# **Boxplot**

A box plot is a standardized graphical representation that summarizes a dataset's distribution using its key statistics: the minimum value (lower whisker), the first quartile (Q1), the median, the third quartile (Q3), and the maximum value (higher whisker). It provides insights into various aspects of the data, such as the presence and values of outliers, the symmetry of the data, the degree of data clustering, and any skewness in the distribution.

The five-point summary comprises five crucial statistical measures employed to encapsulate the characteristics of a dataset's distribution. These statistics offer a succinct overview of both the data's typical value and its variability. The five elements in this summary include:

1. Minimum (Q1 - 1.5 \* IQR): This represents the dataset's smallest value, signifying the lower end of the data range.

2. First Quartile (Q1): Also referred to as the 25th percentile, this is the value below which 25% of the data points lie. It demarcates the lower boundary of the first quarter of the dataset when sorted in ascending order.

3. Median (Q2): The median corresponds to the middle value in the sorted dataset. It signifies the 50th percentile, indicating the point below which 50% of the data values are situated.

4. Third Quartile (Q3): Commonly known as the 75th percentile, Q3 delineates the value beneath which 75% of the dataset falls. It establishes the lower boundary of the third quarter of the sorted dataset.

5. Maximum (Q3 + 1.5 \* IQR): This represents the dataset's largest value, signifying the upper limit of the data range.



# **Kolmogorov Smirnov (K-S Test)**

The Kolmogorov-Smirnov test (Chakravart, Laha, and Roy, 1967) is used to decide if a sample comes from a population with a specific distribution.

The Kolmogorov Smirnov test is defined by -:

-: The data follows a specified distribution.

-: The data does not follow a specified distribution.

Test Statistic is defined as –

Where *F* is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution in our case it is the distribution being tested is *Normal Distribution.*

Critical Values: The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table.

*Characteristics -:* One notable advantage is that the distribution of the K-S test statistic remains independent of the specific cumulative distribution function under examination. Additionally, it offers the benefit of being an exact test, unlike the chi-square goodness-of-fit test, which requires a sufficiently large sample size for its approximations to hold.

*Limitations* -: It is exclusively applicable to continuous distributions and tends to exhibit higher sensitivity in the central part of the distribution, as opposed to the tails. Perhaps the most significant drawback is that the distribution must be fully defined. In other words, if parameters related to location, scale, or shape are estimated from the data, the critical region for the K-S test loses its validity. Typically, it must be determined through simulation.

# Mann Whitney U test

Mann-Whitney U test is the non-parametric alternative test to the independent sample t-test.  It is a non-parametric test that is used to compare two sample means of rank sums that come from the same population, and used to test whether two sample means of rank sums are equal or not.  Usually, the Mann-Whitney U test is used when the data is ordinal or when the assumptions of the t-test are not met.

**Calculation of the Mann-Whitney U Test:**

Where:  
U=Mann-Whitney U test  
= sample size one  
= Sample size two  
 = Rank of the sample size

**The Hypothesis of Mann-Whitney U-Test is defined as,**

-: There is no significant difference among the medians of the two groups in the population.

-: There is a significant difference among the medians of the two groups in the population.

**Decision** -: If the U-value is greater than the tabulated value, reject the null hypothesis. This indicates that there is a significant difference among the group medians. If the U-value is less than tabulated value, fail to reject the null hypothesis, suggesting no significant difference among the group medians.

**WHAT IF THERE IS A TIE WITHIN THE RANKS:**

**Assign Average Ranks**: When you encounter tied ranks, calculate the average rank for the tied values. For example, if three data points have the same value, assign them an average rank of (r1 + r2 + r3) / 3, where r1, r2, and r3 are their individual ranks.

**Calculate the U Statistic**: Calculate the U statistic for each group, using the average ranks instead of individual ranks. The U statistic is the sum of ranks for one group. It is used to determine whether one group's values tend to be higher or lower than the other group. Use the Smaller U as the Test Statistic. Compare the U statistics for both groups, and use the smaller one as the test statistic. This smaller U value is then used to look up the critical value from the Mann-Whitney U table or calculate the p-value.

**Adjust for Ties in Sample Size**: If there are significant ties in the data, you may need to adjust the U statistic for tied ranks. This adjustment is usually made for small to moderate sample sizes and involves using a correction factor.

# Kruskal Wallis H-Test

The Kruskal-Wallis H test, often referred to simply as the Kruskal-Walli’s test, is a non-parametric statistical test used to determine whether there are statistically significant differences among the medians of three or more independent groups. It is an extension of the Mann-Whitney U test, which compares two groups, to scenarios with multiple groups.

**Calculation of Kruskal Wallis H Test:**

n = sum of sample sizes for all samples,

c = number of samples,

 = sum of ranks in the jth sample,

 = size of the jth sample.

**The hypotheses for the Kruskal-Walli’s test are as follows:**

**Null Hypothesis ():** There are no significant differences among the medians of the groups being compared. In other words, all the groups come from the same population, and any observed differences are due to random chance.

**Alternative Hypothesis ():** There are significant differences among the group medians. In practical terms, this means that at least one group differs from the others in terms of the central tendency of the data.

**Decision**-: If the H-value is greater than the tabulated value, reject the null hypothesis. This indicates that there are significant differences among the group medians. If the H-value is less than tabulated value, fail to reject the null hypothesis, suggesting no significant differences among the group medians.

**WHAT IF THERE IS A TIE WITHIN THE RANKS:**

**Assign Average Ranks for Ties**: When you have tied data points within or across groups, calculate the average rank for each set of tied values. For example, if multiple data points share the same value, assign them the average rank of (r1 + r2 + ... + rn) / n, where r1, r2, ..., rn are the individual ranks for the tied values, and 'n' is the number of tied data points.

**Calculate the Adjusted Kruskal-Wallis H Statisti**c: Use the average ranks to calculate the adjusted Kruskal-Wallis H statistic. This statistic considers the ranked data with tied ranks.

**Determine Degrees of Freedom**: Determine the degrees of freedom for the Kruskal-Wallis test, which is calculated based on the number of groups (k) and the number of tied ranks. The formula for degrees of freedom in this case is usually (k - 1) since tied ranks reduce the effective number of independent values.

**Critical Value or P-Value:** Compare the calculated H statistic to a chi-squared distribution with (k - 1) degrees of freedom or calculate the associated p-value to determine whether the result is statistically significant.

**Make a Decision:** If the p-value is less than the chosen significance level (alpha), reject the null hypothesis. This indicates that there are significant differences among the group medians. If the p-value is greater than alpha, fail to reject the null hypothesis, suggesting no significant differences among the group medians

# Dunns Test

**WHY DUNNS TEST:**

If the results of a Kruskal-Walli’s test are statistically significant, then it’s appropriate to conduct **Dunn’s Test** to determine exactly which groups are different. Dunn’s Test performs **pairwise comparisons** between each independent group and tells you which groups are statistically significantly different at some level of α.

**The hypotheses for Dunn's test can be framed as follows:**

**Null Hypothesis (H0):** There is no significant difference in the population medians between those two groups. In other words, the medians of the compared groups are equal.

**Alternative Hypothesis (Ha):** There is a significant difference in the population medians between the two groups being compared. This means that the medians of the compared groups are not equal.

**To perform Dunn's test, follow these steps:**

**Conduct the Kruskal-Wallis H Test:** Start by conducting the Kruskal-Walli’s test to determine whether there are significant differences among the groups. If the Kruskal-Walli’s test is significant (rejecting the null hypothesis), proceed to Dunn's test.

**Calculate Ranks for All Data Points**: Rank all data points across all groups. If there are tied ranks, calculate average ranks as explained earlier.

**Calculate the H Value:** For each pair of groups, calculate the H value using the formula:

**Where:**

is the H value for the pair of groups.

is the sum of the ranks in the first group.

is the sum of the ranks in the second group.

is the number of observations in the smaller of the two groups being compared.

**Adjust for Multiple Comparisons**: Adjust the significance level (alpha) for multiple comparisons. This is often done using methods like the Bonferroni correction, Holm-Bonferroni method, or false discovery rate (FDR) correction.

**Compare H Values to Critical Values**: Compare the H values calculated for each pair of groups to the critical value obtained after adjusting for multiple comparisons. If H > critical value, you can conclude that there is a significant difference between the groups being compared.

**Repeat for Each Pair of Groups:** Repeat steps 3 to 5 for all possible pairs of groups. This will help you identify which specific pairs of groups have significant differences.