$) lead to lower power.

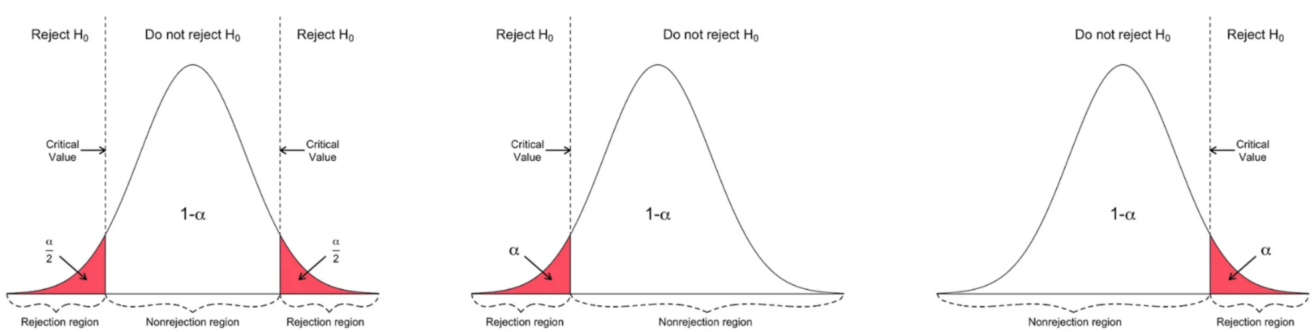
Framework for Hypothesis Testing:

1. Define null and alternate hypotheses.
2. Decide a test statistic and a corresponding distribution.
3. Determine whether the test should be left-tailed, right-tailed, or two-tailed.
4. Determine the p-value.
5. Choose a significance level.
6. Accept or reject the null hypothesis by comparing the obtained p-value with the chosen significance level.

Central Limit Theorem (CLT):

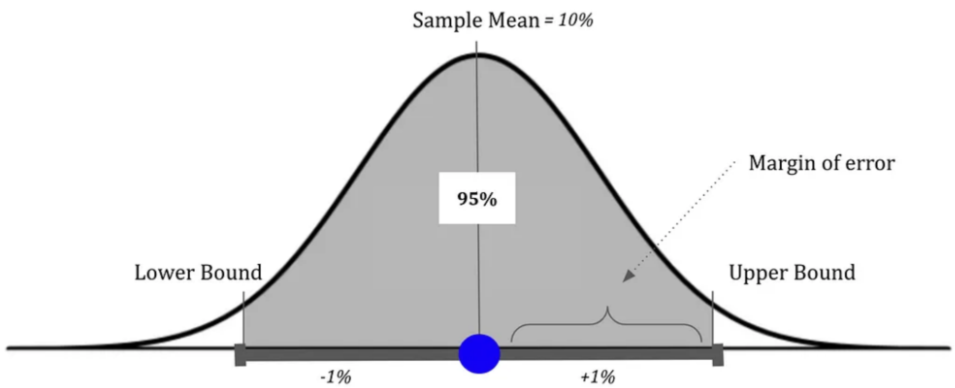
- It states that the sampling distribution of sample means is approximately Gaussian, no matter what the shape of the original distribution is.
- Assumptions:
    - Population standard deviation should be finite
    - The sample size is sufficiently large (typically n >=30.)
    - Data is sampled randomly and independently.

Critical value: A cut-off value used to mark the start of a region where the test statistic is unlikely to fall in.



Confidence Intervals:

- This gives a range of values where you're reasonably sure the true result lies.
- If this interval includes the null hypothesis value, you accept the null hypothesis; otherwise, you reject it.
- $Confidence\ Interval\ =\ Sample\ Mean \pm (Critical\ Value\ *\ Standard\ Error)$



One sample Z-test:

- Used to determine whether the population mean is significantly different from an assumed value.
- It uses Standard normal distribution as the baseline.
- **Assumptions**: Either the population's standard deviation should be known or we should estimate them well when the sample size is not too small (n>30)
- Test statistic = $z\ =\ \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

Two sample Z-test:

- Used to compare the means of two populations.
- **Assumption:** Either the standard deviation (σ_1, σ_2) of the populations should be known or we should estimate them when the sample sizes are not too small ($n_1, n_2\ \geq\ 30$).
- Test statistic = z = $\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

One sample Z-proportion test:

- Used to assess if the proportion of a single sample is significantly different from a given value.
- **Assumptions**:

- The sample is randomly selected and the sample size is large enough (usually when $n * p_0$ and $n * (1 - p_0)$ are both greater than 10).
- The population can be assumed to be normally distributed or the sample size is large enough for the Central Limit Theorem to apply.

- $Test\ Statistic\ =\ z\ =\ \frac{\hat{p}-p_0}{\sqrt{(p_0(1-p_0)/n}}$

Two-sample Z-proportion test:

- Employed to compare the proportions of two independent samples.
- **Assumptions**:
  - The samples are randomly selected and independent of each other and the sample sizes are large enough (usually when $(n_1 * \hat{p}_1)$, $(n_1 * (1 - \hat{p}_1))$, $(n_2 * \hat{p}_2)$, and $(n_2 * (1 - \hat{p}_2))$ are all greater than 10).
  - The populations can be assumed to be normally distributed or the sample sizes are large enough for the Central Limit Theorem to apply.
- $Test\ Statistic\ =\ z\ =\ \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) + (\frac{1}{n_1}+\frac{1}{n_2})}}$

  Note: $\hat{p} = \frac{x_1+x_2}{n_1\ + n_2}$ is the pooled sample proportion.

One sample t-test:

- The test statistic follows a t-distribution
- It is used when the sample size is too small (n < 30) and/or the population standard deviation (σ) is unknown.
- Test statistic = t = $\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
- Degree of freedom = n -1

Two sample t-test:

- It is used when the sample sizes are too small ($n_1, n_2\ < 30$) and/or the population standard deviations (σ_1, σ_2) are unknown.
- Test statistic = t = $\frac{(\bar{x}_1 - \bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
- Degree of freedom = n1 + n2 - 2

Paired t-test:

- Used to compare the means of two related groups (e.g., before and after treatment).
- **Assumptions**:
    - The differences between the paired samples are normally distributed.
    - The differences are independent of each other.

Chi-square goodness of fit test:

- Used to determine if the distribution of categorical data fits a theoretical distribution (expected behavior).
- Formula: $ChiSquare\ Statistic = \sum_i \frac{(O_i - E_i)^2}{E_i}$
- **Assumptions:**
    - Categorical data (data that can be divided into categories).
    - Random sample & Independent observations.
    - Expected frequencies in each category ≥ 5.

Chi-square test of independence:

- Used to assess whether there is a significant association between two categorical variables.
- The assumptions for this test are similar to the goodness of fit test.

One way-ANOVA (Analysis of variance):

- Used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.
- Test Statistic Formula: $F\ statistic = \frac{MSB}{MSW}$, where:
    - MSB is the mean square between groups (measures variability between group means)
    - MSW is the mean square within groups (measures variability within each group)
- **Assumptions:**
    - Normality: The data within each group is normally distributed.
        - To check normality we perform the Shapiro-Wilk test
    - Homogeneity of variances: The variances of the groups are equal.
        - To check the homogeneity of variances we perform Levene's test
    - Independence: The observations within each group are independent of each other.

Kruskal-Wallis test:

- A non-parametric test is used to determine if there are statistically significant differences between two or more independent groups.

- If One-way ANOVA’s assumption of normality fails, we can perform the Kruskal-Wallis test.
- Instead of using sample means to compare the groups, it uses sample medians

Two-way ANOVA:

- Used to analyze the influence of two categorical independent variables on a dependent variable.
- **Assumptions:**
    - The populations from which the samples are drawn should be approximately normally distributed.
    - Homogeneity of variances within each combination of the two independent variables.
    - Independence of observations.

KS (Kolmogorov - Smirnov) test:

- It is a non - parametric test used for determining whether the distributions of two samples are the same or not.
- The test statistic T_{ks} follows a distribution called the Kolmogorov Distribution.

T_{KS} = the maximum absolute value of the difference in the CDFs of the two samples X and Y.

- **Assumptions:**
    - The data is continuous.
    - The data is independent and identically distributed

A/B Testing:

- Used to compare the performance of two versions (A and B) of something. Various statistical tests can be used, depending on the type of data and goals of the test
- **Assumptions:**
    - Randomization: Users are randomly assigned to each version.
    - Independence: The behavior of users in one group doesn't affect those in the other group.
    - Sample size: Each version has a sufficient number of users to detect meaningful differences.

Covariance:

- Measures how two variables change together.
- It doesn't indicate the strength of the relationship or its direction, only the tendency to move together or in opposite directions.

Correlation:

- Measures the strength and direction of a linear relationship between two continuous variables.
- Values range from -1 (negative correlation) to +1 (positive correlation), with 0 indicating no correlation.

Pearson correlation coefficient(PCC):

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}$$

The limitation of PCC is that it only captures the linear relationship between the variables. It fails to capture the non-linear patterns.

Spearman Rank Correlation Coefficient:

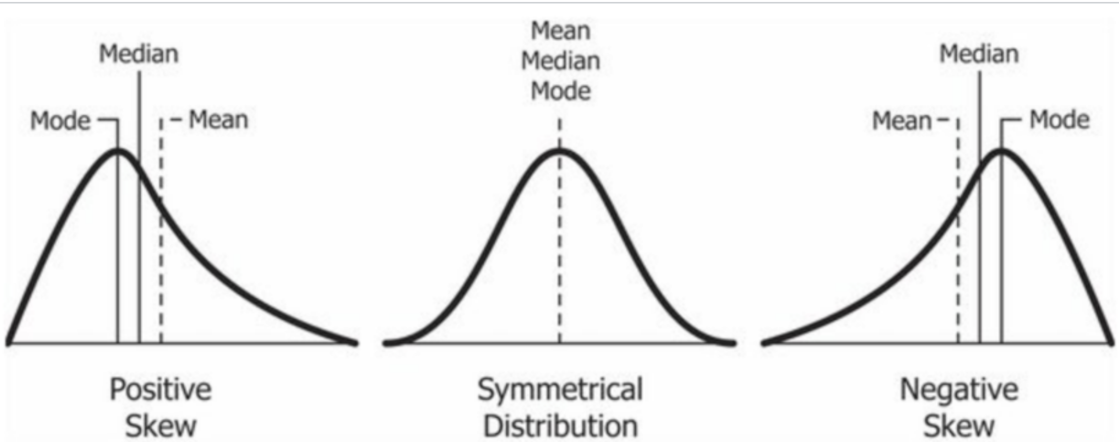
- It is a statistical measure of the strength of a monotonic relationship between paired data.
- It captures the monotonicity of the variables rather than the linearity.

Feature Engineering:

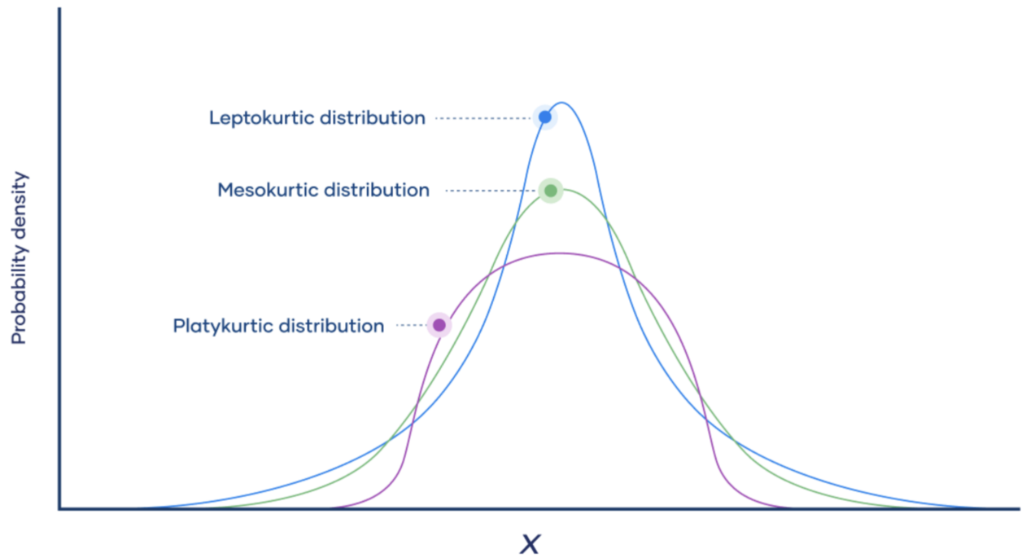
- Interaction terms: Create new features by combining existing ones to uncover hidden relationships.
- Domain-specific transformations: Apply transformations based on domain knowledge.

Binning: Convert numerical values into categorical bins for simplicity or to capture non-linear relationships.

Skewness: Measures the asymmetry of a probability distribution.



Kurtosis: Measures the sharpness of the peak and the tails of a probability distribution.



Handling Missing Values:

- Identify missing values: Use functions like `isnull()` or `info()` to identify missing values.
- Imputation:
  - One advanced technique is the Simple Imputer replaces missing values with the mean or median or mode of the column.

Outlier Treatment:

- Identifies and removes extreme values based on predefined thresholds using statistical methods like Z-scores, and IQR (Interquartile Range)

Encoding: Convert categorical variables to numerical representations using encoding techniques

- One-hot encoding: transforms categorical data into binary values, creating new columns for each category and indicating the presence (1) or absence (0).
- Label encoding: Assigns a unique number to each category
- Target encoding: Replacing categorical values with the mean of the target variable within each category.

Scaling:

- Min-Max scaling: Scale values between 0 and 1.
- Standardization (Z-score normalization): Transform data to have a mean of 0 and a standard deviation of 1. Scale values between -3 and +3.