Multiplicity Adjustments: Understanding the Nuance of Post hoc Tests

Paul Schmidt

2023-12-13

What are multiplicity adjustments? What are the various methods for multiple mean comparisons, including the t-test, Tukey test, and others, highlighting their purposes and differences.

Table of contents

1 The motivating issue

In statistical analysis, conducting multiple mean comparisons is a common practice, especially following ANOVA tests. This practice, however, introduces a complexity: the risk of committing Type I errors (false positives) increases with each additional comparison.

ℹ Brief Intro: Type I & Type II errors

In statistical analysis, understanding the concepts of Type I and Type II errors is crucial. A Type I error, denoted as α, occurs when a true null hypothesis is incorrectly rejected (= false positive), such as falsely identifying a difference where none exists. Conversely, a Type II error, denoted as β, happens when a false null hypothesis is not rejected, essentially missing a real difference. Controlling these error rates is vital in any statistical test, but it's important to note that it's impossible to completely eliminate them. When a test is said to be "performed at the 5% level," it means that the Type I error rate (α)

> is controlled at 5%, accepting a 5% risk of a false positive.

Taken directly from Wikipedia's article "Multiple comparisons problem":

> For example, if one test is performed at the 5% level and the corresponding null hypothesis is true, there is only a 5% risk of incorrectly rejecting the null hypothesis. However, if 100 tests are each conducted at the 5% level and all corresponding null hypotheses are true, the expected number of incorrect rejections (also known as false positives or Type I errors) is 5. If the tests are statistically independent from each other (i.e. are performed on independent samples), the probability of at least one incorrect rejection is approximately 99.4%.
>
> The multiple comparisons problem also applies to confidence intervals. A single confidence interval with a 95% coverage probability level will contain the true value of the parameter in 95% of samples. However, if one considers 100 confidence intervals simultaneously, each with 95% coverage probability, the expected number of non-covering intervals is 5. If the intervals are statistically independent from each other, the probability that at least one interval does not contain the population parameter is 99.4%.

As a consequence, several methods have been developed to deal with this issue, each with its strengths and weaknesses.

2 The Landscape of Post hoc Tests

Here are a few of the most common post hoc tests:

- **Multiple t-test** (Fisher, 1940): A pairwise comparison test that looks at each pair of means separately.
- **Tukey's HSD Test** (John Tukey, 1949): Designed to compare all possible pairs of means while controlling the family-wise error rate.
- **Dunnett's Test** (Charles Dunnett, 1955): Compares multiple treatments with a control.
- **Scheffé's Test** (Henry Scheffé, 1959): Allows comparison of any set of contrasts.
- **Holm-Bonferroni Method** (Sture Holm & Carlo Emilio Bonferroni, 1979): A stepwise approach to control the family-wise error rate.

Each of these tests brings a unique approach to handling multiple comparisons, and the choice of which to use depends largely on the specific context and objectives of your analysis.

2.1 The Core Difference: Multiple t-test vs. Tukey's Test

To better grasp the fundamental differences between the multiple t-test and Tukey's HSD test, let's compare their approaches side by side, focusing on how they control the Type I error rate and their implications:

Aspect	Multiple t-test	Tukey's HSD Test
Type I Error Rate Control	Controls the Type I error rate (α) for **each individual comparison** without adjustment for other tests.	Controls the Type I error rate (α) **across all comparisons collectively**, incorporating a multiplicity adjustment.
Multiplicity Adjustment	**Does not have any multiplicity adjustment**. Individual comparisons are treated independently.	**Has a multiplicity adjustment** to account for the number of comparisons, reducing the risk of collective Type I errors.
Conservativeness	More liberal, tending to find smaller p-values. Quicker to declare a difference as statistically significant.	More conservative, resulting in larger p-values. Less likely to declare a difference as statistically significant.
Implication on False Positives	Higher probability of false positives (Type I errors) in multiple comparisons due to no collective adjustment.	Lower probability of false positives across the set of comparisons, due to adjustment for collective error rate.
Implication on False Negatives	Lower risk of false negatives (Type II errors) due to its liberal approach.	Higher risk of false negatives (Type II errors), as its conservative approach may overlook some actual differences.

The multiple t-test evaluates each comparison in isolation, which makes it more likely to find significant differences, but with an increased risk of Type I errors when many comparisons are made. In contrast, Tukey's test takes a more holistic approach, adjusting for the collective risk across all comparisons, which makes it more stringent but also more reliable in the context of multiple testing.

2.2 So which one should I use?

While multiplicity adjustments are essential in controlling the overall error rate in multiple comparisons, they are not universally preferred in every situation. The decision between tests like the multiple t-test and the Tukey test is a balance of statistical rigor, research objectives, data characteristics, and field-specific conventions. This complexity is why statisticians often stop short of declaring a one-size-fits-all best choice, emphasizing instead the importance of context and research goals in guiding the selection of appropriate statistical methods.

In conclusion, it is true that the decision between tests such as the multiple t-test and the Tukey test is nuanced and context-dependent, but ultimately the most critical aspect of your analysis is transparency. It's essential to clearly report which test you've used when showing your results. Equally important is the principle of deciding on your test choice before examining the results. This approach ensures that your analysis is guided by your research question and methodological considerations, rather than being influenced by the data patterns you observe.

Additional Resources

- Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol. 2018 Oct;71(5):353-360. doi: 10.4097/kja.d.18.00242