11.2: Pairwise Comparisons of Means (Post-Hoc Tests)

If you do in fact reject H_0, then you know that at least two of the means are different. The ANOVA test does not tell which of those means are different, only that a difference exists. Most likely your sample means will be different from each other, but how different do they need to be for there to be a statistically significant difference?

To determine which means are significantly different, you need to conduct further tests. These post-hoc tests include the range test, multiple comparison tests, Duncan test, Student-Newman-Keuls test, Tukey test, Scheffé test, Dunnett test, Fisher's least significant different test, and the Bonferroni test, to name a few. There are more options, and there is no consensus on which test to use. These tests are available in statistical software packages such as R, Minitab and SPSS.

One should **never** use two-sample t-tests from the previous chapter. This would inflate the type I error.

The probability of at least one type I error increases exponentially with the number of groups you are comparing. Let us assume that $\alpha = 0.05$, then the probability that an observed difference between two groups that does not occur by chance is $1 - \alpha = 0.95$. If two comparisons are made, the probability that the observed difference is true is no longer 0.95. The probability is $(1 - \alpha)^2 = 0.9025$, and the P(Type I Error) = $1 - 0.9025 = 0.0975$. Therefore, the P(Type I Error) occurs if m comparisons are made is $1 - (1 - \alpha)m$.

For instance, if we are comparing the means of four groups: There would be $m = {}_4C_2 = 6$ different ways to compare the 4 groups: groups (1,2), (1,3), (1,4), (2,3), (2,4), and (3,4). The P(Type I Error) = $1 - (1 - \alpha)6 = 0.2649$. This is why a researcher should use ANOVA for comparing means, instead of independent t-tests.

There are many different methods to use. Many require special tables or software. We could actually just start with post-hoc tests, but they are a lot of work. If we run an ANOVA and we fail to reject the null hypothesis, then there is no need for further testing and it will save time if you were doing these steps by hand. Most statistical software packages give you the ANOVA table followed by the pairwise comparisons with just a change in the options menu. Keep in mind that Excel is not a statistical software and does not give pairwise comparisons.

We will use the Bonferroni Test, named after the mathematician Carlo Bonferroni. The Bonferroni Test uses the t-distribution table and is similar to previous t-tests that we have used, but adjusts α to the number of comparisons being made.

Carlo Bonferroni

The Bonferroni test is a statistical test for testing the difference between two population means (only done after an ANOVA test shows not all means are equal).

The formula for the Bonferroni test statistic is $t = \dfrac{\bar{x}_i - \bar{x}_j}{\sqrt{\left(MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right)}}$.

where $\bar{x}_i$ and $\bar{x}_j$ are the means of the samples being compared, n_i and n_j are the sample sizes, and MSW is the within-group variance from the ANOVA table.

The Bonferroni test critical value or p-value is found by using the t-distribution with within degrees of freedom $df_W = N - k$, using an adjusted $\frac{\alpha}{m}$ two-tail area under the t-distribution, where k = number of groups and $m = {}_kC_2$, all the combinations of pairs out of k groups.

Critical Value Method

According to the ANOVA test that we previously performed, there does appear to be a difference in the average age of assistant professors (μ_1), associate professors (μ_2), and full professors (μ_3) at this university.

Source	SS	*df*	MS	F
Between	1208.6667	2	604.3333	12.6195
Within	862	18	47.8889	
Total	2070.6667	20		

The hypotheses were:

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least one mean differs.

The decision was to reject H_0, which means there is a significant difference in the mean age. The ANOVA test does not tell us, though, where the differences are. Determine which of the difference between each pair of means is significant. That is, test if $\mu_1 \neq \mu_2$, if $\mu_1 \neq \mu_3$, and if $\mu_2 \neq \mu_3$.

Solution

The alternative hypothesis for the ANOVA was "at least one mean is different." There will be ${}_3C_2 = 3$ subsequent hypothesis tests to compare all the combinations of pairs (Group 1 vs. Group 2, Group 1 vs. Group 3, and Group 2 vs. Group 3). Note that if you have 4 groups then you would have to do ${}_4C_2 = 6$ comparisons, etc.

Use the t-distribution to find the critical value for the Bonferroni test. The total of all the individual sample sizes $N = 21$ and $k = 3$, and $m = {}_3C_2 = 3$, then the area for both tails would be $\frac{\alpha}{m} = \frac{0.01}{m} = 0.003333$.

This is a two-tailed test so the area in one tail is $\frac{0.003333}{2}$ with $df_W = N - k = 21 - 3 = 18$ gives $\text{C.V.} = \pm 3.3804$. The critical values are really far out in the tail so it is hard to see the shaded area. See Figure 11-3.

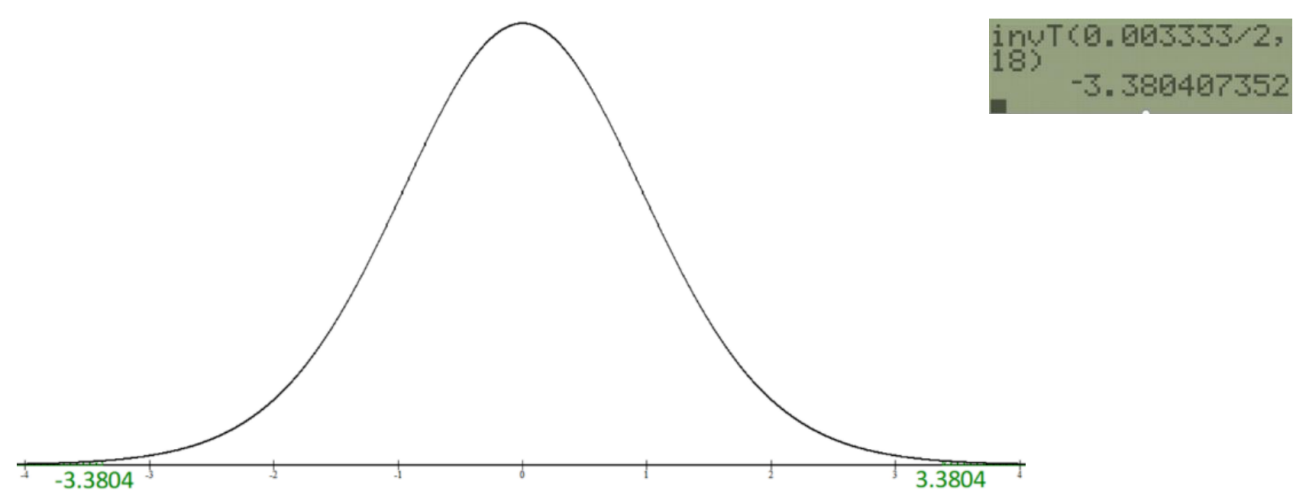


Figure 11-3: t-distribution with critical values.

Compare μ_1 and μ_2:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

The test statistic is $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \frac{37 - 52}{\sqrt{\left(47.8889\left(\frac{1}{7} + \frac{1}{7}\right)\right)}} = -4.0552$.

Compare the test statistic to the critical value. Since the test statistic $-4.0552 <$ critical value $= -3.3804$ we reject H_0.

There is enough evidence to conclude that there is a difference in the average age of assistant and associate professors.

Compare μ_1 and μ_3:

$H_0 : \mu_1 = \mu_3$

$H_1 : \mu_1 \neq \mu_3$

The test statistic is $t = \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_1}+\frac{1}{n_3}\right)\right)}} = \frac{37-54}{\sqrt{\left(47.8889\left(\frac{1}{7}+\frac{1}{7}\right)\right)}} = -4.5958$.

Compare the test statistic to the critical value. Since the test statistic $-4.5958 <$ critical value $= -3.3804$ we reject H_0.

Reject H_0, since the test statistic is in the lower tail. There is enough evidence to conclude that there is a difference in the average age of assistant and full professors.

Compare μ_2 and μ_3:

$H_0 : \mu_2 = \mu_3$

$H_1 : \mu_2 \neq \mu_3$

The test statistic is $t = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_2}+\frac{1}{n_3}\right)\right)}} = \frac{52-54}{\sqrt{\left(47.8889\left(\frac{1}{7}+\frac{1}{7}\right)\right)}} = -0.5407$

Compare the test statistic to the critical value. Since the test statistic is between the critical values $-3.3804 < -0.5407 < 3.3804$ we fail to reject H_0.

Do not reject H_0, since the test statistic is between the two critical values. There is enough evidence to conclude that there is not a difference in the average age of associate and full professors.

Note: you should get at least one group that has a reject H_0, since you only do the Bonferroni test if you reject H_0 for the ANOVA. Also, note that the transitive property does not apply. It could be that group 1 = group 2 and group 2 = group 3; this does not mean that group 1 = group 3.

P-Value Method

A research organization tested microwave ovens. At α = 0.10, is there a significant difference in the average prices of the three types of oven?

1000-watts	270	245	190	215	250	230		
900-watts	240	135	160	230	250	200	200	210
800-watts	180	155	200	120	140	180	140	130

Solution

The ANOVA was run in Excel.

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
1000-watts	6	1400	233.3333	796.6667
900-watts	8	1625	203.125	1549.554
800-watts	8	1245	155.625	795.9821

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	21729.73	2	10864.87	10.1182	0.001019	2.61
Within Groups	20402.08	19	1073.794			
Total	42131.82	21				

To test if there is a significant difference in the average prices of the three types of oven, the hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least one mean differs.

Use the Excel output to find the p-value in the ANOVA table of 0.001019, which is less than α so reject H_0; there is at least one mean that is different in the average oven prices.

There is a statistically significant difference in the average prices of the three types of oven. Use the Bonferroni test p-value method to see where the differences are.

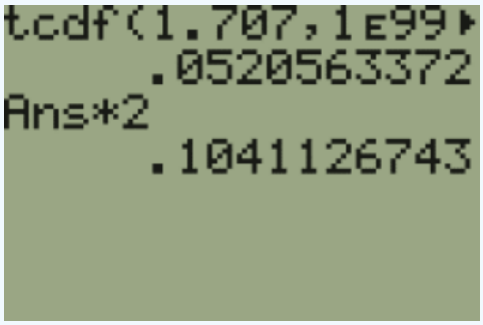
Compare μ_1 and μ_2:

$H_0 : \mu_1 = \mu_2$

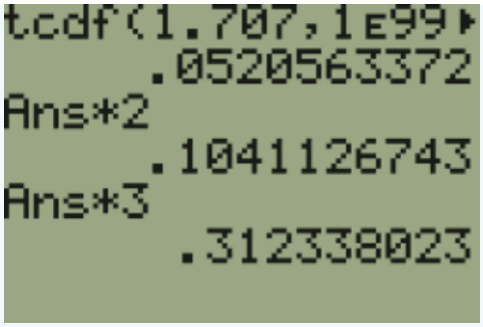
$H_1 : \mu_1 \neq \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \frac{233.3333 - 203.125}{\sqrt{\left((1073.794\left(\frac{1}{6} + \frac{1}{8}\right)\right)}} = 1.7070$$

To find the p-value, find the area in both tails and multiply this area by m. The area to the right of $t = 1.707$, using $df_W = 19$, is 0.0520563. Remember these are always two-tail tests, so multiply this area by 2, to get both tail areas of 0.104113.



Then multiply this area by $m = {}_3C_2 = 3$ to get a p-value = 0.3123.



Since the p-value = $0.3123 > \alpha = 0.10$, we do not reject H_0. There is a statistically significant difference in the average price of the 1,000- and 900-watt ovens.

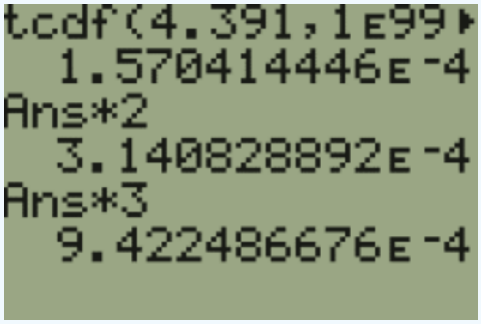
Compare μ_1 and μ_3:

$H_0 : \mu_1 = \mu_3$

$H_1 : \mu_1 \neq \mu_3$

$t = \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_3}\right)\right)}} = \frac{233.3333 - 155.625}{\sqrt{\left((1073.794\left(\frac{1}{6} + \frac{1}{8}\right)\right)}} = 4.3910$

Use df_W = 19 to find the p-value.



Since the p-value = (tail areas)*3 = $0.00094 < \alpha = 0.10$, we reject H_0. There is a statistically significant difference in the average price of the 1,000- and 800-watt ovens.

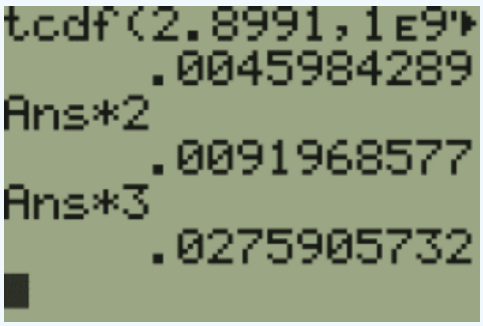
Compare μ_2 and μ_3:

$H_0 : \mu_2 = \mu_3$

$H_1 : \mu_2 \neq \mu_3$

$t = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_2} + \frac{1}{n_3}\right)\right)}} = \frac{203.125 - 155.625}{\sqrt{\left((1073.794\left(\frac{1}{8} + \frac{1}{8}\right)\right)}} = 2.8991$

Use $df_W = 19$ to find the p-value (remember that these are always two-tail tests).



Since the p-value = $0.0276 < \alpha = 0.10$, we reject H_0. There is a statistically significant difference in the average price of the 900- and 800-watt ovens.

There is a chance that after we multiply the area by the number of comparisons, the p-value would be greater than one. However, since the p-value is a probability we would cap the probability at one.

This is a lot of math! The calculators and Excel do not have post-hoc pairwise comparisons shortcuts, but we can use the statistical software called SPSS to get the following results. We will look specifically at interpreting the SPSS output for Example 11-4.

Descriptives

Price

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
800	8	155.6250	28.21316	9.97486	132.0382	179.2118	120.00	200.00
900	8	203.1250	39.36437	13.91741	170.2156	236.0344	135.00	250.00
1000	6	233.3333	28.22528	11.52292	203.7127	262.9540	190.00	270.00
Total	22	194.0909	44.79148	9.54958	174.2315	213.9503	120.00	270.00

ANOVA

Price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	21729.735	2	10864.867	10.118	.001
Within Groups	20402.083	19	1073.794		
Total	42131.818	21			

Multiple Comparisons

Dependent Variable: Price

Bonferroni

(I) Watts	(J) Watts	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval	
					Lower Bound	Upper Bound
1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126
	800-Watts	77.70833*	17.69717	.001	37.1041	118.3126
900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959
	800-Watts	47.50000*	16.38440	.028	9.9078	85.0922
800-Watts	1000-Watts	-77.70833*	17.69717	.001	-118.3126	-37.1041
	900-Watts	-47.50000*	16.38440	.028	-85.0922	-9.9078

*. The mean difference is significant at the 0.10 level.

Figure 11-4: Multiple Comparisons table.

The first table, labeled "Descriptives", gives descriptive statistics; the second table is the ANOVA table, and note that the p-value is in the column labeled Sig. The Multiple Comparisons table is where we want to look. There are repetitive pairs in the last table, just in a different order.

The first two rows in Figure 11-4 are comparing group 1 with groups 2 and 3. If we follow the first row across under the Sig. column, this gives the p-value = 0.312 for comparing the 1,000- and 900-watt ovens.

1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126

The second row in Figure 11-4 compares the 1,000- and 800-watt ovens, p-value = 0.001.

1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126
	800-Watts	77.70833*	17.69717	.001	37.1041	118.3126

The third row in Figure 11-4 compares the 900- and 1000-watt ovens in the reverse order as the first row; note that the difference in the means is negative but the p-value is the same.

900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959

The fourth row in Figure 11-4 compares the 900- and 800-watt ovens, p-value = 0.028.

900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959
	800-Watts	47.50000*	16.38440	.028	9.9078	85.0922

The last set of rows in Figure 11-4 are again repetitive and give the 800-watt oven compared to the 900- and 1000-watt ovens.

Keep in mind that post-hoc is defined as occurring after an event. A post-hoc test is done after an ANOVA test shows that there is a statistically significant difference. You should get at least one group that has a result of "reject H_0", since you only do the Bonferroni test if you reject H_0 for the ANOVA.