

## Chapter 11: Comparing several means

### Smart Alex's Solutions

#### Task 1

To test how different teaching methods affected students' knowledge I took three statistics courses where I taught the same material. For one course I wandered around with a large cane and beat anyone who asked daft questions or got questions wrong (punish). In the second I encouraged students to discuss things that they found difficult and gave anyone working hard a nice sweet (reward). For the final course I remained indifferent and neither punished nor rewarded students' efforts (indifferent). As the dependent measure I took the students' percentage exam marks. The data are in the file **Teach.sav**. Carry out a one-way ANOVA and use planned comparisons to test the hypotheses that: (1) reward results in better exam results than either punishment or indifference; and (2) indifference will lead to significantly better exam results than punishment.

Descriptives

Exam Mark	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
					Punish	10		
Indifferent	10	56.0000	7.10243	2.24598	50.9192	61.0808	46.00	67.00
Reward	10	65.4000	4.29987	1.35974	62.3241	68.4759	58.00	71.00
Total	30	57.1333	8.26181	1.50839	54.0483	60.2183	45.00	71.00

#### Output 1

Output shows the table of descriptive statistics from the one-way ANOVA; we're told the means, standard deviations and standard errors of the means for each experimental condition. The means should correspond to those plotted in the graph. These diagnostics are important for interpretation later on. It looks as though marks are highest after reward and lowest after punishment.

**Test of Homogeneity of Variances**

Exam Mark			
Levene Statistic	df1	df2	Sig.
2.569	2	27	.095

**Output 2**

The next part of the output (Output ) reports a test of the assumption of homogeneity of variance (Levene's test). For these data, the assumption of homogeneity of variance has been met, because our significance is .095, which is bigger than the criterion of .05.

**ANOVA**

Exam Mark					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1205.067	2	602.533	21.008	.000
Within Groups	774.400	27	28.681		
Total	1979.467	29			

**Output 3**

Output is the main ANOVA summary table; it shows us that because the observed significance value is less than .05 we can say that there was a significant effect of teaching style on exam marks. However, at this stage we still do not know exactly what the effect of the teaching style was (we don't know which groups differed).

**Robust Tests of Equality of Means**

Exam Mark				
	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	32.235	2	17.336	.000
Brown-Forsythe	21.008	2	20.959	.000

a. Asymptotically F distributed.

**Output 4**

Output shows the Welch and Brown-Forsythe *F*s, but we can ignore these because the homogeneity of variance assumption was met.

**Contrast Coefficients**

Contrast	Type of Teaching Method		
	Punish	Indifferent	Reward
1	1	1	-2
2	1	-1	0

**Output 5**

Because there were specific hypotheses I specified some contrasts. Output shows the codes I used. The first contrast compares reward (coded with -2) against punishment and

indifference (both coded with 1). The second contrast compares punishment (coded with 1) against indifference (coded with -1). Note that the codes for each contrast sum to zero, and that in contrast 2, reward has been coded with a 0 because it is excluded from that contrast.

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Exam Mark	Assume equal variances	1	-24.8000	4.14836	-5.978	27	.000
		2	-6.0000	2.39506	-2.505	27	.019
	Does not assume equal variances	1	-24.8000	3.76180	-6.593	21.696	.000
		2	-6.0000	2.59915	-2.308	14.476	.036

#### Output 6

Output shows the significance of the two contrasts specified above. Because homogeneity of variance was met, we can ignore the part of the table labelled *Does not assume equal variances*. The *t*-test for the first contrast tells us that reward was significantly different from punishment and indifference (it's significantly different because the value in the column labelled *Sig.* is less than .05). Looking at the means, this tells us that the average mark after reward was significantly higher than the average mark for punishment and indifference combined. The second contrast (together with the descriptive statistics) tells us that the marks after punishment were significantly lower than after indifference (again, significantly different because the value in the column labelled *Sig.* is less than .05). As such we could conclude that reward produces significantly better exam grades than punishment and indifference, and that punishment produces significantly worse exam marks than indifference. So lecturers should reward their students, not punish them.

## Task 2

*Compute the effect sizes for the previous task.*

### Calculating the effect size

Output provides us with three measures of variance: the between-group effect ( $SS_M$ ), the within-subject effect ( $MS_R$ ) and the total amount of variance in the data ( $SS_T$ ). We can use these to calculate omega squared ( $\omega^2$ ):

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$$

Substituting from Output 3:

$$\omega^2 = \frac{1205.067 - (2)28.681}{1979.467 + 28.681}$$

$$\begin{aligned}
 &= \frac{1147.705}{2008.148} \\
 &= .57 \\
 \omega &= .76
 \end{aligned}$$

For the contrasts the effect sizes will be:

$$r_{\text{contrast}} = \sqrt{\frac{t^2}{t^2 + df}}$$

For contrast 1 we get:

$$\begin{aligned}
 r_{\text{contrast1}} &= \sqrt{\frac{(-5.978)^2}{(-5.978)^2 + 27}} \\
 &= .75
 \end{aligned}$$

If you think back to our benchmarks for effect sizes this represents a huge effect (it is well above .5, the threshold for a large effect). Therefore, as well as being statistically significant, this effect is large and so represents a substantive finding. For contrast 2 we get:

$$\begin{aligned}
 r_{\text{contrast2}} &= \sqrt{\frac{(-2.505)^2}{(-2.505)^2 + 27}} \\
 &= .43
 \end{aligned}$$

This too is a substantive finding and represents a medium to large effect size.

## Interpreting and writing the result

The correct way to report the main finding would be:

- ✓ All significant values are reported at  $p < .05$ . There was a significant effect of teaching style on exam marks,  $F(2, 27) = 21.01$ ,  $\omega^2 = .57$ . Planned contrasts revealed that reward produced significantly better exam grades than punishment and indifference,  $t(27) = -5.98$ ,  $r = .75$ , and that punishment produced significantly worse exam marks than indifference,  $t(27) = -2.51$ ,  $r = .43$ .

## Task 3

Children wearing superhero costumes are more likely to harm themselves because of the unrealistic impression of invincibility that these costumes could create. For example, children have reported to hospital with severe injuries because of trying 'to initiate flight without having planned for landing strategies' (Davies, SurrIDGE, Hole, & Munro-Davies, 2007). I can relate to the imagined power that a costume bestows upon you; even now, I have been known to dress up as Fisher by donning a beard and glasses and trailing a goat around on a lead in the hope that it might make me more knowledgeable about statistics. Imagine we had data (**Superhero.sav**) about the severity of **injury** (on a scale from 0, no injury, to 100, death) for children reporting to the emergency centre at hospitals and information on which superhero costume they were wearing (**hero**): Spiderman, Superman, the Hulk or a teenage mutant ninja turtle. Use one-way ANOVA and multiple comparisons to test the hypotheses that different costumes give rise to more severe injuries.

### Descriptives

Injury Severity (0-100)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Spiderman	8	41.6250	12.21167	4.31748	31.4158	51.8342	20.00	58.00
Superman	6	60.3333	17.85124	7.28774	41.5996	79.0671	32.00	85.00
Hulk	8	35.3750	13.38376	4.73187	24.1859	46.5641	10.00	53.00
Ninja Turtle	8	26.2500	8.15475	2.88314	19.4325	33.0675	18.00	41.00
Total	30	39.6000	17.15769	3.13255	33.1932	46.0068	10.00	85.00

### Output 7

Looking at the means in Output , it seems that children wearing a Ninja Turtle costume had the least severe injuries ( $M = 26.25$ ), whereas children wearing a Superman costume had the most severe injuries ( $M = 60.33$ ).

### Test of Homogeneity of Variances

Injury Severity (0-100)

Levene Statistic	df1	df2	Sig.
.891	3	26	.459

### Output 8

Looking at Output , we can see that the assumption of homogeneity of variance has been met because our significance value is .46, which is much bigger than the criterion of .05.

### ANOVA

#### Injury Severity (0-100)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4180.617	3	1393.539	8.317	.000
Within Groups	4356.583	26	167.561		
Total	8537.200	29			

#### Output 9

In the ANOVA output (Output ), the observed significance value is much less than .05 and so we can say that there was a significant effect of superhero costume on injury severity. However, at this stage we still do not know exactly what the effect of superhero costume was (we don't know which groups differed).

Because there were no specific hypotheses, only that the groups would differ, we can't look at planned contrasts but we can conduct some *post hoc* tests. I am going to use Gabriel's *post hoc* test because the group sizes are slightly different (Spiderman,  $N = 8$ ; Superman,  $N = 6$ ; Hulk,  $N = 8$ ; Ninja Turtle,  $N = 8$ ).

### Multiple Comparisons

Dependent Variable: Injury Severity (0-100)  
Gabriel

(I) Type of Superhero Costume	(J) Type of Superhero Costume	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Spiderman	Superman	-18.70833	6.99085	.070	-38.4621	1.0455
	Hulk	6.25000	6.47227	.907	-12.0855	24.5855
	Ninja Turtle	15.37500	6.47227	.136	-2.9605	33.7105
Superman	Spiderman	18.70833	6.99085	.070	-1.0455	38.4621
	Hulk	24.95833*	6.99085	.008	5.2045	44.7121
	Ninja Turtle	34.08333*	6.99085	.000	14.3295	53.8371
Hulk	Spiderman	-6.25000	6.47227	.907	-24.5855	12.0855
	Superman	-24.95833*	6.99085	.008	-44.7121	-5.2045
	Ninja Turtle	9.12500	6.47227	.650	-9.2105	27.4605
Ninja Turtle	Spiderman	-15.37500	6.47227	.136	-33.7105	2.9605
	Superman	-34.08333*	6.99085	.000	-53.8371	-14.3295
	Hulk	-9.12500	6.47227	.650	-27.4605	9.2105

\*. The mean difference is significant at the 0.05 level.

#### Output 10

Output tells us that wearing a Superman costume was significantly different from wearing either a Hulk or Ninja Turtle costume in terms of injury severity, but that none of the other groups differed significantly.

The *post hoc* test has shown us which differences between means are significant; however, if we want to see the direction of the effects we can look back to the means in the table of descriptives (Output 7). We can conclude that wearing a Superman costume resulted in significantly more severe injuries than wearing either a Hulk or a Ninja Turtle costume.

### Calculating the effect size

Output provides us with three measures of variance: the between-group effect ( $SS_M$ ), the within-subject effect ( $MS_R$ ) and the total amount of variance in the data ( $SS_T$ ). We can use these to calculate omega squared ( $\omega^2$ ):

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$$

Substituting from Output 9:

$$\begin{aligned}\omega^2 &= \frac{4180.617 - (3)167.561}{8537.20 + 167.561} \\ &= \frac{3677.934}{8704.761} \\ &= .42 \\ \omega &= .65\end{aligned}$$

### Interpreting and writing the result

The correct way to report the main finding would be:

- ✓ All significant values are reported at  $p < .05$ . There was a significant effect of superhero costume on severity of injury,  $F(3, 26) = 8.32$ ,  $\omega^2 = .42$ . Gabriel's *post hoc* tests revealed that wearing a Superman costume resulted in significantly more severe injuries than wearing either a Hulk costume  $p = .008$ , or a Ninja Turtle costume.

## Task 4

*In Chapter 6 (Section 6.6) there are some data looking at whether eating soya meals reduces your sperm count. Have a look at this section, access the data for that example,*

*but analyse them with ANOVA. What's the difference between what you find and what is found in section 6.6.5? Why do you think this difference has arisen?*

**Descriptives**

Sperm Count (Millions)									
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	
					Lower Bound	Upper Bound			
No Soya Meals	20	4.9868	5.08437	1.13690	2.6072	7.3663	.35	21.08	
1 Soya Meal Per Week	20	4.6052	4.67263	1.04483	2.4184	6.7921	.33	18.47	
4 Soya Meals Per Week	20	4.1101	4.40991	.98609	2.0462	6.1740	.40	18.21	
7 Soya Meals Per Week	20	1.6530	1.10865	.24790	1.1341	2.1719	.31	4.11	
Total	80	3.8388	4.26048	.47634	2.8906	4.7869	.31	21.08	

**Output 1**

Output shows the table of descriptive statistics from the one-way ANOVA. It looks as though as soya intake increases, sperm counts do indeed decrease.

**Test of Homogeneity of Variances**

Sperm Count (Millions)			
Levene Statistic	df1	df2	Sig.
5.117	3	76	.003

**Output 2**

The next part of the output (Output ) reports a test of the assumption of homogeneity of variance (Levene's test). For these data, the assumption of homogeneity of variance has been broken, because our significance is .003, which is smaller than the criterion of .05. In fact, these data also violate the assumption of normality (see Chapter 6, on non-parametric statistics).

**ANOVA**

Sperm Count (Millions)					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	135.130	3	45.043	2.636	.056
Within Groups	1298.853	76	17.090		
Total	1433.983	79			

**Output 3**

Output is the main ANOVA summary table; it shows us that because the observed significance value is greater than .05 we can say that there was no significant effect of soya intake on men's sperm count. This is strange because if you read Chapter 6, from where this example came, the Kruskal-Wallis test produced a significant result! The reason for this difference is that the data violate the assumptions of normality and homogeneity of variance. As I mention in Chapter 6, although parametric tests have more power to detect effects when



their assumptions are met, when their assumptions are violated non-parametric tests have more power! This example was arranged to prove this point: because the parametric assumptions are violated, the non-parametric tests produced a significant result and the parametric test did not because, in these circumstances, the non-parametric test has the greater power!

**Robust Tests of Equality of Means**

Sperm Count (Millions)

	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	6.284	3	34.657	.002
Brown-Forsythe	2.636	3	58.236	.058

a. Asymptotically F distributed.

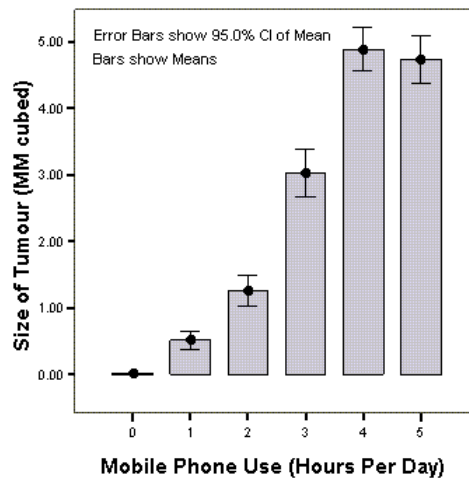
#### Output 14

Output shows the Welch and Brown–Forsythe  $F$ s; note that the Welch test agrees with the non-parametric test in that the significance of  $F$  is below the .05 threshold. However, the Brown-Forsythe  $F$  is non-significant (it is just above the threshold). This illustrates the relative superiority of the Welch procedure. However, in these circumstances, because normality *and* homogeneity of variance have been violated we'd use a non-parametric test anyway!

## Task 5

*Mobile phones emit microwaves, and so holding one next to your brain for large parts of the day is a bit like sticking your brain in a microwave oven and pushing the 'cook until well done' button. If we wanted to test this experimentally, we could get six groups of people and strap a mobile phone on their heads (so that they can't remove it). Then, by remote control, we turn the phones on for a certain amount of time each day. After six months, we measure the size of any tumour (in mm<sup>3</sup>) close to the site of the phone antenna (just behind the ear). The six groups experienced 0, 1, 2, 3, 4 or 5 hours per day of phone microwaves for six months. Carry out an ANOVA to see if tumours increased with greater daily exposure. The data are in **Tumour.sav**.*

The following figure displays the error bar chart of the mobile phone data shows the mean size of brain tumour in each condition, and the funny 'I' shapes show the confidence interval of these means:



Note that in the control group (0 hours), the mean size of the tumour is virtually zero (we wouldn't actually expect them to have a tumour) and the error bar shows that there was very little variance across samples. We'll see later that this is problematic for the analysis.

**Descriptives**

Size of Tumour (MM cubed)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
0	20	.0175	.01213	.00271	.0119	.0232	.00	.04
1	20	.5149	.28419	.06355	.3819	.6479	.00	.94
2	20	1.2614	.49218	.11005	1.0310	1.4917	.48	2.34
3	20	3.0216	.76556	.17118	2.6633	3.3799	1.77	4.31
4	20	4.8878	.69625	.15569	4.5619	5.2137	3.04	6.05
5	20	4.7306	.78163	.17478	4.3648	5.0964	2.70	6.14
Total	120	2.4056	2.02662	.18500	2.0393	2.7720	.00	6.14

**Output 15**

Output shows the table of descriptive statistics from the one-way ANOVA; we're told the means, standard deviations and standard errors of the means for each experimental condition. The means should correspond to those plotted in the graph. These diagnostics are important for interpretation later on.

**Test of Homogeneity of Variances**

Size of Tumour (MM cubed)

Levene Statistic	df1	df2	Sig.
10.245	5	114	.000

**Output 16**

Output reports a test of this assumption, Levene's test. For these data, the assumption of homogeneity of variance has been violated, because our significance is .000, which is considerably smaller than the criterion of .05. In these situations, we have to try to correct the problem, and we can either transform the data or choose the Welch *F*.

#### ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	450.664	5	90.133	269.733	.000
Within Groups	38.094	114	.334		
Total	488.758	119			

#### Output 17

Output is the main ANOVA summary table and shows us that because the observed significance value is less than .05 we can say that there was a significant effect of mobile phones on the size of tumour. However, at this stage we still do not know exactly what the effect of the phones was (we don't know which groups differed).

#### Robust Tests of Equality of Means

	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	414.926	5	44.390	.000
Brown-Forsythe	269.733	5	75.104	.000

a. Asymptotically F distributed.

#### Output 18

Output shows the Welch and Brown-Forsythe *F*s, which are useful because homogeneity of variance was violated. Luckily our conclusions remain the same: both *F*s have significance values less than .05.

## Multiple Comparisons

Dependent Variable: Size of Tumour (MM cubed)

Games-Howell

(I) Mobile Phone Use (Hours Per Day)	(J) Mobile Phone Use (Hours Per Day)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
0	1	-.4973*	.18280	.000	-.6982	-.2964
	2	-1.2438*	.18280	.000	-1.5916	-.8960
	3	-3.0040*	.18280	.000	-3.5450	-2.4631
	4	-4.8702*	.18280	.000	-5.3622	-4.3783
	5	-4.7130*	.18280	.000	-5.2653	-4.1608
1	0	.4973*	.18280	.000	.2964	.6982
	2	-.7465*	.18280	.000	-1.1327	-.3603
	3	-2.5067*	.18280	.000	-3.0710	-1.9424
	4	-4.3729*	.18280	.000	-4.8909	-3.8549
	5	-4.2157*	.18280	.000	-4.7908	-3.6406
2	0	1.2438*	.18280	.000	.8960	1.5916
	1	.7465*	.18280	.000	.3603	1.1327
	3	-1.7602*	.18280	.000	-2.3762	-1.1443
	4	-3.6264*	.18280	.000	-4.2017	-3.0512
	5	-3.4692*	.18280	.000	-4.0949	-2.8436
3	0	3.0040*	.18280	.000	2.4631	3.5450
	1	2.5067*	.18280	.000	1.9424	3.0710
	2	1.7602*	.18280	.000	1.1443	2.3762
	4	-1.8662*	.18280	.000	-2.5607	-1.1717
	5	-1.7090*	.18280	.000	-2.4429	-.9751
4	0	4.8702*	.18280	.000	4.3783	5.3622
	1	4.3729*	.18280	.000	3.8549	4.8909
	2	3.6264*	.18280	.000	3.0512	4.2017
	3	1.8662*	.18280	.000	1.1717	2.5607
	5	.1572	.18280	.984	-.5455	.8599
5	0	4.7130*	.18280	.000	4.1608	5.2653
	1	4.2157*	.18280	.000	3.6406	4.7908
	2	3.4692*	.18280	.000	2.8436	4.0949
	3	1.7090*	.18280	.000	.9751	2.4429
	4	-.1572	.18280	.984	-.8599	.5455

\*. The mean difference is significant at the .05 level.

## Output 19

Because there were no specific hypotheses I just carried out *post hoc* tests and stuck to my favourite Games–Howell procedure (because variances were unequal). It is clear from Output that each group of participants is compared to all of the remaining groups. First, the control group (0 hours) is compared to the 1, 2, 3, 4 and 5 hour groups and reveals a significant difference in all cases (all the values in the column labelled *Sig.* are less than .05). In the next part of the table, the 1 hour group is compared to all other groups. Again all comparisons are significant (all the values in the column labelled *Sig.* are less than .05). In fact, all of the comparisons appear to be highly significant except the comparison between the 4 and 5 hour groups, which is non-significant because the value in the column labelled *Sig.* is bigger than .05.

## Calculating the effect size

Output provides us with three measures of variance: the between-group effect ( $SS_M$ ), the within-subject effect ( $MS_R$ ) and the total amount of variance in the data ( $SS_T$ ). We can use these to calculate omega squared ( $\omega^2$ ):

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$$

Substituting from Output 17:

$$\begin{aligned}\omega^2 &= \frac{450.664 - (5)0.334}{488.758 + 0.334} \\ &= \frac{448.994}{488.424} \\ &= .92 \\ \omega &= .96\end{aligned}$$

## Interpreting and Writing the Result

We could report the main finding as follows:

- ✓ Levene's test indicated that the assumption of homogeneity of variance had been violated,  $F(5, 114) = 10.25, p < .001$ , so Welch's  $F$  is reported. The results show that using a mobile phone significantly affected the size of brain tumour found in participants,  $F(5, 44.39) = 414.93, p < .001, \omega^2 = .69$ . The effect size indicated that the effect of phone use on tumour size was substantial.

The next thing that needs to be reported are the *post hoc* comparisons. It is customary just to summarize these tests in very general terms like this:

- ✓ Games–Howell *post hoc* tests revealed significant differences between all groups ( $p < .001$  for all tests) except between 4 and 5 hours (*ns*).

If you do want to report the results for each *post hoc* test individually, then at least include the 95% confidence intervals for the test as these tell us more than just the significance value. In this example, though, when there are many tests it might be as well to summarize these confidence intervals as a table:

Mobile Phone Use (Hours Per Day)		Sig.	95% Confidence Interval	
			Lower Bound	Upper Bound
0	1	< .001	-0.6982	-0.2964
	2	< .001	-1.5916	-0.8960
	3	< .001	-3.5450	-2.4631
	4	< .001	-5.3622	-4.3783
	5	< .001	-5.2653	-4.1608
1	2	< .001	-1.1327	-0.3603
	3	< .001	-3.0710	-1.9424
	4	< .001	-4.8909	-3.8549
	5	< .001	-4.7908	-3.6406
2	3	< .001	-2.3762	-1.1443
	4	< .001	-4.2017	-3.0512
	5	< .001	-4.0949	-2.8436
3	4	< .001	-2.5607	-1.1717
	5	< .001	-2.4429	-0.9751
4	5	= .984	-0.5455	0.8599

## Task 6

Using the Glastonbury data from Chapter 8 (*GlastonburyFestival.sav*), carry out a one-way ANOVA on the data to see if the change in hygiene (*change*) is significant across people with different musical tastes (*music*). Do a simple contrast to compare each group against 'No Affiliation'. Compare the results to those described in Section 10.5.

### Levene's Test of Equality of Error Variances<sup>a</sup>

Dependent Variable: Change in Hygiene Over The Festival

F	df1	df2	Sig.
.866	3	119	.461

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + music

### Output 4

Looking at Output , we can see that Levene's test is non-significant, indicating that variances were roughly equal,  $F(3, 119) = 0.87, p > .05$ , across crusties, metallers, indie kids and people with no affiliation.

**ANOVA**

Change in Hygiene Over The Festival

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.646	3	1.549	3.270	.024
Within Groups	56.358	119	.474		
Total	61.004	122			

**Output 5**

Output 5 is the main ANOVA table. We could say that the change in hygiene scores was significantly different across the different musical groups,  $F(3, 119) = 3.27, p < .05$ . Compare this table to the one in Section 7.11 (Output 4), in which we analysed these data as a regression:

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.646	3	1.549	3.270	.024 <sup>a</sup>
	Residual	56.358	119	.474		
	Total	61.004	122			

a. Predictors: (Constant), No Affiliation vs. Indie Kid, No Affiliation vs. Crusty, No Affiliation vs. Metaller


b. Dependent Variable: Change in Hygiene Over The Festival

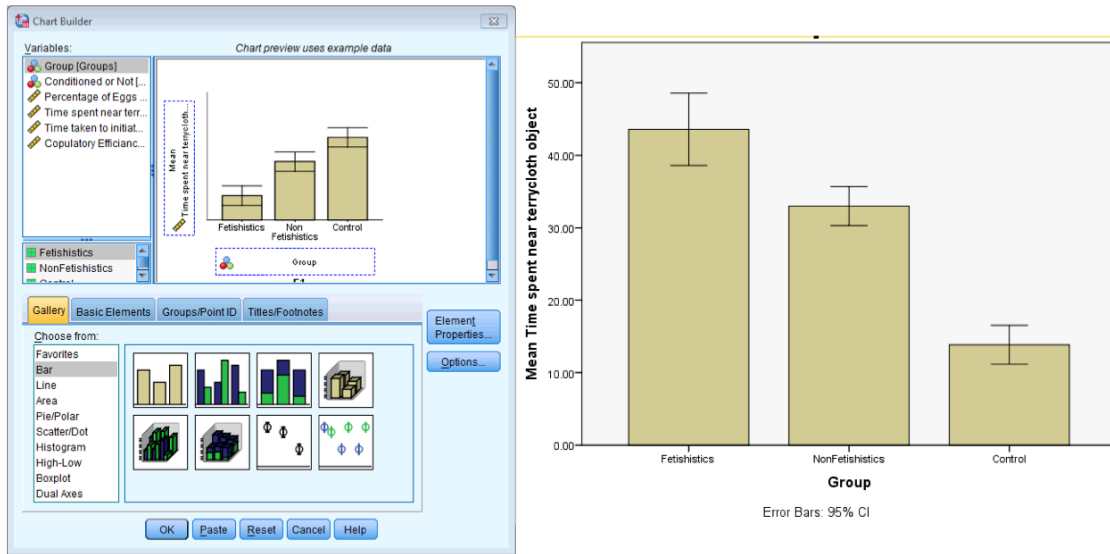
**Output 6**



It's exactly the same! This should, I hope, re-emphasize to you that regression and ANOVA are the same analytic system!

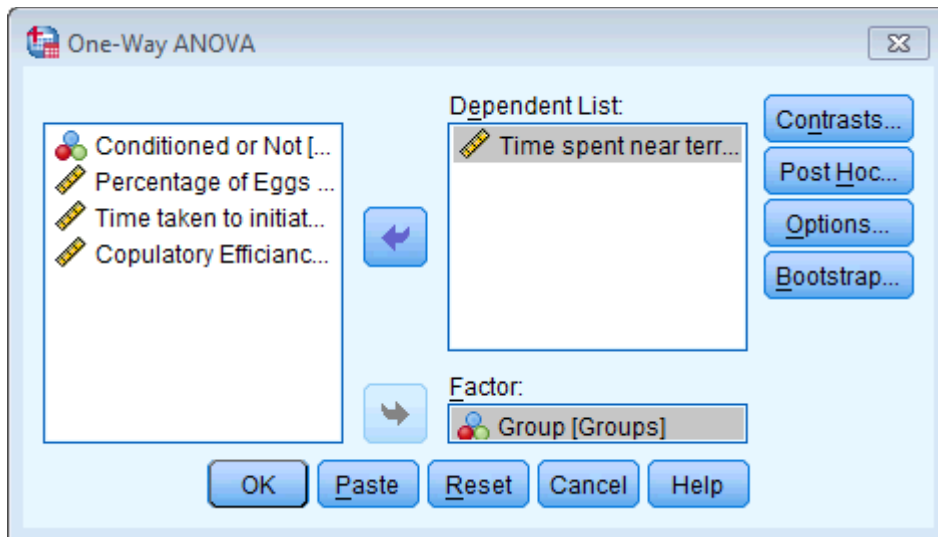
**Task 7**

*Labcoat Leni's Real Research 6.2 describes an experiment (Çetinkaya & Domjan, 2006) on quails with fetishes for terrycloth objects. (Really, it does.) You were asked to analyse two of the variables that they measured with a Kruskal–Wallis test. However, there were two other outcome variables (time spent near the terrycloth object and copulatory efficiency). These data can be analysed with one-way ANOVA. Carry out a one-way ANOVA and Bonferroni post hoc tests on the time spent near the terrycloth object.*

Let's begin by using the Chart Builder ([Graphs](#)  [Chart Builder...](#)) to do an error bar chart:

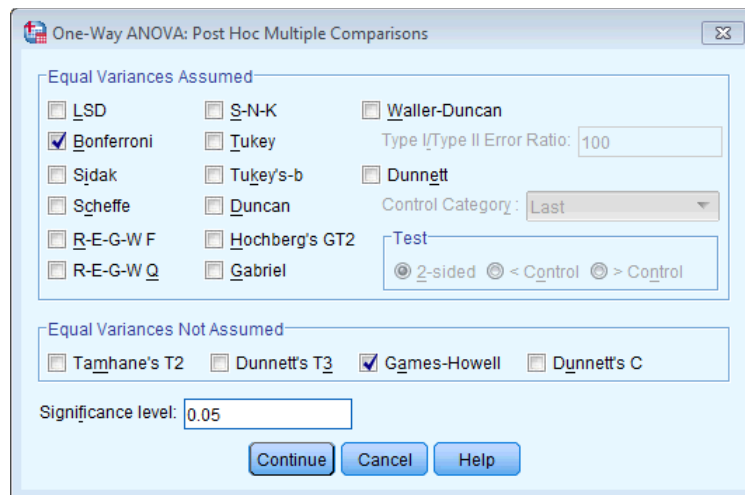


To conduct a one-way ANOVA we have to access the main dialog box by selecting **Analyze** **Compare Means** **One-Way ANOVA...**. This dialog box has a space in which you can list one or more dependent variables and a second space to specify a grouping variable, or *factor*. For these data we need to select **Duration** from the variables list and drag it to the box labelled *Dependent List* (or click on ). Then select the grouping variable **Group** and drag it to the box labelled *Factor* (or click on .

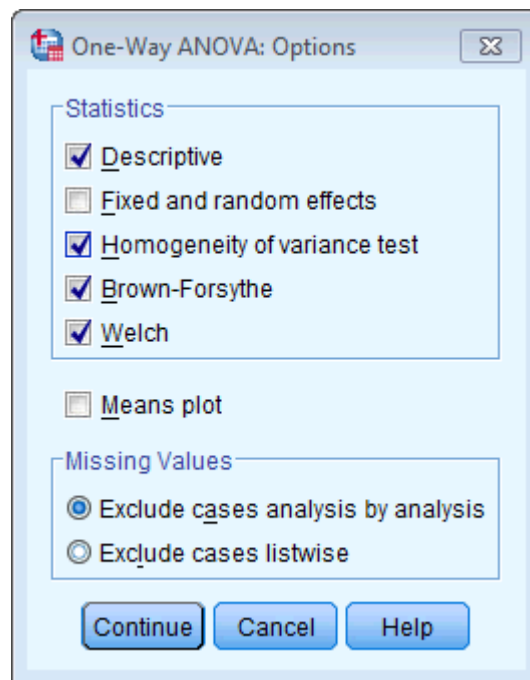




You were asked to do *post hoc* tests so we can skip the contrast options. Click on **Post Hoc...** in the main dialog box to access the *post hoc* tests dialog box. You were asked to do a Bonferroni *post hoc* test so select this, but let's also select Games–Howell in case of problems in homogeneity (which of course we would have checked before running this main analysis!). Click on **Continue** to return to the main dialog box.



Select to test for homogeneity of variance and also to obtain the Brown–Forsythe *F* and Welch *F*. Click on **Continue** to return to the main dialog box, and then click on **OK** to run the analysis.



The output should look like this:

**Descriptives**

Time spent near terrycloth object

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Fetishistics	17	43.5882	9.68929	2.35000	38.6065	48.5700	30.00	66.00
NonFetishistics	15	33.0000	4.86973	1.25736	30.3032	35.6968	26.00	42.00
Control	27	13.8519	6.77497	1.30384	11.1718	16.5319	4.00	25.00
Total	59	27.2881	14.91823	1.94219	23.4004	31.1758	4.00	66.00

Output 7

**Test of Homogeneity of Variances**

Time spent near terrycloth object

Levene Statistic	df1	df2	Sig.
1.891	2	56	.160

Output 24

Output tells us that the homogeneity of variance assumption is met for the *time spent near terrycloth object* outcome variable. This means that we can ignore (just as the authors did) the corrected *F*s and Games–Howell *post hoc* tests. Instead we can look at the normal *F*s and Bonferroni *post hoc* tests (which is what the authors of this paper reported).

#### ANOVA

Time spent near terrycloth object

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9880.577	2	4940.288	91.380	.000
Within Groups	3027.525	56	54.063		
Total	12908.102	58			

#### Output 25

Output tells us that the group (fetishistic, non-fetishistic or control group) had a significant effect on the time spent near the terrycloth object. To find out exactly what's going on we can look at our *post hoc* tests (Output ).

#### Multiple Comparisons

Dependent Variable: Time spent near terrycloth object

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	Fetishistics	NonFetishistics	10.58824 <sup>*</sup>	2.60468	.000	4.1598	17.0167
		Control	29.73638 <sup>*</sup>	2.27651	.000	24.1179	35.3549
	NonFetishistics	Fetishistics	-10.58824 <sup>*</sup>	2.60468	.000	-17.0167	-4.1598
		Control	19.14815 <sup>*</sup>	2.36781	.000	13.3043	24.9920
	Control	Fetishistics	-29.73638 <sup>*</sup>	2.27651	.000	-35.3549	-24.1179
		NonFetishistics	-19.14815 <sup>*</sup>	2.36781	.000	-24.9920	-13.3043
Games-Howell	Fetishistics	NonFetishistics	10.58824 <sup>*</sup>	2.66523	.002	3.9360	17.2404
		Control	29.73638 <sup>*</sup>	2.68747	.000	23.0561	36.4166
	NonFetishistics	Fetishistics	-10.58824 <sup>*</sup>	2.66523	.002	-17.2404	-3.9360
		Control	19.14815 <sup>*</sup>	1.81134	.000	14.7266	23.5697
	Control	Fetishistics	-29.73638 <sup>*</sup>	2.68747	.000	-36.4166	-23.0561
		NonFetishistics	-19.14815 <sup>*</sup>	1.81134	.000	-23.5697	-14.7266

\*. The mean difference is significant at the 0.05 level.

#### Output 26


The authors reported as follows:

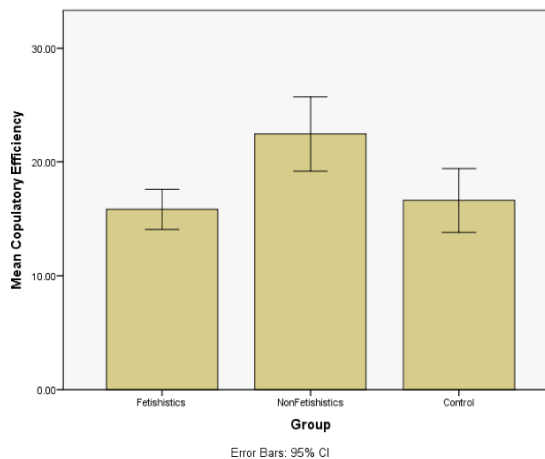
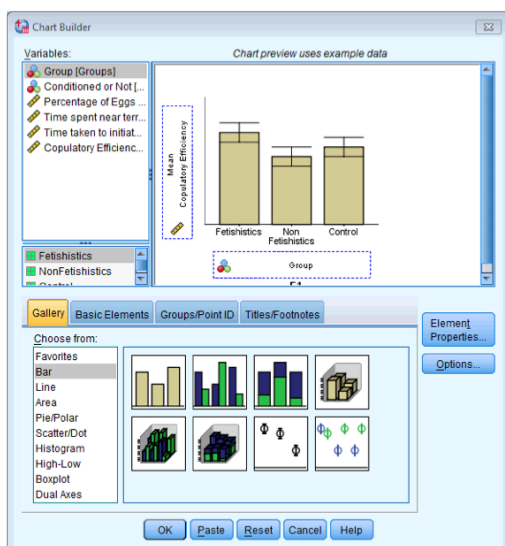
A one-way ANOVA indicated significant group differences,  $F(2, 56) = 91.38, p < .05, \eta^2 = 0.76$ . Subsequent pairwise comparisons (with the Bonferroni correction) revealed that fetishistic male quail stayed near the CS longer than both the nonfetishistic male quail (mean difference = 10.59 s; 95% CI = 4.16, 17.02;  $p < .05$ ) and the control male quail (mean difference = 29.74 s; 95% CI = 24.12, 35.35;  $p < .05$ ). In addition, the nonfetishistic male quail spent more time near the CS than did the control male quail (mean difference = 19.15 s; 95% CI = 13.30, 24.99;  $p < .05$ ). (pp. 429–430)

These results show that male quails do show fetishistic behaviour (the time spent with the terrycloth). Note that the CS is the terrycloth object. Look at the graph, the ANOVA table and the *post hoc* tests to see from where the values that they report come.



## Task 8

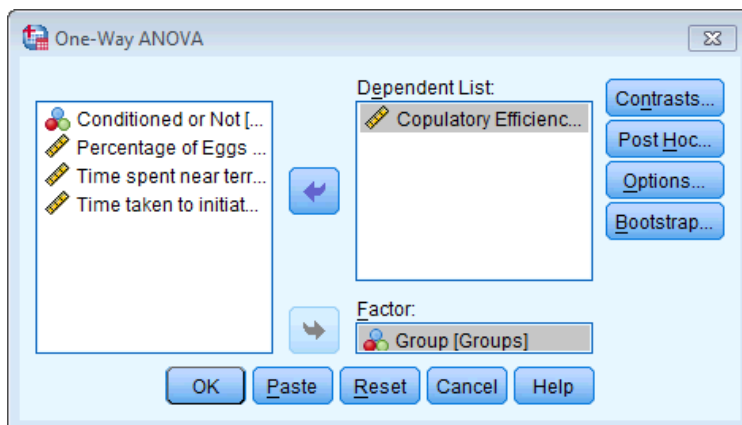
*Repeat the analysis above but using copulatory efficiency as the outcome.*



Let's begin by using the Chart Builder ( [Graphs](#)  [Chart Builder...](#) ) to do an error bar chart:

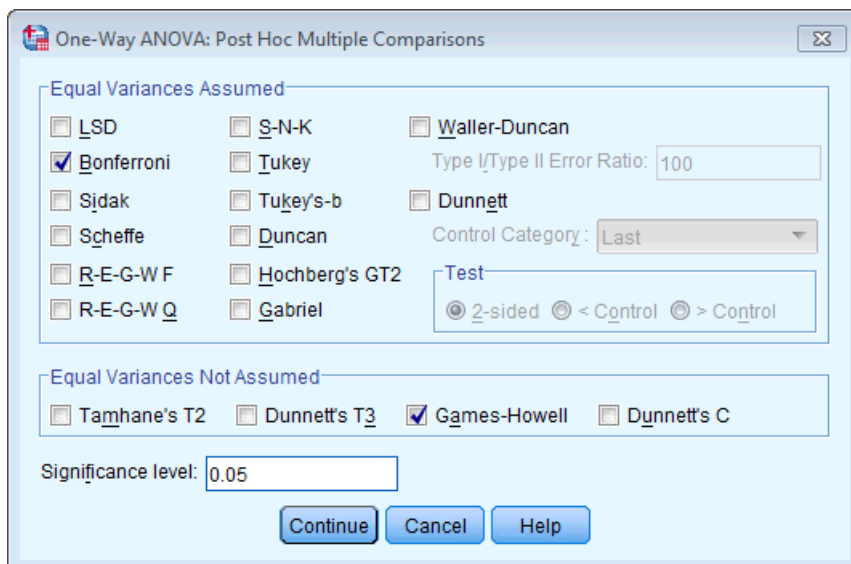


To conduct a one-way ANOVA we have to access the main dialog box by selecting [Analyze](#) [Compare Means](#) [One-Way ANOVA...](#). This dialog box has a space in which you can list one

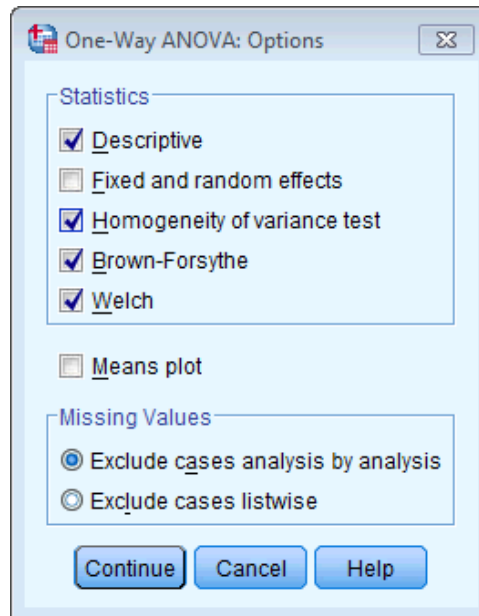
or more dependent variables and a second space to specify a grouping variable, or *factor*. For these data we need to select **Efficiency** from the variables list and drag it to the box labelled *Dependent List* (or click on ). Then select the grouping variable **Group** and drag it to the box labelled *Factor* (or click on ).



You were asked to do *post hoc* tests, so we can skip the contrast options. Click on  in the main dialog box to access the *post hoc* tests dialog box. You were asked to do a Bonferroni *post hoc* test so select this, but let's also select Games–Howell in case of problems in homogeneity (which of course we would have checked before running this main analysis!). Click on  to return to the main dialog box.



Select to test for homogeneity of variance and also to obtain the Brown–Forsythe  $F$  and Welch  $F$ . Click on **Continue** to return to the main dialog box and then click on **OK** to run the analysis.



The output should look like this:

#### Descriptives

Copulatory Efficiency

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Fetishistics	17	15.8447	3.43087	.83211	14.0807	17.6087	11.11	22.22
NonFetishistics	15	22.4593	5.88174	1.51866	19.2021	25.7165	4.35	28.57
Control	27	16.6241	7.09905	1.36621	13.8158	19.4324	.00	33.33
Total	59	17.8831	6.44678	.83930	16.2030	19.5631	.00	33.33

Output 27

**Test of Homogeneity of Variances**

Copulatory Efficiency

Levene Statistic	df1	df2	Sig.
1.498	2	56	.232

**Output 28**

Output tells us that the homogeneity of variance assumption is met for copulatory efficiency. This means that we can ignore (just as the authors did) the corrected *F*s and Games–Howell *post hoc* tests. Instead we can look at the normal *F*s and Bonferroni *post hoc* tests (which is what the authors of this paper reported).

**ANOVA**

Copulatory Efficiency

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	427.563	2	213.782	6.037	.004
Within Groups	1982.970	56	35.410		
Total	2410.533	58			

**Output 29**

Output tells us that the group (fetishistic, non-fetishistic or control group) had a significant effect on copulatory efficiency. To find out exactly what's going on we can look at our *post hoc* tests (Output 30).

## Multiple Comparisons

Dependent Variable: Copulatory Efficiency

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	Fetishistics	NonFetishistics	-6.61463*	2.10799	.008	-11.8172	-1.4121
		Control	-.77937	1.84240	1.000	-5.3265	3.7677
	NonFetishistics	Fetishistics	6.61463*	2.10799	.008	1.4121	11.8172
		Control	5.83526*	1.91629	.011	1.1058	10.5647
	Control	Fetishistics	.77937	1.84240	1.000	-3.7677	5.3265
		NonFetishistics	-5.83526*	1.91629	.011	-10.5647	-1.1058
Games-Howell	Fetishistics	NonFetishistics	-6.61463*	1.73168	.003	-10.9656	-2.2637
		Control	-.77937	1.59967	.878	-4.6731	3.1143
	NonFetishistics	Fetishistics	6.61463*	1.73168	.003	2.2637	10.9656
		Control	5.83526*	2.04276	.019	.8288	10.8417
	Control	Fetishistics	.77937	1.59967	.878	-3.1143	4.6731
		NonFetishistics	-5.83526*	2.04276	.019	-10.8417	-.8288

\*. The mean difference is significant at the 0.05 level.

## Output 8

The authors reported as follows:

A one-way ANOVA yielded a significant main effect of groups,  $F(2, 56) = 6.04$ ,  $p < .05$ ,  $\eta^2 = 0.18$ . Paired comparisons (with the Bonferroni correction) indicated that the nonfetishistic male quail copulated with the live female quail (US) more efficiently than both the fetishistic male quail (mean difference = 6.61; 95% CI = 1.41, 11.82;  $p < .05$ ) and the control male quail (mean difference = 5.83; 95% CI = 1.11, 10.56;  $p < .05$ ). The difference between the efficiency scores of the fetishistic and the control male quail was not significant (mean difference = 0.78; 95% CI = -5.33, 3.77;  $p > .05$ ). (p. 430)

These results show that male quails do show fetishistic behaviour (the time spent with the terrycloth – see Task 7 above) and that this affects their copulatory efficiency (they are less efficient than those that don't develop a fetish, but it's worth remembering that they are no worse than quails that had no sexual conditioning – the controls). If you look at Labcoat Leni's box then you'll also see that this fetishistic behaviour may have evolved because the quails with fetishistic behaviour manage to fertilize a greater percentage of eggs (so their genes are passed on).

## Task 9

A sociologist wanted to compare murder rates (**Murder**) each month in a year at three high profile locations in London (**Street**). Run an ANOVA with bootstrapping on the post hoc tests to see in which streets the most murders happened (**Murder.sav**).



**Descriptives**

Number of Murders

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
					Ruskin Avenue	12		
Acacia Avenue	12	1.25	1.138	.329	.53	1.97	0	3
Rue Morgue	12	2.92	2.275	.657	1.47	4.36	0	6
Total	36	1.67	1.740	.290	1.08	2.26	0	6

**Output 91**

Looking at the means in the Descriptives table (Output 9), we can see that Rue Morgue had the highest mean number of murders ( $M = 2.92$ ) and Ruskin Avenue had the smallest mean number of murders ( $M = 0.83$ ). These means will be important in interpreting the *post hoc* tests later.

**Test of Homogeneity of Variances**

Number of Murders

Levene Statistic	df1	df2	Sig.
11.382	2	33	.000

**Output 10**

Output displays the results of Levene's test. For these data, the assumption of homogeneity of variance has been violated, because our significance is .000, which is considerably smaller than the criterion of .05. In these situations, we have to try to correct the problem, and we can either transform the data or choose the Welch  $F$ .

**ANOVA**

Number of Murders

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	29.167	2	14.583	6.264	.005
Within Groups	76.833	33	2.328		
Total	106.000	35			

**Output 11**

The main ANOVA summary table (Output ) shows us that because the observed significance value is less than .05 we can say that there was a significant effect of street on the number of murders. However, at this stage we still do not know exactly which streets had significantly more murders (we don't know which groups differed).

**Robust Tests of Equality of Means**

Number of Murders

	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	4.595	2	19.285	.023
Brown-Forsythe	6.264	2	18.689	.008

a. Asymptotically F distributed.

**Output 34**

Output shows the Welch and Brown–Forsythe *F*s, which are useful because homogeneity of variance was violated. Luckily our conclusions remain the same: both *F*s have significance values less than .05.

**Multiple Comparisons**

Dependent Variable: Number of Murders  
Games-Howell

(I) Street	(J) Street	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Ruskin Avenue	Acacia Avenue	-.417	.388	.542	-1.41	.57
	Rue Morgue	-2.083*	.689	.024	-3.90	-.27
Acacia Avenue	Ruskin Avenue	.417	.388	.542	-.57	1.41
	Rue Morgue	-1.667	.734	.089	-3.56	.23
Rue Morgue	Ruskin Avenue	2.083*	.689	.024	.27	3.90
	Acacia Avenue	1.667	.734	.089	-.23	3.56

\*. The mean difference is significant at the 0.05 level.

**Output 35**

Because there were no specific hypotheses I just carried out *post hoc* tests and stuck to my favourite Games–Howell procedure (because variances were unequal). It is clear from Output that each street is compared to all of the remaining streets. If we look at the values in the column labelled *Sig.* we can see that the only significant comparison was between Ruskin Avenue and Rue Morgue ( $p = .024$ , which is less than .05); all other comparisons were non-significant because all the other values in this column are greater than .05. However, Acacia Avenue and Rue Morgue were close to being significantly different ( $p = .089$ ).

**Bootstrap for Multiple Comparisons**

Dependent Variable: Number of Murders  
Games-Howell

(I) Street	(J) Street	Mean Difference (I-J)	Bootstrap <sup>a</sup>			
			Bias	Std. Error	BCa 95% Confidence Interval	
					Lower	Upper
Ruskin Avenue	Acacia Avenue	-.417	-.021	.383	-1.115	.254
	Rue Morgue	-2.083	-.013	.677	-3.506	-.604
Acacia Avenue	Ruskin Avenue	.417	.021	.383	-.384	1.242
	Rue Morgue	-1.667	.008	.740	-3.164	-.151
Rue Morgue	Ruskin Avenue	2.083	.013	.677	.755	3.422
	Acacia Avenue	1.667	-.008	.740	.171	3.129

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Output 36**

The question asked us to bootstrap the *post hoc* tests and this has been done in Output . The columns of interest are the ones containing the BCa 95% confidence intervals (lower and upper limits). We can see that the difference between Ruskin Avenue and Rue Morgue

remains significant after bootstrapping the confidence intervals; we can tell this because the confidence intervals do not cross zero for this comparison. Surprisingly, it appears that the difference between Acacia Avenue and Rue Morgue is now significant after bootstrapping the confidence intervals, because again the confidence intervals do not cross zero. This seems to contradict the  $p$ -values in the previous output; however, the  $p$ -value was close to being significant ( $p = .089$ ). The mean values in the table of descriptives tell us that Rue Morgue had a significantly higher number of murders than Ruskin Avenue and Acacia Avenue; however, Acacia Avenue did not differ significantly in the number of murders compared to Ruskin Avenue.

### Calculating the effect size

Output provides us with three measures of variance: the between-group effect ( $SS_M$ ), the within-subject effect ( $MS_R$ ) and the total amount of variance in the data ( $SS_T$ ). We can use these to calculate omega squared ( $\omega^2$ ):

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$$

Substituting from Output 33:

$$\begin{aligned}\omega^2 &= \frac{29.167 - (2)2.328}{106.00 + 2.328} \\ &= \frac{24.511}{108.328} \\ &= .23 \\ \omega &= .48\end{aligned}$$

### Interpreting and writing the result

We could report the main finding as:

- ✓ Levene's test indicated that the assumption of homogeneity of variance had been violated,  $F(2, 33) = 11.38, p < .001$ , so Welch's  $F$  is reported. The results show that the streets measured differed significantly in the number of murders,  $F(2, 19.29) = 4.60, p < .05, \omega^2 = .23$ .

The next thing that needs to be reported are the *post hoc* comparisons:

- ✓ Games–Howell *post hoc* tests with 95% bias corrected confidence intervals on the mean differences revealed that Rue Morgue experienced a significantly greater number of murders than either Ruskin Avenue, 95% BCa CI [ 0.76, 3.42] or Acacia Avenue, 95% BCa CI [0.17, 3.13]. However, Acacia Avenue and Ruskin Avenue did not

differ significantly in the number of murders that had occurred, 95% BCa CI [ -0.38, 1.24].