

Pairwise Comparisons (Correlated Observations)

Prerequisites

[Difference between Two Means \(Independent Groups\)](#), [All Pairwise Comparisons Among Means](#), [Difference Between Two Means, Correlated Pairs](#), [Specific Comparisons](#)

In the section on all [pairwise comparisons among independent groups](#), the [Tukey HSD Test](#) was the recommended procedure. When you have one group with several scores from the same subjects, the Tukey test makes an assumption that is unlikely to hold: The variance of difference scores is the same for all pairwise differences between means.

The standard practice for pairwise comparisons with *correlated observations* is to compare each pair of means using the method outlined in the section "[Difference Between Two Means, Correlated Pairs](#)" with the addition of the *Bonferroni correction* described in the section "[Specific Comparisons](#)." For example, suppose you were going to do all pairwise comparisons among four means and hold the [familywise error rate](#) at 0.05. Since there are six possible pairwise comparisons among four means, you would use the $0.05/6 = 0.0083$ for the per comparison error rate.

As an example, consider the case study "[Stroop](#)." There were three tasks each performed by 47 subjects. In the "words" task, subjects read the names of 60 color words written in black ink; in the "color" task, subjects named the colors of 60 rectangles; in the "interference" task, subjects named the ink color of 60 conflicting color words. The times to read the stimuli were recorded. In order to do compute all pairwise comparisons, the difference in times for each pair of conditions for each subject is calculated. Table 1 shows these scores for 5 of the 42 subjects.

Table 1. Pairwise Differences

W-C	W-I	C-I
-3	-24	-21
2	-41	-43
-1	-18	-17
-4	-23	-19
-2	-17	-15

[Data for all 47 subjects](#)

The means, standard deviations, and [*standard error of the mean*](#) (Sem), t, and p for all 47 subjects are shown in Table 2. The t's are computed by dividing the means by the standard errors of the mean. Since there are 47 subject, the degrees of freedom is 46. Notice how different the standard deviations are. For the Tukey test to be valid, all population values of the standard deviation would have to be the same.

Table 2. Distribution of colors.

Comparison	Mean	Sd	Sem	t	p
W-C	-4.15	2.99	0.43	-9.53	<0.001
W-I	-20.51	7.84	1.14	-17.93	<0.001
C-I	-16.36	7.47	1.09	-15.02	<0.001

Using the Bonferroni correction for three comparisons, the p value has to be below $0.05/3 = 0.0167$ for an effect to be significant at the 0.05 level. For these data, all p values are far below that, and therefore all pairwise differences are significant.