Chapter 8: Answers

Task 1

Imagine that I was interested in how different teaching methods affected students' knowledge. I noticed that some lecturers were aloof and arrogant in their teaching style and humiliated anyone who asked them a question, while others were encouraging and supporting of questions and comments. I took three statistics courses where I taught the same material. For one group of students I wandered around with a large cane and beat anyone who asked daft questions or got questions wrong (punish). In the second group I used my normal teaching style which is to encourage students to discuss things that they find difficult and to give anyone working hard a nice sweet (reward). The final group I remained indifferent to and neither punished nor rewarded their efforts (indifferent). As the dependent measure I took the students' exam marks (percentage). Based on theories of operant conditioning, we expect punishment to be a very unsuccessful way of reinforcing learning, but we expect reward to be very successful. Therefore, one prediction is that reward will produce the best learning. A second hypothesis is that punishment should actually retard learning such that it is worse than an indifferent approach to learning. 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.

SPSS Output

Descriptives

Exam Mark

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Punish	10	50.0000	4.13656	1.30809	47.0409	52.9591	45.00	57.00
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

This 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

The next part of the 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 0.095, which is bigger than the criterion of 0.05.

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			

The main ANOVA summary table shows us that because the observed significance value is less than 0.05 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[a]	df1	df2	Sig.
Welch	32.235	2	17.336	.000
Brown-Forsythe	21.008	2	20.959	.000

a. Asymptotically F distributed.

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

Contrast Coefficients

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

Because there were specific hypotheses I specified some contrasts. This table 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 Tests

		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

This table 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 0.05). 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 (and the descriptive statistics) tells us that the marks after punishment were significantly lower than after indifference (again, it's significantly different because the value in the column labelled *Sig.* is less than 0.05). 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!

Calculating the Effect Size

The output provides us with three measures of variance: the between group effect (SS_M), the within subject effect (SS_R) and the total amount of variance in the data (SS_T). We can use these to calculate omega squared (ω^2):

$$\omega^2 = \frac{MS_M - MS_R}{MS_M + ((n-1) \times MS_R)}$$

$$\omega^2 = \frac{602.53 - 28.68}{602.53 + ((10-1) \times 28.68)}$$

$$= \frac{573.85}{602.53 + 258.12}$$

$$= 0.67$$

$$\omega = \sqrt{0.67} = 0.82$$

For the contrasts the effect sizes will be:

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

$$r_{contrast1} = \sqrt{\frac{-5.978^2}{-5.978^2 + 27}}$$

$$= 0.75$$

If you think back to our benchmarks for effect sizes this represents a huge effect (it is well above 0.5—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:

$$r_{contrast2} = \sqrt{\frac{-2.505^2}{-2.505^2 + 27}}$$

$$= 0.43$$

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, $F(2, 27) = 21.01$, $\omega = .82$. 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 2

In Chapter 11 (section 11.4) 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 11.4? Why do you think this difference has arisen?

SPSS Output

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 Soyal 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

This 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

The next part of the 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 0.003, which is smaller than the criterion of 0.05. In fact, these data also violate the assumption of normality (see the Chapter on nonparametric 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			

The main ANOVA summary table shows us that because the observed significance value is greater than 0.05 we can say that there was no significant effect of Soya intake on men's sperm count. This is strange because if you read the chapter on nonparametric statistics 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 the chapter on nonparametric statistics, although parametric tests have more power to detect effects when their assumptions are met, when their assumptions are violated nonparametric tests have more power! This example was arranged to prove this point: because the parametric assumptions are violated, the nonparametric tests produced a significant result and the parametric test did not because, in these circumstances, the nonparametric test has the greater power!

Robust Tests of Equality of Means

Sperm Count (Millions)

	Statistic[a]	df1	df2	Sig.
Welch	6.284	3	34.657	.002
Brown-Forsythe	2.636	3	58.236	.058

a. Asymptotically F distributed.

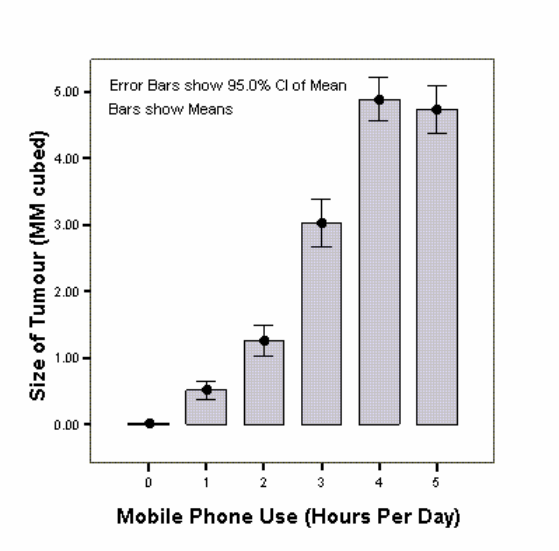
This table shows the Welch and Brown-Forsythe *F*s, note that the Welch test agrees with the nonparametric test in that the significance of F is below the 0.05 threshold. However, the Brown-Forsythe F is non-significant (it is just above the threshold). This illustrates the relative superiority of the Welch procedure. However, in these circumstances because normality *and* homogeneity of variance have been violated we'd use a nonparametric test anyway!

Task Three

Students (and lecturers for that matter) love their mobile phones, which is rather worrying given some recent controversy about links between mobile phone use and brain tumours. The basic idea is that 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 selecting the 'cook until well done' button. If we wanted to test this experimentally, we could get 6 groups of people and strap a mobile phone on their heads (that they can't remove). Then, by remote control, we turn the phones on for a certain amount of time each day. After 6 months, we measure the size of any tumour (in mm3) close to the site of the phone antennae (just behind the ear). The six groups experienced 0, 1, 2, 3, 4 or 5 hours per day of phone microwaves for 6 months. The data are in **Tumour.sav.** (From Field & Hole, 2003, so there is a very detailed answer in there).

SPSS Output

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 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

This 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

The next part of the 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 0.000, which is considerably smaller than the criterion of 0.05. In these situations, we have to try to correct the problem and we can either transform the data or choose the Welch *F*.

ANOVA

Size of Tumour (MM cubed)

	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			

The main ANOVA summary table shows us that because the observed significance value is less than 0.05 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

Size of Tumour (MM cubed)

	Statistic[a]	df1	df2	Sig.
Welch	414.926	5	44.390	.000
Brown-Forsythe	269.733	5	75.104	.000

a. Asymptotically F distributed.

This table 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 0.05.

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.

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 the table that each group of participants is compared to all of the remaining groups. First, the control group (0 hours) is compared to the 1-hour, 2-hour, 3-hour, 4-hour and 5-hour groups and reveals a significant difference in all cases (all the values in the column labeled Sig. are less than 0.05). 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 labeled Sig. are less than 0.05). In fact, all of the comparisons appear to be highly significant except the comparison between the 4-hour and 5-hour groups, which is non-significant because the value in the column labeled Sig. Is bigger than 0.05.

Calculating the Effect Size

The output provides us with three measures of variance: the between group effect (SS_M), the within subject effect (SS_R) and the total amount of variance in the data (SS_T). We can use these to calculate omega squared (ω^2):

$$\omega^2 = \frac{MS_M - MS_R}{MS_M + ((n-1) \times MS_R)}$$

$$\omega^2 = \frac{90.13 - 0.33}{90.13 + ((20-1) \times 0.33)}$$

$$= \frac{89.8}{90.13 + 6.27}$$

$$= 0.93$$

$$\omega = \sqrt{0.93} = 0.96$$

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 (5, 114) = 10.25, p < .001). Transforming the data did not rectify this problem and so F-tests are reported nevertheless. The results show that using a mobile phone significantly affected the size of brain tumour found in participants (F(5, 114) = 269.73, p < .001, r = .96). 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 summarise 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 summarise these confidence intervals as a table:

Mobile Phone Use (Hours Per Day)		Sig.	95% Confidence Interval	
			Lower Bound	Upper Bound
0	1	< .001	-.6982	-.2964
	2	< .001	-1.5916	-.8960
	3	< .001	-3.5450	-2.4631
	4	< .001	-5.3622	-4.3783
	5	< .001	-5.2653	-4.1608
1	2	< .001	-1.1327	-.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	-.9751
4	5	= .984	-.5455	.8599