# 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.



Output 1

Output 1 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.



Output 2

The next part of the output (Output 2) 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.



Output 3

Output 3 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).



Output 4

Output 4 shows the Welch and Brown–Forsythe *F*s, but we can ignore these because the homogeneity of variance assumption was met.



Output 5

Because there were specific hypotheses I specified some contrasts. Output 5 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.



Output 6

Output 6 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 3 provides us with three measures of variance: the between-group effect (SSM), the within-subject effect (MSR) and the total amount of variance in the data (SST). We can use these to calculate omega squared (*ω*2):

Substituting from Output 3:

For the contrasts the effect sizes will be:



For contrast 1 we get:



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:



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

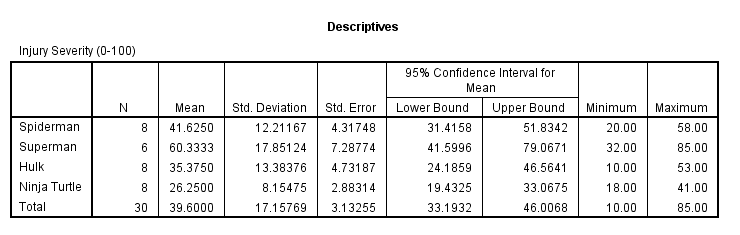
#### Interpreting and writing the result

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, *ω*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.

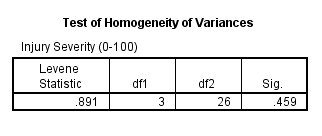
### Task 3

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](#_ENREF_53)). 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.



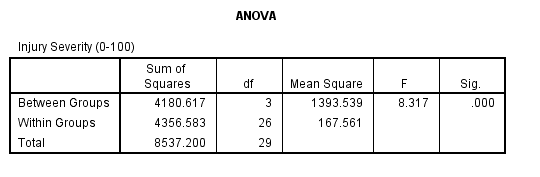
Output 7

Looking at the means in Output 7, 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).



Output 8

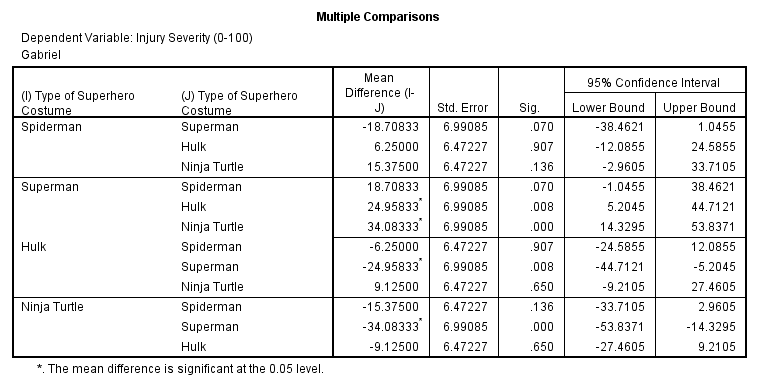
Looking at Output 8, we can see that the assumption of homogeneity of variance has been met because our significance value is .46, which is much bigger than the criterion of .05.



Output 9

In the ANOVA output (Output 9), 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).



Output 10

Output 10 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 9 provides us with three measures of variance: the between-group effect (SSM), the within-subject effect (MSR) and the total amount of variance in the data (SST). We can use these to calculate omega squared (*ω*2):

Substituting from Output 9:

#### 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, *ω*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?



Output 11

Output 11 shows the table of descriptive statistics from the one-way ANOVA. It looks as though as soya intake increases, sperm counts do indeed decrease.



Output 12

The next part of the output (Output 12) 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).



Output 13

Output 13 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!



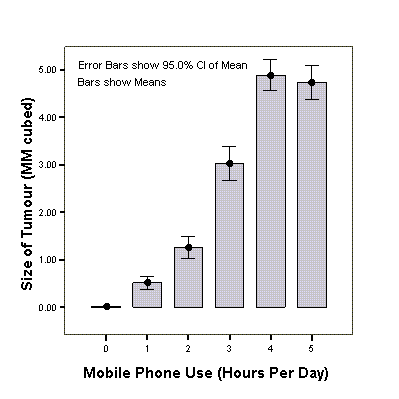
Output 14

Output 14 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 mm3) 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.



Output 15

Output 15 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.



Output 16

Output 16 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*.



Output 17

Output 17 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).



Output 18

Output 18 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.



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 19 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 17 provides us with three measures of variance: the between-group effect (SSM), the within-subject effect (MSR) and the total amount of variance in the data (SST). We can use these to calculate omega squared (*ω*2):

Substituting from Output 17:

#### 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, *ω*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:

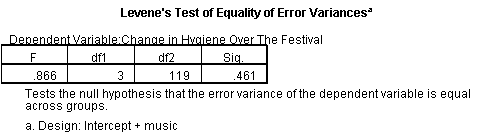
* Games–Howell *post hoc* tests revealed significant differences between all groups (*p* < .001 for all tests) except between 4 and 5 hours (*ns*).

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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 95% Confidence Interval | |
| Mobile Phone Use (Hours Per Day) | | Sig. | 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.



Output 20

Looking at Output 20, 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.



Output 21

Output 21 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 22), in which we analysed these data as a regression:



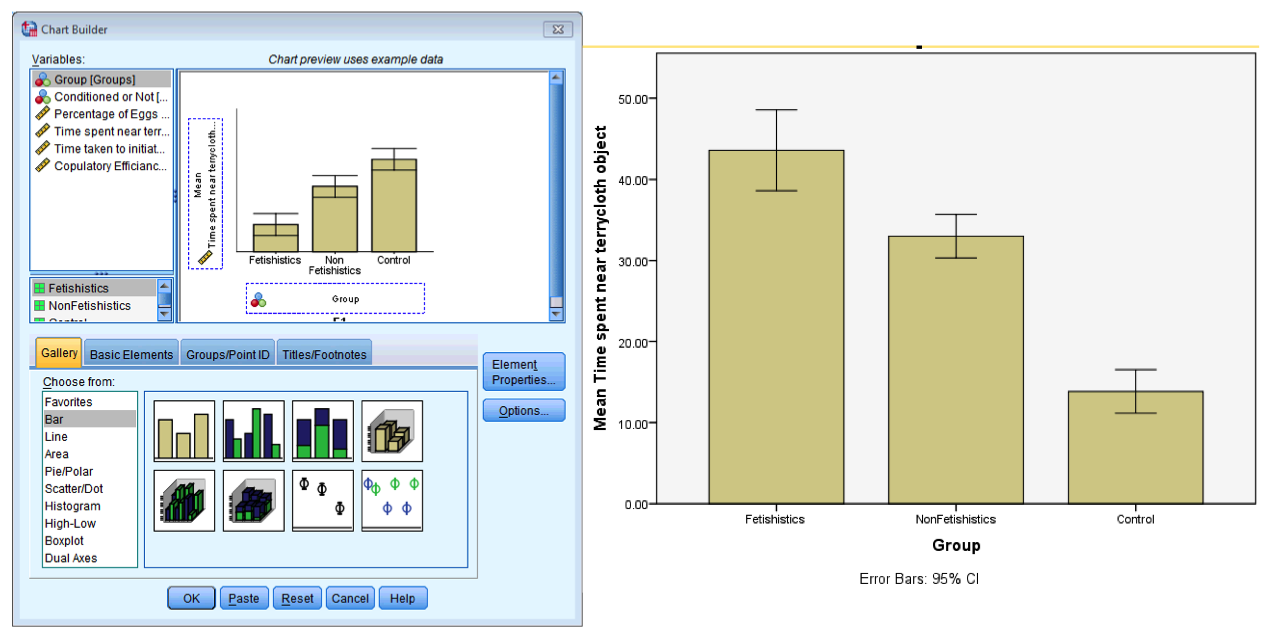
Output 22

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

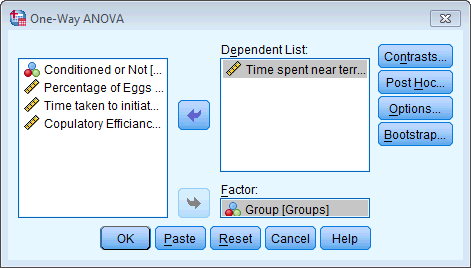
### Task 7

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.

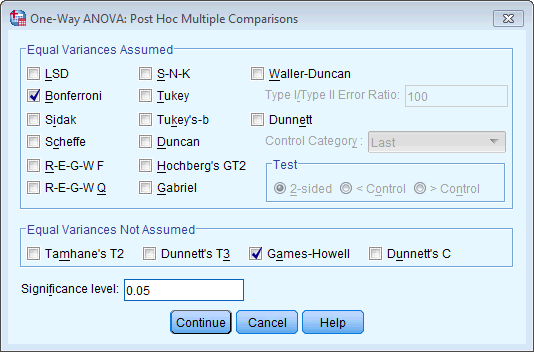
Let’s begin by using the Chart Builder ( ) to do an error bar chart:



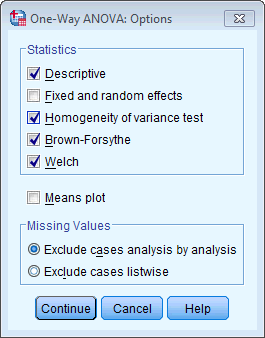
To conduct a one-way ANOVA we have to access the main dialog box by selecting Macintosh HD:Users:Andy:Dropbox:Zoe:DSUS:DSUS 4:Button Images:Analyse.png. This dialog box has a space in which you can list one or more dependent variables and a second space to specify a grouping variable, or *factor*. For 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 ).



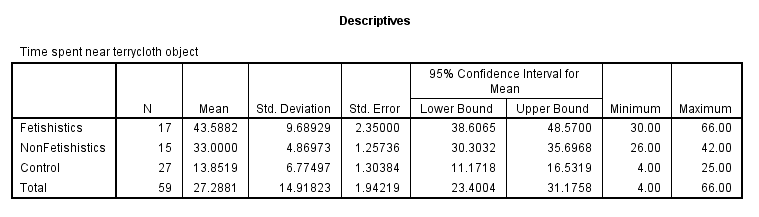
You were asked to do *post hoc* tests so we can skip the contrast options. Click on  in the main dialog box to access the *post hoc* tests dialog box. You were asked to do a Bonferroni *post hoc* test so select this, but let’s also select Games–Howell in case of problems in homogeneity (which of course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



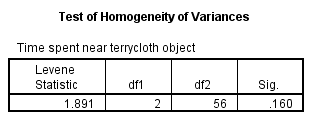
Select to test for homogeneity of variance and also to obtain the Brown–Forsythe *F* and Welch *F*. Click on  to return to the main dialog box, and then click on Macintosh HD:Users:Andy:Dropbox:Zoe:DSUS:DSUS 4:Button Images:ok.png to run the analysis.



The output should look like this:

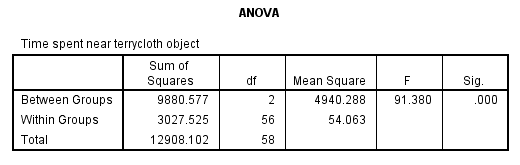


Output 23



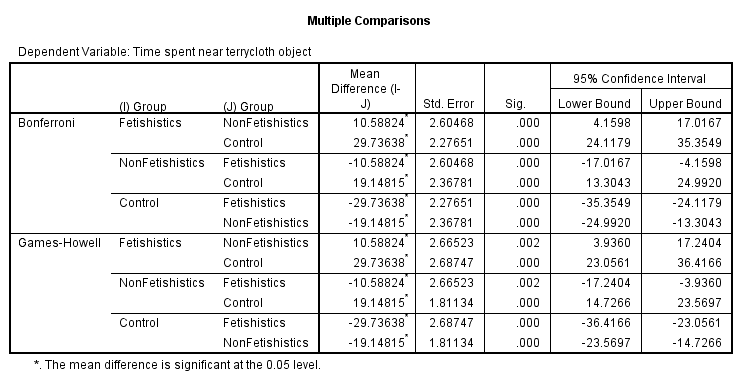
Output 24

Output 24 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).



Output 25

Output 25 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 26).



Output 26

The authors reported as follows:

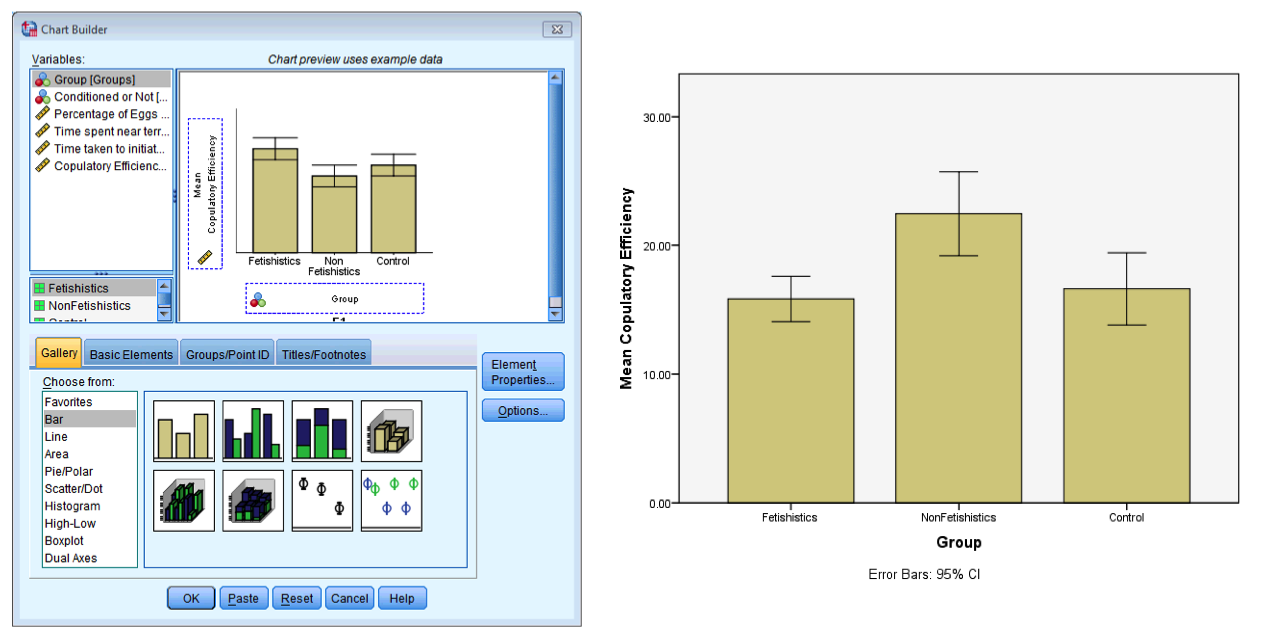
A one-way ANOVA indicated significant group differences, *F*(2, 56) = 91.38, *p* < .05, = 0.76. Subsequent pairwise comparisons (with the Bonferroni correction) revealed that fetishistic male quail stayed near the CS longer than both the nonfetishistic male quail (mean difference = 10.59 s; 95% CI = 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.

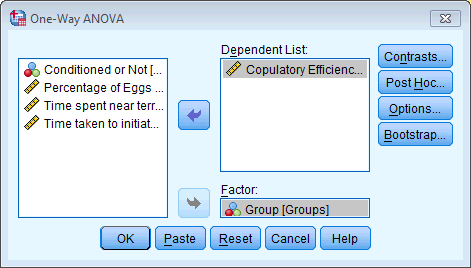
### Task 8

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

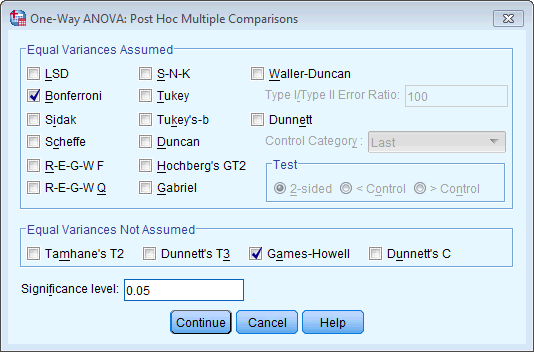
Let’s begin by using the Chart Builder ( ) to do an error bar chart:



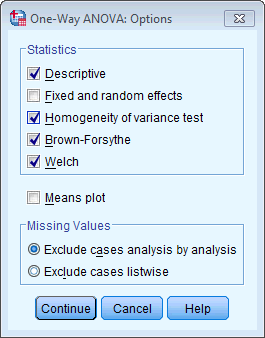
To conduct a one-way ANOVA we have to access the main dialog box by selecting Macintosh HD:Users:Andy:Dropbox:Zoe:DSUS:DSUS 4:Button Images:Analyse.png. This dialog box has a space in which you can list one or more dependent variables and a second space to specify a grouping variable, or *factor*. For 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 ).



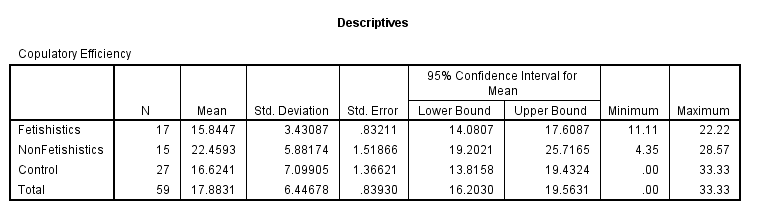
You were asked to do *post hoc* tests, so we can skip the contrast options. Click on  in the main dialog box to access the *post hoc* tests dialog box. You were asked to do a Bonferroni *post hoc* test so select this, but let’s also select Games–Howell in case of problems in homogeneity (which of course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



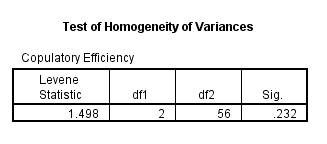
Select to test for homogeneity of variance and also to obtain the Brown–Forsythe *F* and Welch *F*. Click on  to return to the main dialog box and then click on Macintosh HD:Users:Andy:Dropbox:Zoe:DSUS:DSUS 4:Button Images:ok.png to run the analysis.



The output should look like this:

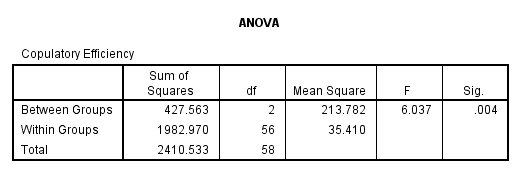


Output 27



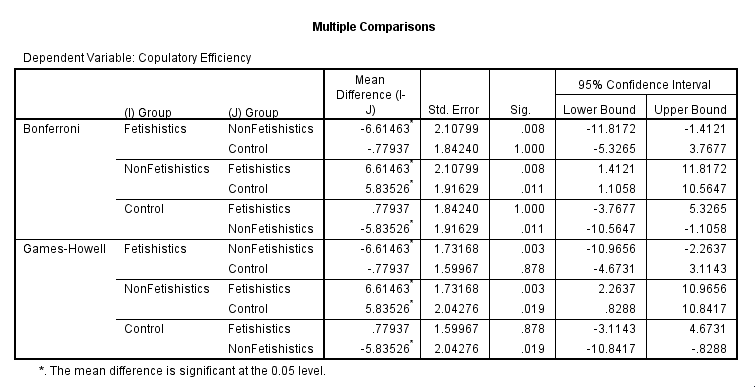
Output 28

Output 28 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).



Output 29

Output 29 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).



Output 30

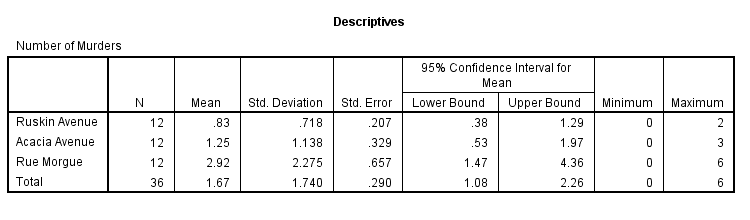
The authors reported as follows:

A one-way ANOVA yielded a significant main effect of groups, *F*(2, 56) = 6.04, *p* < .05, = 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).

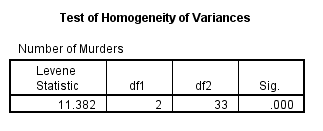
### Task 9

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**).



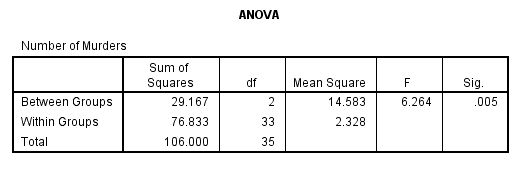
Output 31

Looking at the means in the Descriptives table (Output 31), 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.



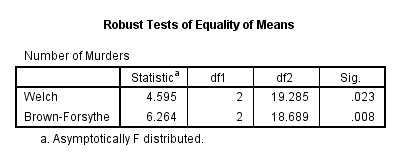
Output 32

Output 32 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*.



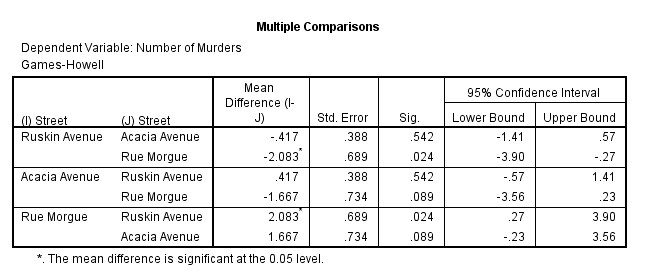
Output 33

The main ANOVA summary table (Output 33) 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).



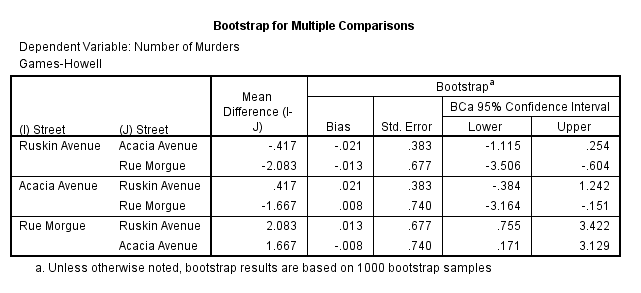
Output 34

Output 34 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.



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 35 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).



Output 36

The question asked us to bootstrap the *post hoc* tests and this has been done in Output 36. 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 33 provides us with three measures of variance: the between-group effect (SSM), the within-subject effect (MSR) and the total amount of variance in the data (SST). We can use these to calculate omega squared (*ω*2):

Substituting from Output 33:

#### 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, *ω*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].