Hypothesis testing Cheatsheet

• **Central Limit Theorem (CLT):**
the distribution of sample means is Gaussian, no matter what the shape of the original distribution is.
Assumptions: population mean and standard deviation should be finite and sample size >=30.

• **Hypothesis Testing:** a method of statistical inference to decide whether the data at hand sufficiently support a particular hypothesis. A test statistic directs us to either reject or not reject the null hypothesis.

• **Null hypothesis** (H_0) represents the assumption that is made about the data sample whereas the **alternative hypothesis** (Ha) represents a counterpoint.

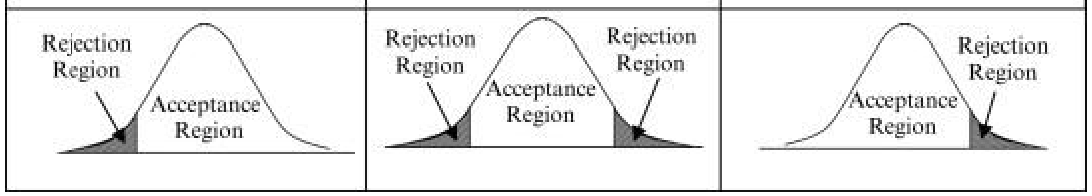
• **p-value**: Probability of observing the Test statistic as extreme or more than $T_{observed}$ considering the null hypothesis as true.

If p-value < significance level; reject the null hypothesis, else fail to reject the null hypothesis.

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

• **Types of Hypothesis testing:**

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



Type I error (α) - Reject a null hypothesis that is true.
Type II error (β) - Not reject a null hypothesis that is false.

Framework for Hypothesis testing:

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

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 standard deviation of the population should be known or we should estimate them well when the sample size is not too small (n>30).

$$\text{Test statistic} = Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

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 ($n_1, n_2 \geq 30$).

$$\text{Test statistic} = t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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.

$$\text{Test statistic} = z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

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.

$$\text{Test statistic} = t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Degree of freedom = The smaller of $(n_1 - 1)$ and $(n_2 - 1)$

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.

The test statistic f follows the F distribution represented by two parameters (k-1) and (n-k).
k = No. of groups, n = Total sample size.

$$\text{Test statistic} = f = \frac{MSB}{MSW}$$

where, MSB = mean of the squared distances between the groups
and MSW = the mean of the squared distances within the groups.

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1} \qquad MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_ij - \bar{X}_i)^2}{n-k}$$

Assumptions of ANOVA:

- The variance of each group should be the same or close to each other.
- The total n observations should be independent of each other.

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**.

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

Correlation is the degree of the mutual relationship between two variables.

Pearson correlation coefficient(PCC):

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

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.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where, d =difference between the two ranks of each observation
and n = number of observations