All Pairwise Comparisons Among Means

Prerequisites
Difference Between Two Means (Independent Groups)

Many experiments are designed to compare more than two conditions. We will take as an example the case study Smiles and Leniency. In this study, the effect of different types of smiles on the leniency showed to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each other group mean. This procedure would lead to the six comparisons shown in Table 1.

Table 1. Six Comparisons among Means.

false vs felt		
false vs miserable		
false vs neutral		
felt vs miserable		

felt vs neutral		
miserable vs neutral		

The problem with this approach is that if you did this analysis, you would have six chances to make a *Type I error*. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the Type I error rate is inflated. Figure 1 shows the number of possible comparisons between pairs of means (*pairwise comparisons*) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

Figure 1. Number of Comparisons as a Function of the Number of Means.

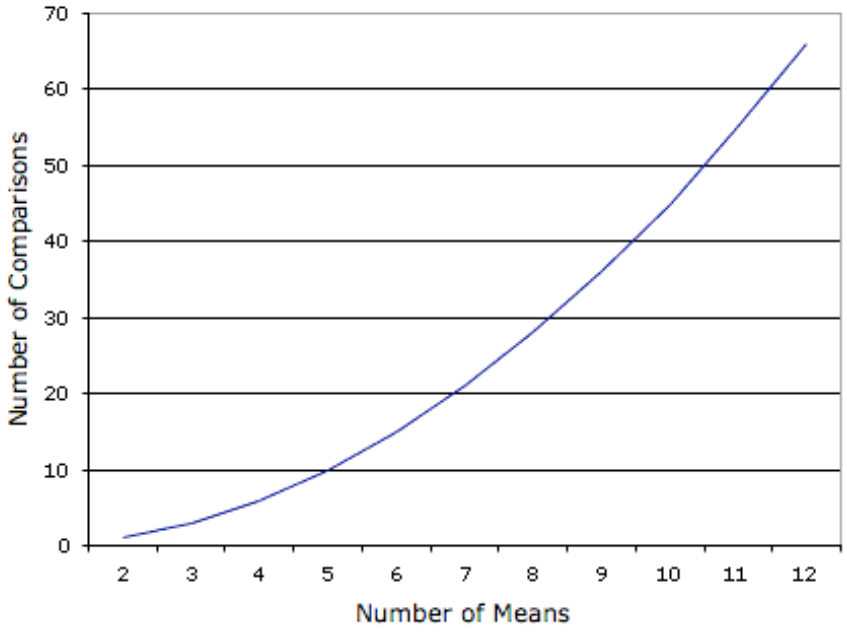
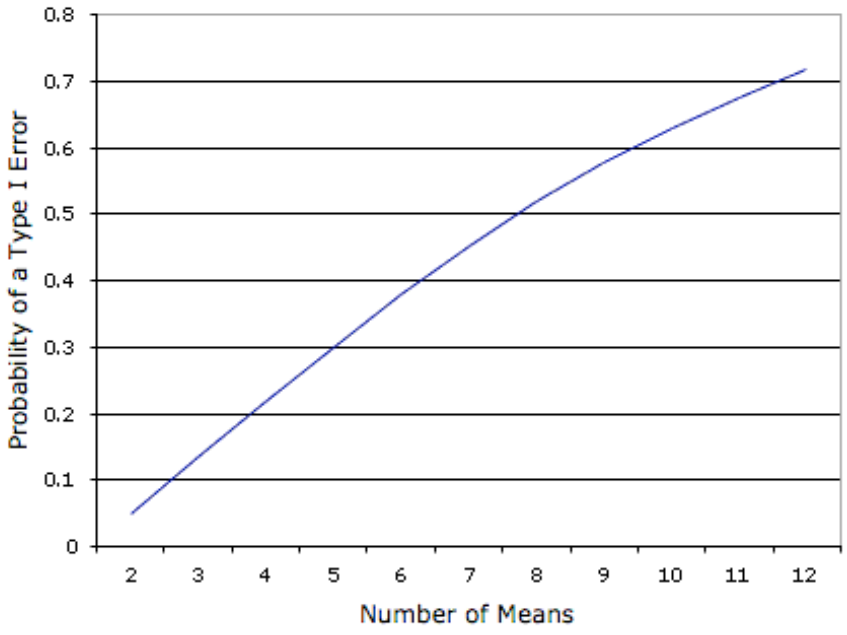


Figure 2 shows probability of a Type I error as a function of the number of means. As you can see in Figure 2, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even though all 12 population means are the same.

Figure 2. Probability of a Type I Error as a Function of the Number of Means.



The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or *Tukey HSD* for short. The Tukey HSD is based on a variation of the *t distribution* that takes into account the number of means being compared. This distribution is called the *studentized range distribution*.

Let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

Condition	Mean	Variance
False	5.37	3.34
Felt	4.91	2.83
Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE which is simply the mean of the variances. It is equal to 2.65.

3. Compute:

$$Q = \frac{M_i - M_j}{\sqrt{MSE/n}}$$

for each pair of means where M_i is one mean, M_j is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute p for each comparison using the Studentized Range Calculator. The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, df = 136 - 4 = 132.

Studentized Range Calculator

The tests for these data are shown in Table 1.

Table 1. Six Pairwise Comparisons.

Comparison	$M_i - M_j$	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt- Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

The only significant comparison is between the false smile and the neutral smile.

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in between.

Assumptions

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

Computer Analysis

For most computer programs, you should format your data the same way you

do for independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

Although full-featured statistics programs such as SAS, SPSS, R, and others can compute Tukey's test, smaller programs (including Analysis Lab) may not. However, these programs are generally able to compute a procedure known as Analysis of Variance (ANOVA). This procedure will be described in detail in a later chapter. Its relevance here is that an ANOVA computes the MSE that is used in the calculation of Tukey's test. For example, the following shows the ANOVA summary table for the smiles and leniency data.

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.4650	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The column labeled MS stands for "Mean Square" and therefore the value 2.6489 in the "Error" row and the MS column stands for Mean Squared Error or MSE. Recall that this is the same value computed here (2.65) when rounded off.

Tukey's Test Need not be a Follow-Up to ANOVA

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Leland Wilkinson and APA Board of Scientific Affairs Task Force on Statistical Inference published in the American Psychologist, August 1999, Vol. 54, No. 8, 594–604.

Computations for Unequal Sample Sizes (optional)

The calculation of MSE for unequal sample sizes is similar to its calculation in independent-groups t test. Here are the steps:

1. Compute a Sum of Squares Error (SSE) using the following formula

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2 + ... + \sum (X - M_k)^2$$

where M_i is the mean of the ith group and is the number of means.

2. Compute the degrees of freedom error (dfe) by subtracting the number of groups (k) from the total number of observations (N). Therefore:

dfe N - k.

3. Compute MSE by dividing SSE by dfe:
   MSE = SSE/dfe.

4. For each comparison of means, use the *harmonic mean* of the n's for the two means (n_h).

   All other aspects of the calculations are the same as when you have equal sample sizes.