Dunn’s Test for Multiple Comparisons

Authored by

Mohammed looti

November 7, 2025

Understanding Non-Parametric Hypothesis Testing

The Kruskal-Wallis test is a fundamental tool in non-parametric statistics. It is utilized when researchers need to assess whether there are statistically significant differences among the medians of three or more independent groups. This test serves as the non-parametric equivalent of the standard One-Way ANOVA, which typically requires strict assumptions about data distribution that the Kruskal-Wallis test does not.

If the initial Kruskal-Wallis analysis yields a statistically significant result (i.e., the null hypothesis is rejected), we only know that at least one group median differs from the others. Crucially, it does not specify **which** pairs of groups are the source of this detected difference. To rigorously identify the specific groups causing the variation, a dedicated post-hoc procedure is required.

The Purpose and Application of Dunn's Test

When the Kruskal-Wallis test indicates an overall effect, it becomes necessary and appropriate to conduct **Dunn's Test**. This statistical procedure is specifically designed for multiple comparisons following a non-parametric omnibus test. It systematically performs pairwise comparisons between every combination of independent groups, thereby identifying exactly which pairs are statistically significantly different at a predefined level of alpha (α).

Consider a practical example in experimental design: Suppose a researcher is investigating whether three different experimental drugs (Drug A, Drug B, and Drug C) have varying effects on chronic back pain relief. The study involves recruiting 30 subjects, randomly assigning them to one of the three drug regimens for one month, and then measuring their final back pain score.

The researcher would first perform the Kruskal-Wallis test to determine if the median back pain scores are equal across the three drug groups. If the resulting p-value is below the predetermined significance threshold, the conclusion is that the three drugs produce statistically different overall effects. Following this initial finding, the researcher would then execute **Dunn's Test** to determine precisely *which* pairs of drugs (A vs. B, A vs. C, or B vs. C) generate statistically significant differences in pain reduction.

Dunn's Test: The Formula

While statistical software packages (such as R, Python, Stata, or SPSS) are routinely used to execute Dunn's Test, understanding the mathematical basis provides essential clarity. The test relies on calculating a z-test statistic for the comparison of average ranks between any two specified groups.

The general formula used to calculate the z-test statistic (zi) for the *i*th comparison is defined as:

$$z_i = y_i / \sigma_i$$

In this formula, *i* represents the specific pairwise comparison being performed among the total *m* comparisons. The term yi represents the difference between the average of the sum of the ranks for the two groups being compared (yi = WA - WB, where WA is the average of the sum of the ranks for the ith group). The term σi represents the estimated standard deviation of the difference in average ranks, which is calculated incorporating adjustments for potentially tied ranks.

The calculation of σi is complex, especially when accounting for data points that share the same rank (tied ranks). The standard deviation calculation is defined as follows:

$$\sigma_i = \sqrt{((N(N+1)/12) - (\Sigma T3s - Ts/(12(N-1)) / ((1/nA)+(1/nB))}$$

Within this variance formula, *N* denotes the total number of observations pooled across all groups, *r* is the number of tied ranks present in the data set, and Ts signifies the number of observations tied at the *s*th specific tied value. This incorporation of tied ranks ensures the test retains its statistical validity under real-world data conditions.

Controlling the Family-wise Error Rate

A fundamental challenge inherent in conducting multiple comparisons is the inflation of the experiment-wise or family-wise error rate (FWER). If we conduct a single test at a significance level (α) of 0.05, there is a 5% chance of committing a Type I error (a false positive). However, when multiple tests are performed simultaneously, the overall probability of incorrectly rejecting the null hypothesis in at least one comparison increases dramatically, potentially undermining the integrity of the research findings.

To maintain statistical rigor and ensure that the overall probability of a Type I error remains at or below the predefined alpha level, it is absolutely essential to apply a rigorous correction method. These methods adjust the p-values resulting from the multiple comparisons to account for the increased risk, thereby ensuring the set of findings is statistically sound.

Popular P-Value Adjustment Methods

Several established statistical methodologies exist to correct for the inflation of the FWER in post-hoc tests. These methods modify the p-value threshold, making it more conservative for each individual comparison. The two methods most frequently used in conjunction with Dunn's Test are the Bonferroni and the Sidak adjustments.

1. The Bonferroni Adjustment

The Bonferroni correction is one of the most widely known and simplest adjustment techniques. It

achieves FWER control by dividing the nominal alpha level by the total number of comparisons (m), or equivalently, by multiplying the observed p-value by the number of comparisons. This is a highly conservative approach.

Adjusted p-value = p*m

p: The original, unadjusted p-value derived from the specific pairwise comparison.
m: The total number of comparisons being made within the statistical family.

2. The Sidak Adjustment

The Sidak (or Dunn-Sidak) adjustment provides an alternative to the Bonferroni method. It is generally less conservative, which can lead to greater statistical power, particularly when the number of comparisons is large. The Sidak method assumes that the individual tests are independent of one another.

Adjusted p-value = 1 - (1-p)m

where:

p: The original p-value calculated for the specific pairwise comparison.
m: The total number of comparisons being evaluated simultaneously.

By successfully applying either of these p-value adjustments, researchers can dramatically reduce the overall probability of committing a Type I error across the entire set of hypothesis tests, thus ensuring the reliability and validity of the final conclusions drawn from the data.

Additional Resources

For those seeking to implement Dunn's Test using modern statistical tools, the following guides offer detailed instructions for practical application:

Detailed Guide on How to Perform Dunn's Test in R
Step-by-Step Instructions for Dunn's Test in Python