Repeated-measures designs

Self-test answers

✓ What is a repeated-measures design?

Repeated-measures is a term used when the same participants participate in all conditions of an experiment.

- Use *ggplot2* to plot a bar graph (with error bars) of the time to retch with the type of animal eaten on the *x*-axis.

```
bushBar <- ggplot(longBush, aes(Animal, Retch))
bushBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black") +
stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Type of Animal Eaten", y
= "Mean Time to Retch (Seconds)")
```

- Use *ggplot2* to plot boxplots of the time to retch after eating each animal (*x*-axis).

```
bushBoxplot <- ggplot(longBush, aes(Animal, Retch))
bushBoxplot + geom_boxplot() + labs(x = "Type of Animal Eaten", y = "Mean Time to Retch
(Seconds)")
```

✓ Using what you learnt earlier in the chapter and the commands that we have just used to create **drink** and **imagery**, can you work out how to enter the data into **R** directly?

If we wanted to enter the data directly into **R**, we would first need to create the variable that identifies participants by using the *gl()* function (Chapter 3). Remember that this function takes the general form:

```
factor<-gl(number of levels, cases in each level, total cases, labels = c("label1", "label2"…))
```

This function creates a factor variable called *factor*; you specify the number of levels or groups of the factor, how many cases are in each level/group, optionally the total number of cases (the default is to multiply the number of groups by the number of cases per group), and you can also use the *labels* option to list names for each level/group.

For **participant**, we want nine scores for each of the 20 participants, so we can specify it as:

```
participant<-gl(20, 9, labels = c("P01", "P02", "P03", "P04", "P05", "P06", "P07", "P08", "P09",
"P10", "P11", "P12", "P13", "P14", "P15", "P16", "P17", "P18", "P19", "P20" ))
```

The numbers in the function tell **R** that we had 20 sets of nine scores, the labels option then specifies the names to attach to these 20 sets, which correspond to their participant number. A quicker way to do this is to use the *paste()* function to create the labels for you. We can execute this command instead:

```
participant<-gl(20, 9, labels = c(paste("P", 1:20, sep = "_")))
```

The *paste()* function takes the things in brackets and pastes them together, the *sep* option specifying how to separate the bits that have been pasted together. So the *"P"* means that we begin with the letter P and then we paste a number after it separated by an underscore. The *1:20* creates a sequence of numbers from 1 to 20. Therefore, we create a sequence of text strings that are P then an underscore then a number, where the number starts at 1 and goes to 20. Therefore, we'll get a sequence of strings P_1, P_2, P_3, …, P_20. To see for yourself, just execute the *paste()* function as we have specified it above:

```
paste("P", 1:20, sep = "_")
```

The resulting sequence is:

```
"P_1"  "P_2"  "P_3"  "P_4"  "P_5"  "P_6"  "P_7"  "P_8"  "P_9"  "P_10" "P_11" "P_12" "P_13"
"P_14" "P_15" "P_16" "P_17" "P_18" "P_19" "P_20"
```

Therefore, by placing this paste command within the *gl()* function we automatically generate the labels for each person, which when you have a lot of participants is quicker then typing them all in.

To create the **drink** variable we follow the same procedure as in the chapter. We currently have nine rows per person that we need to identify based on levels of **drink** and **imagery**. Within each person, for each of the three types of drink (beer, wine and water) there are three scores (positive imagery, negative imagery and neutral imagery). Therefore, we want three groups that each contains three scores. This will create the codes within a person, and we need these codes to be repeated for all 20 cases, and to do this we include the total number of cases (20 cases $\times$ 9 scores per case = 180 scores). Including this information in the *gl()* function we would execute:

```
drink<-gl(3, 3, 180, labels = c("Beer", "Wine", "Water"))
```

This creates a variable **drink**; the numbers in the function tell **R** that we had three sets of three scores, the labels option then specifies the names to attach to these three sets, which correspond to the type of drink. The *180* tells **R** to repeat this sequence for 180 cases. Essentially, this will create three rows with the label *Beer* then three labelled *Wine*, then three labelled *Water*, and then repeats this sequence for 180 cases.

We also need a variable that tells us the type of imagery that was used. To do this we want three sets of one score (positive, negative, neutral). This will create three cases, or, put another way, it will create the codes for the first level (beer) of the **drink** variable. We want this pattern to be repeated for the remaining two levels of **drink** (i.e., wine and water). We can do this by adding a third value to the function that is the total number of cases (i.e., 180). By specifying the total number of cases the *gl()* function will repeat the pattern of codes until it reaches this total number of cases

```
imagery<-gl(3, 1, 180, labels = c("Positive", "Negative", "Neutral"))
```

If the interaction turns out to be significant and we want *post hoc* tests for this interaction, then it's necessary to have a variable that codes combinations of drink and imagery into a single variable:

```
groups<-gl(9, 1, labels = c( "beerpos", "beerneg", "beerneut", "winepos", "wineneg", "wineneut",
"waterpos", "waterneg", "waterneut"))
```

This command creates nine sets of one row and then labels them according to the nine combinations of the drink and imagery variables.

We can add the attitude scores by creating a numeric variable in the usual way:

```
attitude<-c(1, 6, 5, 38, -5, 4, 10, -14, -2, 26, 27, 27, 23, -15, 14, 21, -6, 0, 1, -19, -10,
28, -13, 13, 33, -2, 9, 7, -18, 6, 26, -16, 19, 23, -17, 5, 22, -8, 4, 34, -23, 14, 21, -19, 0,
30, -6, 3, 32, -22, 21, 17, -11, 4, 40, -6, 0, 24, -9, 19, 15, -10, 2, 15, -9, 4, 29, -18, 7,
13, -17, 8, 20, -17, 9, 30, -17, 12, 16, -4, 10, 9, -12, -5, 24, -15, 18, 17, -4, 8, 14, -11, 7,
34, -14, 20, 19, -1, 12, 43, 30, 8, 20, -12, 4, 9, -10, -13, 15, -6, 13, 23, -15, 15, 29, -1,
10, 15, 15, 12, 20, -15, 6, 6, -16, 1, 40, 30, 19, 28, -4, 0, 20, -10, 2, 8, 12, 8, 11, -2, 6,
27, 5, -5, 17, 17, 15, 17, -6, 6, 9, -6, -13, 30, 21, 21, 15, -2, 16, 19, -20, 3, 34, 23, 28,
27, -7, 7, 12, -12, 2, 34, 20, 26, 24, -10, 12, 12, -9, 4)
```

Finally, we can merge these variables into a dataframe called *longAttitude* by executing:

```
longAttitude<-data.frame(participant, drink, imagery, groups, attitude)
```

The data should look like this:

```
    participant drink  imagery    groups attitude
1           P01  Beer Positive   beerpos        1
```

2	P01	Beer	Negative	beerneg	6
3	P01	Beer	Neutral	beerneut	5
4	P01	Wine	Positive	winepos	38
5	P01	Wine	Negative	wineneg	-5
6	P01	Wine	Neutral	wineneut	4
7	P01	Water	Positive	waterpos	10
8	P01	Water	Negative	waterneg	-14
9	P01	Water	Neutral	waterneut	-2
10	P02	Beer	Positive	beerpos	26
11	P02	Beer	Negative	beerneg	27
12	P02	Beer	Neutral	beerneut	27
13	P02	Wine	Positive	winepos	23
14	P02	Wine	Negative	wineneg	-15
15	P02	Wine	Neutral	wineneut	14
16	P02	Water	Positive	waterpos	21
17	P02	Water	Negative	waterneg	-6
18	P02	Water	Neutral	waterneut	0
19	P03	Beer	Positive	beerpos	1
20	P03	Beer	Negative	beerneg	-19
21	P03	Beer	Neutral	beerneut	-10
22	P03	Wine	Positive	winepos	28
23	P03	Wine	Negative	wineneg	-13
24	P03	Wine	Neutral	wineneut	13
25	P03	Water	Positive	waterpos	33
26	P03	Water	Negative	waterneg	-2
27	P03	Water	Neutral	waterneut	9
28	P04	Beer	Positive	beerpos	7
29	P04	Beer	Negative	beerneg	-18
30	P04	Beer	Neutral	beerneut	6
31	P04	Wine	Positive	winepos	26
32	P04	Wine	Negative	wineneg	-16
33	P04	Wine	Neutral	wineneut	19
34	P04	Water	Positive	waterpos	23
35	P04	Water	Negative	waterneg	-17
36	P04	Water	Neutral	waterneut	5
37	P05	Beer	Positive	beerpos	22
38	P05	Beer	Negative	beerneg	-8
39	P05	Beer	Neutral	beerneut	4
40	P05	Wine	Positive	winepos	34
41	P05	Wine	Negative	wineneg	-23
42	P05	Wine	Neutral	wineneut	14
43	P05	Water	Positive	waterpos	21
44	P05	Water	Negative	waterneg	-19
45	P05	Water	Neutral	waterneut	0
46	P06	Beer	Positive	beerpos	30
47	P06	Beer	Negative	beerneg	-6
48	P06	Beer	Neutral	beerneut	3
49	P06	Wine	Positive	winepos	32
50	P06	Wine	Negative	wineneg	-22
51	P06	Wine	Neutral	wineneut	21
52	P06	Water	Positive	waterpos	17
53	P06	Water	Negative	waterneg	-11
54	P06	Water	Neutral	waterneut	4
55	P07	Beer	Positive	beerpos	40
56	P07	Beer	Negative	beerneg	-6
57	P07	Beer	Neutral	beerneut	0
58	P07	Wine	Positive	winepos	24
59	P07	Wine	Negative	wineneg	-9
60	P07	Wine	Neutral	wineneut	19
61	P07	Water	Positive	waterpos	15
62	P07	Water	Negative	waterneg	-10
63	P07	Water	Neutral	waterneut	2
64	P08	Beer	Positive	beerpos	15
65	P08	Beer	Negative	beerneg	-9
66	P08	Beer	Neutral	beerneut	4
67	P08	Wine	Positive	winepos	29
68	P08	Wine	Negative	wineneg	-18
69	P08	Wine	Neutral	wineneut	7
70	P08	Water	Positive	waterpos	13
71	P08	Water	Negative	waterneg	-17
72	P08	Water	Neutral	waterneut	8
73	P09	Beer	Positive	beerpos	20
74	P09	Beer	Negative	beerneg	-17
75	P09	Beer	Neutral	beerneut	9
76	P09	Wine	Positive	winepos	30
77	P09	Wine	Negative	wineneg	-17
78	P09	Wine	Neutral	wineneut	12
79	P09	Water	Positive	waterpos	16
80	P09	Water	Negative	waterneg	-4
81	P09	Water	Neutral	waterneut	10
82	P10	Beer	Positive	beerpos	9
83	P10	Beer	Negative	beerneg	-12
84	P10	Beer	Neutral	beerneut	-5

85	P10	Wine	Positive	winepos	24
86	P10	Wine	Negative	wineneg	-15
87	P10	Wine	Neutral	wineneut	18
88	P10	Water	Positive	waterpos	17
89	P10	Water	Negative	waterneg	-4
90	P10	Water	Neutral	waterneut	8
91	P11	Beer	Positive	beerpos	14
92	P11	Beer	Negative	beerneg	-11
93	P11	Beer	Neutral	beerneut	7
94	P11	Wine	Positive	winepos	34
95	P11	Wine	Negative	wineneg	-14
96	P11	Wine	Neutral	wineneut	20
97	P11	Water	Positive	waterpos	19
98	P11	Water	Negative	waterneg	-1
99	P11	Water	Neutral	waterneut	12
100	P12	Beer	Positive	beerpos	43
101	P12	Beer	Negative	beerneg	30
102	P12	Beer	Neutral	beerneut	8
103	P12	Wine	Positive	winepos	20
104	P12	Wine	Negative	wineneg	-12
105	P12	Wine	Neutral	wineneut	4
106	P12	Water	Positive	waterpos	9
107	P12	Water	Negative	waterneg	-10
108	P12	Water	Neutral	waterneut	-13
109	P13	Beer	Positive	beerpos	15
110	P13	Beer	Negative	beerneg	-6
111	P13	Beer	Neutral	beerneut	13
112	P13	Wine	Positive	winepos	23
113	P13	Wine	Negative	wineneg	-15
114	P13	Wine	Neutral	wineneut	15
115	P13	Water	Positive	waterpos	29
116	P13	Water	Negative	waterneg	-1
117	P13	Water	Neutral	waterneut	10
118	P14	Beer	Positive	beerpos	15
119	P14	Beer	Negative	beerneg	15
120	P14	Beer	Neutral	beerneut	12
121	P14	Wine	Positive	winepos	20
122	P14	Wine	Negative	wineneg	-15
123	P14	Wine	Neutral	wineneut	6
124	P14	Water	Positive	waterpos	6
125	P14	Water	Negative	waterneg	-16
126	P14	Water	Neutral	waterneut	1
127	P15	Beer	Positive	beerpos	40
128	P15	Beer	Negative	beerneg	30
129	P15	Beer	Neutral	beerneut	19
130	P15	Wine	Positive	winepos	28
131	P15	Wine	Negative	wineneg	-4
132	P15	Wine	Neutral	wineneut	0
133	P15	Water	Positive	waterpos	20
134	P15	Water	Negative	waterneg	-10
135	P15	Water	Neutral	waterneut	2
136	P16	Beer	Positive	beerpos	8
137	P16	Beer	Negative	beerneg	12
138	P16	Beer	Neutral	beerneut	8
139	P16	Wine	Positive	winepos	11
140	P16	Wine	Negative	wineneg	-2
141	P16	Wine	Neutral	wineneut	6
142	P16	Water	Positive	waterpos	27
143	P16	Water	Negative	waterneg	5
144	P16	Water	Neutral	waterneut	-5
145	P17	Beer	Positive	beerpos	17
146	P17	Beer	Negative	beerneg	17
147	P17	Beer	Neutral	beerneut	15
148	P17	Wine	Positive	winepos	17
149	P17	Wine	Negative	wineneg	-6
150	P17	Wine	Neutral	wineneut	6
151	P17	Water	Positive	waterpos	9
152	P17	Water	Negative	waterneg	-6
153	P17	Water	Neutral	waterneut	-13
154	P18	Beer	Positive	beerpos	30
155	P18	Beer	Negative	beerneg	21
156	P18	Beer	Neutral	beerneut	21
157	P18	Wine	Positive	winepos	15
158	P18	Wine	Negative	wineneg	-2
159	P18	Wine	Neutral	wineneut	16
160	P18	Water	Positive	waterpos	19
161	P18	Water	Negative	waterneg	-20
162	P18	Water	Neutral	waterneut	3
163	P19	Beer	Positive	beerpos	34
164	P19	Beer	Negative	beerneg	23
165	P19	Beer	Neutral	beerneut	28
166	P19	Wine	Positive	winepos	27
167	P19	Wine	Negative	wineneg	-7

```
168         P19  Wine  Neutral  wineneut        7
169         P19 Water Positive  waterpos       12
170         P19 Water Negative  waterneg      -12
171         P19 Water  Neutral waterneut        2
172         P20  Beer Positive   beerpos       34
173         P20  Beer Negative   beerneg       20
174         P20  Beer  Neutral  beerneut       26
175         P20  Wine Positive   winepos       24
176         P20  Wine Negative   wineneg      -10
177         P20  Wine  Neutral  wineneut       12
178         P20 Water Positive  waterpos       12
179         P20 Water Negative  waterneg       -9
180         P20 Water  Neutral waterneut        4
```

✓ Use *ggplot2* to plot boxplots of the attitude scores for each type of drink (*x*-axis) after adverts using different imagery (different plots).

```
attitudeBoxplot <- ggplot(longAttitude, aes(drink, attitude))

attitudeBoxplot + geom_boxplot() + facet_wrap(~imagery, nrow = 1) + labs(x = "Type of Drink", y
= "Mean Preference Score")
```

✓ Using *ggplot2* and *stat.desc*, plot an error bar graph and get the means for the main effect of **drink**.

Graph:

```
drinkBar <- ggplot(longAttitude, aes(drink, attitude))

drinkBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black") +
stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Type of Drink", y = "Mean
Attitude")
```

Descriptive statistics:

```
by(longAttitude$attitude, longAttitude$drink, stat.desc, basic = FALSE)
```

✓ Using *ggplot2* and *stat.desc*, plot an error bar graph and get the means for the main effect of **imagery**.

Graph:

```
imageryBar <- ggplot(longAttitude, aes(imagery, attitude))

imageryBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black") +
stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Type of Imagery", y =
"Mean Attitude")
```

Descriptive statistics:

```
by(longAttitude$attitude, longAttitude$imagery, stat.desc, basic = FALSE)
```

✓ Using *ggplot2*, plot a line graph with error bars of the means for the **drink** × **imagery** interaction.

```
attitudeInt <- ggplot(longAttitude, aes(drink, attitude, colour = imagery))

attitudeInt + stat_summary(fun.y = mean, geom = "point") + stat_summary(fun.y = mean, geom =
"line", aes(group= imagery)) + stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width =
0.2) + labs(x = "Type of Drink", y = "Mean Attitude", colour = "Type of Imagery")
```

Oliver Twisted

Please Sir, can I have some more … sphericity?

The following article appears in:

Field, A. P. (1998). A bluffer's guide to sphericity. *Newsletter of the Mathematical, Statistical and Computing Section of the British Psychological Society*, 6(1), 13–22

It appears in adapted form below.

A bluffer's guide to sphericity

The use of repeated measures, where the same subjects are tested under a number of conditions, has numerous practical and statistical benefits. For one thing it reduces the error variance caused by between-group individual differences; however, this reduction of error comes at a price because repeated-measures designs potentially introduce covariation between experimental conditions (this is because the same people are used in each condition and so there is likely to be some consistency in their behaviour across conditions). In between-group ANOVA we have to assume that the groups we test are independent for the test to be accurate (Scariano & Davenport, 1987, have documented some of the consequences of violating this assumption). As such, the relationship between treatments in a repeated-measures design creates problems with the accuracy of the test statistic. The purpose of this article is to explain, as simply as possible, the issues that arise in analysing repeated-measures data with ANOVA: specifically, what is sphericity and why is it important?

What is Sphericity?

Most of us are taught during our degrees that it is crucial to have homogeneity of variance between conditions when analysing data from *different* subjects, but often we are left to assume that this problem 'goes away' in repeated-measures designs. This is not so, and the assumption of sphericity can be likened to the assumption of homogeneity of variance in between-group ANOVA.

Sphericity (denoted by ε and sometimes referred to as *circularity*) is a more general condition of *compound symmetry*. Imagine you had a population covariance matrix $\boldsymbol{\Sigma}$, where:

$$\Sigma = \begin{bmatrix} s_{11}^2 & \alpha_{12} & \alpha_{13} & \ldots & \alpha_{1n} \\ \alpha_{21} & s_{22}^2 & \alpha_{23} & \ldots & \alpha_{2n} \\ \alpha_{31} & \alpha_{32} & s_{33}^2 & \ldots & \alpha_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \alpha_{n3} & \ldots & s_{nn}^2 \end{bmatrix}$$

Equation 1

This matrix represents two things: (1) the off-diagonal elements represent the covariances between the treatments 1, ..., *n* (you can think of this as the unstandardized correlation between each of the repeated-measures conditions); and (2) the diagonal elements signify the variances within each treatment. As such, the assumption of homogeneity of variance between treatments will hold when:

$$s_{11}^2 \approx s_{22}^2 \approx s_{33}^2 \approx \ldots \approx s_{nn}^2$$

Equation 2

(i.e. when the diagonal components of the matrix are approximately equal). This is comparable to the situation we would expect in a between-group design. However, in repeated-measures designs there is the added complication that the experimental conditions covary with each other. The end result is that we have to consider the effect of these covariances when we analyse the data, and specifically we need to assume that all of the

covariances are approximately equal (i.e. all of the conditions are related to each other to the same degree and so the effect of participating in one treatment level after another is also equal). Compound symmetry holds when there is a pattern of constant variances along the diagonal (i.e. homogeneity of variance — see Equation 2) and constant covariances off of the diagonal (i.e. the covariances between treatments are equal — see Equation 3). While compound symmetry has been shown to be a sufficient condition for conduction ANOVA on repeated-measures data, it is not a necessary condition.

$$\alpha_{12} \approx \alpha_{13} \approx \alpha_{23} \approx \ldots \approx \alpha_{1n} \approx \alpha_{2n} \approx \alpha_{3n} \approx \ldots$$

Equation 3

Sphericity is a less restrictive form of compound symmetry (in fact much of the early research into repeated-measures ANOVA confused compound symmetry with sphericity). Sphericity refers to the equality of variances of the *differences* between treatment levels. Whereas compound symmetry concerns the covariation between treatments, sphericity is related to the variance of the differences between treatments. So, if you were to take each pair of treatment levels, and calculate the differences between each pair of scores, then it is necessary that these differences have equal variances. Imagine a situation where there are 4 levels of a repeated-measures treatment (A, B, C, D). For sphericity to hold, one condition must be satisfied:

$$s^2_{A-B} \approx s^2_{A-C} \approx s^2_{A-D} \approx s^2_{B-C} \approx s^2_{B-D} \approx s^2_{C-D}$$

Equation 4

Sphericity is violated when the condition in Equation 4 is not met (i.e. the differences between pairs of conditions have unequal variances).

How is Sphericity Measured?

The simplest way to see whether or not the assumption of sphericity has been met is to calculate the differences between pairs of scores in all combinations of the treatment levels. Once this has been done, you can simply calculate the variance of these differences. E.g. Table 1 shows data from an experiment with 3 conditions (for simplicity there are only 5 scores per condition). The differences between pairs of conditions can then be calculated for each subject. The variance for each set of differences can then be calculated. We saw above that sphericity is met when these variances are roughly equal. For this data, sphericity will hold when:

$$s^2_{A-B} \approx s^2_{A-C} \approx s^2_{B-C}$$

where:

$$s^2_{A-B} = 15.7$$
$$s^2_{A-C} = 10.3$$
$$s^2_{B-C} = 10.3$$

As such,

$$s^2_{A-B} \neq s^2_{A-C} = s^2_{B-C}$$

Condition *A*	Condition *B*	Condition *C*	*A*−*B*	*A*−*C*	*B*−*C*
10	12	8	−2	2	5
15	15	12	0	3	3
25	30	20	−5	5	10
35	30	28	5	7	2
30	27	20	3	10	7
		Variance:	**15.7**	**10.3**	**10.3**

Table 1: Hypothetical data to illustrate the calculation of the variance of the differences between conditions.

So there is at least some deviation from sphericity because the variance of the differences between conditions *A* and *B* is greater than the variance of the differences between conditions *A* and *C*, and between *B* and *C*. However, we can say that this data has *local circularity* (or local sphericity) because two of the variances are identical. This means that for any multiple comparisons involving these differences, the sphericity assumption has been met (for a discussion of local circularity see Rouanet & Lépine, 1970). The deviation from sphericity in the data in Table 1 does not seem too severe (all variances are *roughly* equal). This raises the issue of how we assess whether violations from sphericity are severe enough to warrant action.

Assessing the Severity of Departures from Sphericity

Luckily the advancement of computer packages makes it needless to ponder the details of how to assess departures from sphericity. SPSS produces a test known as Mauchly's test, which tests the hypothesis that the variances of the differences between conditions are equal. Therefore, if Mauchly's test statistic is significant (i.e. has a probability value less than .05) we must conclude that there are significant differences between the variance of differences, *ergo* the condition of sphericity has not been met. If, however, Mauchly's test statistic is non-significant (i.e. $p > .05$) then it is reasonable to conclude that the variances of differences are not significantly different (i.e. they are roughly equal). So, in short, if Mauchly's test is significant then we must be wary of the *F*-ratios produced by the computer.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
FACTOR1	.011	13.485	2	.001	.503	.506	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: FACTOR1

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the layers (by default) of the Tests of Within Subjects Effects table.

Figure 1: Output of Mauchly's test from SPSS version 7.0

Figure 1 shows the result of Mauchly's test on some fictitious data with three conditions (*A*, *B* and *C*). The result of the test is highly significant, indicating that the variance between the differences were significantly different. The table also displays the degrees of freedom (the *df* are simply $N - 1$, where N is the number of variances compared) and three estimates of sphericity (see section on correcting for sphericity).

What is the Effect of Violating the Assumption of Sphericity?

Rouanet and Lépine (1970) provided a detailed account of the validity of the *F*-ratio when the sphericity assumption does not hold. They argued that there are two different *F*-ratios that can be used to assess treatment comparisons. The two types of *F*-ratio were labelled *F′* and *F″* respectively. *F′* refers to an *F*-ratio derived from the mean squares of the comparison in question and the interaction of the subjects with that comparison (i.e. the specific error term for each comparison is used — this is the F-ratio normally used). *F″* is derived not from the specific error mean square but from the total error mean squares for all of the repeated-measures comparisons. Rouanet and Lépine (1970) argued that *F′* is less powerful than *F″* and so it may be the case that this test statistic misses genuine effects. In addition, they showed that for *F′* to be valid the covariation matrix, $\boldsymbol{\Sigma}$, must obey local circularity (i.e. sphericity must hold for the *specific comparison in question*) and Mendoza, Toothaker and Crain (1976) have supported this by demonstrating that the *F*-ratios of an $L \times J \times K$ factorial design with two repeated-measures are valid only if local circularity holds. *F′* requires only *overall* circularity (i.e. the whole data set must be circular) but because of the non-reciprocal nature of circularity and compound symmetry, *F″* does not require compound symmetry whilst *F*′ does. So, given that *F′* is the statistic generally used, the effect of violating sphericity is a loss of power (compared to when *F*″ is used) and a test statistic (*F*-ratio) which simply cannot be validly compared to tabulated values of the *F*-distribution.

Correcting for Violations of Sphericity

If data violates the sphericity assumption there are a number of corrections that can be applied to produce a valid *F*-ratio. SPSS produces three corrections based upon the estimates of sphericity advocated by Greenhouse and Geisser (1959) and Huynh and Feldt (1976). Both of these estimates give rise to a correction factor that is applied to the degrees of freedom used to assess the observed value of *F*. How each estimate is calculated is beyond the scope of this article; for our purposes, all we need know is that each estimate differs slightly from the others. The Greenhouse–Geisser estimate (usually denoted as $\hat{\varepsilon}$) varies between $1/(k-1)$ (where k is the number of repeated-measures conditions) and 1. The closer that $\hat{\varepsilon}$ is to 1.00, the more homogeneous are the variances of differences, and hence the closer the data are to being spherical. Figure 1 shows a situation with three conditions and hence the lower limit of $\hat{\varepsilon}$ is 0.5; it is clear that the calculated value of $\hat{\varepsilon}$ is 0.503 which is very close to 0.5 and so represents a substantial deviation from sphericity. Huynh and Feldt (1976) reported that when $\hat{\varepsilon} > 0.75$ too many false null hypotheses fail to be rejected (i.e. the test is too conservative) and Collier, Baker, Mandeville & Hayes (1967) showed that this was also true with $\hat{\varepsilon}$ as high as 0.90. Huynh and Feldt, therefore, proposed a correction to $\hat{\varepsilon}$ (usually denoted as $\tilde{\varepsilon}$) to make it less conservative. However, Maxwell and Delaney (1990) report that $\tilde{\varepsilon}$ actually overestimates sphericity. Stevens (1992) therefore recommends taking an average of the two and adjusting the *df* by this averaged value. Girden (1992) recommends that when $\hat{\varepsilon} > 0.75$ then the *df* should be corrected using $\tilde{\varepsilon}$; if $\hat{\varepsilon} < 0.75$, or nothing is known about sphericity at all, then the conservative $\hat{\varepsilon}$ should be used to adjust the *df*.

Tests of Within-Subjects Effects

Measure		Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
MEASURE_1	Sphericity Assumed	FACTOR1	2895.600	2	1447.800	5.245	.035	10.489	.662
		Error(FACTOR1)	2208.400	8	276.050				
	Greenhouse-Geisser	FACTOR1	2895.600	1.006	2879.437	5.245	.083	5.274	.418
		Error(FACTOR1)	2208.400	4.022	549.018				
	Huynh-Feldt	FACTOR1	2895.600	1.011	2863.394	5.245	.083	5.304	.420
		Error(FACTOR1)	2208.400	4.045	545.959				
	Lower-bound	FACTOR1	2895.600	1.000	2895.600	5.245	.084	5.245	.417
		Error(FACTOR1)	2208.400	4.000	552.100				

a. Computed using alpha = .05

Figure 2: Output of epsilon corrected *F*-values from SPSS version 7.0.

Figure 2 shows a typical ANOVA table for a set of data that violated sphericity (the same data used to generate Figure 1). The table in Figure 2 shows the *F*-ratio and associated degrees of freedom when sphericity is assumed; as can be seen, this results in a significant *F*-statistic indicating some difference(s) between the means of the three conditions. Underneath are the corrected values (for each of the three estimates of sphericity). Notice that in all cases the *F*-ratios remain the same, it is the degrees of freedom that change (and hence the critical value of *F*). The degrees of freedom are corrected by the estimate of sphericity. How this is done can be seen in Table 2. The new degrees of freedom are then used to ascertain the critical value of *F*. For this data this results in the observed *F* being non-significant at $p < 0.05$. This particular data set illustrates how important it is to use a valid critical value of *F*; it can mean the difference between a statistically significant result and a non-significant result. More importantly, it can mean the difference between making a Type I error and not.

Estimate of Sphericity Used	Value of Estimate	Term	*df*	Correction	New *df*
None		Effect	2		
		Error	8		
	0.503	Effect	2	0.503×2	1.006
		Error	8	0.503×8	4.024
	0.506	Effect	2	0.506×2	1.012
		Error	8	0.506×8	4.048

Table 2: Shows how the sphericity corrections are applied to the degrees of freedom.

Multivariate vs. Univariate Tests

A final option, when you have data that violates sphericity, is to use multivariate test statistics (MANOVA) because they are not dependent upon the assumption of sphericity (see O'Brien & Kaiser, 1985). There is a trade-off of test power between univariate and multivariate approaches although some authors argue that this can be overcome with suitable mastery of the techniques (O'Brien & Kaiser, 1985). MANOVA avoids the assumption of sphericity (and all the corresponding considerations about appropriate *F* ratios and corrections) by using a specific error term for contrasts with 1 *df*, and hence each contrast is only ever associated with its specific error term (rather than the pooled error terms used in ANOVA). Davidson (1972) compared the power of adjusted univariate techniques with those of Hotellings T^2 (a MANOVA test statistic) and found that the univariate technique was relatively powerless to detect small reliable changes between highly correlated conditions when other less correlated conditions were also present. Mendoza, Toothaker and Nicewander (1974) conducted a Monte Carlo study comparing univariate and multivariate techniques under violations of compound symmetry and normality and found that 'as the degree of violation of compound symmetry increased, the empirical power for the multivariate tests also increased. In contrast, the power for the univariate tests generally decreased' (p. 174). Maxwell and Delaney (1990) noted that the univariate test is relatively more powerful than the multivariate test as *n* decreases and proposed that 'the multivariate approach should probably not be used if *n* is less than *a* + 10 (*a* is the number of levels for repeated-measures)' (p. 602). As a general rule it seems that when you have a large violation of sphericity ($\varepsilon < 0.7$) and your sample size is greater than *a* + 10 then multivariate procedures are more powerful whilst with small sample sizes or when sphericity holds ($\varepsilon > 0.7$) the univariate approach is preferred (Stevens, 1992). It is also worth noting that the power of MANOVA increases and decreases as a function of the correlations between dependent variables (Cole, Maxwell, Arvey, & Salas, 1994) and so the relationship between treatment conditions must be considered also.

Multiple Comparisons

So far, I have discussed the effects of sphericity on the omnibus ANOVA. As a final flurry some discussion of the effects on multiple comparison procedures is warranted. Boik (1981) provided an estimable account of the effects

of non-sphericity on *a priori* tests in repeated-measures designs, and concluded that even very small departures from sphericity produce large biases in the *F*-test and recommends against using these tests for repeated-measures contrasts. When experimental error terms are small, the power to detect relatively strong effects can be as low as .05 (when sphericity = .80). He argues that the situation for *a priori* comparisons cannot be improved and concludes by recommending a multivariate analogue. Mitzel and Games (1981) found that when sphericity does not hold ($\varepsilon < 1$) the pooled error term conventionally employed in pairwise comparisons resulted in non-significant differences between two means declared significant (i.e. a lenient Type 1 error rate) or undetected differences (a conservative Type 1 error rate). They therefore recommended the use of separate error terms for each comparison. Maxwell (1980) systematically tested the power and alpha levels for 5 *a priori* tests under repeated-measures conditions. The tests assessed were Tukey's wholly significant difference (WSD) test which uses a pooled error term, Tukey's procedure but with a separate error term with either ($n-1$) *df* [labelled SEP1] or $(n-1)(k-1)$ *df* [labelled SEP2], Bonferroni's procedure (BON), and a multivariate approach — the Roy–Bose simultaneous confidence interval (SCI). Maxwell tested these *a priori* procedures varying the sample size, number of levels of the repeated factor and departure from sphericity. He found that the multivariate approach was always 'too conservative for practical use' (p. 277) and this was most extreme when *n* (the number of subjects) is small relative to *k* (the number of conditions). Tukey's test inflated the alpha rate as the covariance matrix departs from sphericity and even when a separate error term was used (SEP1) alpha was slightly inflated as *k* increased whilst SEP2 also lead to unacceptably high alpha levels. The Bonferroni method, however, was extremely robust (although *slightly* conservative) and controlled alpha levels regardless of the manipulation. Therefore, in terms of Type I error rates the Bonferroni method was best. In terms of test power (the Type II error rate) for a small sample (*n* = 8) WSD was the most powerful under conditions of non-sphericity. This advantage was severely reduced when *n* = 15. Keselman and Keselman (1988) extended Maxwell's work and also investigated unbalanced designs. They too used Tukey's WSD, a modified WSD (with non-pooled error variance), Bonferroni *t*-statistics, and a multivariate approach, and looked at the same factors as Maxwell (with the addition of unequal samples). They found that when unweighted means were used (with unbalanced designs) none of the four tests could control the Type 1 error rate. When weighted means were used only the multivariate tests could limit alpha rates although Bonferroni *t*-statistics were considerably better than the two Tukey methods. In terms of power, they concluded that 'as the number of repeated treatment levels increases, BON is substantially more powerful than SCI' (p. 223).

So, in terms of these studies, the Bonferroni method seems to be generally the most robust of the univariate techniques, especially in terms of power and control of the Type I error rate.

Conclusion

It is more often the rule than the exception that sphericity is violated in repeated-measures designs. For this reason, all repeated-measures designs should be exposed to tests of violations of sphericity. If sphericity is violated then the researcher must decide whether a multivariate or univariate analysis is preferred (with due consideration to the trade-off between test validity on one hand and power on the other). If univariate methods are chosen then the omnibus ANOVA must be corrected appropriately, depending on the level of departure from sphericity. Finally, if pairwise comparisons are required the Bonferroni method should probably be used to control the Type 1 error rate. Finally, to ensure that the group sizes are equal otherwise even the Bonferroni technique is subject to inflations of alpha levels.

References

Boik, R. J. (1981). A priori tests in repeated measures designs: effects of nonsphericity, *Psychometrika*, *46*(3), 241–255.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables, *Psychological Bulletin*, *115*(3), 465–474.

Girden, E. R. (1992). *ANOVA: Repeated measures*. Sage University Paper Series on Qualitative Applications in the Social Sciences, 84. Newbury Park, CA: Sage.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomised block and split-plot designs. *Journal of Educational Statistics*, *1*(1), 69–82.

Keselman, H. J., & Keselman, J. C. (1988). Repeated measures multiple comparison procedures: Effects of violating multisample sphericity in unbalanced designs. *Journal of educational Statistics*, *13*(3), 215–226.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, *5*(3), 269–287.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research*, *9***,** 165-177.

Mendoza, J. L., Toothaker, L. E., & Crain, B. R. (1976). Necessary and sufficient conditions for *F* ratios in the *L* × *J* × *K* factorial design with two repeated factors. *Journal of the American Statistical Association*, *71***,** 992–993.

Mitzel, H. C., & Games, P. A. (1981). Circularity and multiple comparisons in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *34*, 253–259.

O'Brien, M. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*(2), 316–333.

Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: Anova and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, *23***,** 147–163.

Scariano, S. M. ,& Davenport, J. M. (1987). The effects of violations of independence in the one-way ANOVA. *American Statistician*, *41*(2), 123–129.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Labcoat Leni's real research

Who's afraid of the big bad wolf?

Problem

Field, A. P. (2006). *Journal of Abnormal Psychology*, *115*(4), 742–752.

I'm going to let my ego get the better of me and talk about some of my own research. When I'm not scaring my students with statistics, I scare small children with Australian marsupials. There is a good reason for doing this, which is to try to discover how children develop fears (which will help us to prevent them). Most of my research looks at the effect of giving children information about animals or situations that are novel to them (rather like a parent, teacher or TV show would do). In one particular study (Field, 2006), I used three novel animals (the quoll, quokka and cuscus) and children were told negative things about one of the animals, positive things about another, and were given no information about the third (our control). I then asked the children to place their hands in three wooden boxes each of which they believed contained one of the aforementioned animals. My hypothesis was that they would take longer to place their hand in the box containing the animal about which they had heard negative information.

The data from this part of the study are in the file **Field(2006).dat**. Labcoat Leni wants you to carry out a one-way repeated-measures ANOVA on the times taken for children to place their hands in the three boxes (negative information, positive information, no information). First, draw an error bar graph of the means, then do some normality tests on the data, then do a log transformation on the scores, and do the ANOVA on these log-transformed scores (if you read the paper you'll notice that I found that the data were not normal, so I log-transformed them before doing the ANOVA). Do children take longer to put their hands in a box that they believe contains an animal about which they have been told nasty things?

Solution

As you will be well aware by now, we first need to read in the data:

```
fieldData<-read.delim("Field(2006).dat", header = TRUE)
```

If you execute the dataframe name *fieldData* you will see that the data have originally been entered into SPSS in the wide format. However, do conduct this analysis in **R** we need the data to be in the long format. To convert the data we can use the *melt* function as in the book chapter and your command might look like this:

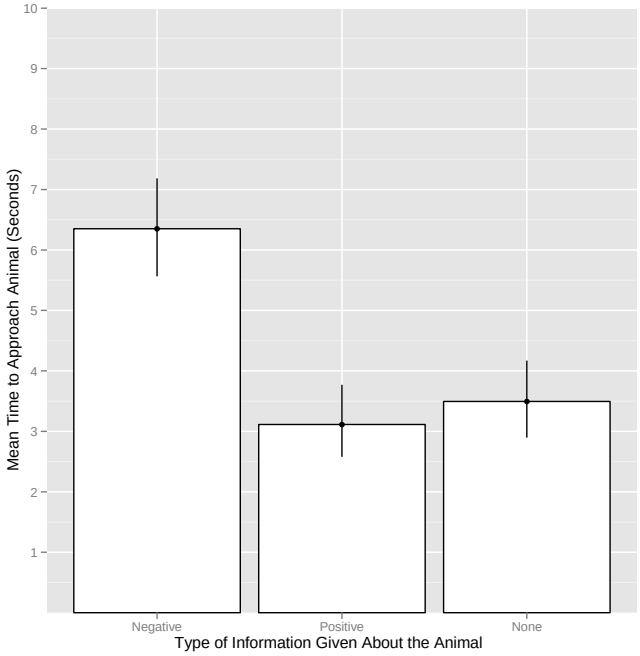
```
longField <-melt(fieldData, id = "code", measured = c("bhvneg", "bhvpos", "bhvnone"))
names(longField)<-c("code", "Information", "Approach")

longField$Information<-factor(longField$Information, labels = c("Negative",
 "Positive", "None"))
longField<-longField[order(longField$code),]
```

If you now execute the dataframe name *longField* you will see that the data are now in the long format. I have changed the labels from **bhvneg**, **bhvpos** and **bhvnone** to **Negative**, **Positive** and **None** because I think it is then clearer what they represent.

We can now plot an error bar graph using the *longField* dataframe by executing:

```
fieldBar <- ggplot(longField, aes(Information, Approach))
fieldBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black")
 + stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Type of
 Information Given About the Animal", y = "Mean Time to Approach Animal (Seconds)") +
 coord_cartesian(ylim=c(0,10)) + scale_y_continuous(breaks = 1:10)
```

Notice that I have added *+ coord_cartesian(ylim=c(0,10)) + scale_y_continuous(breaks = 1:10)*. This line of code specifies the scale of the *y*-axis to be from 0 to 10, which is a more appropriate scale for this particular graph than the default, which was something like 0–25. Basically, because the longest mean time is around 6.4 it seems silly to have a scale that goes up to 25; the graph is clearer with a shorter scale. I have also included breaks from 1 to 10, this makes the graph easier to read. The resulting graph should look like this:



Next we can test for normality using the Shapiro–Wilk test (see Chapter 5). We need to use the *by()* function to conduct the Shapiro–Wilk test for approach time split by type of information. We would execute:

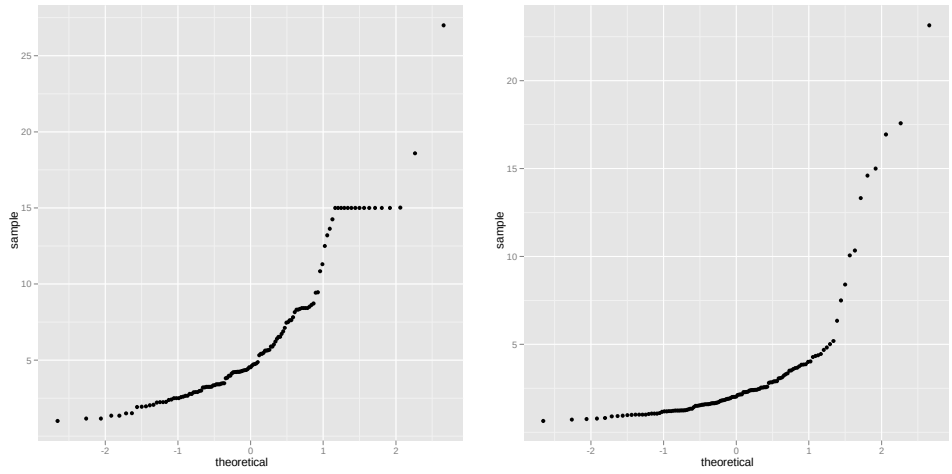
```
by(data=longField$Approach, INDICES=longField$Information, FUN=shapiro.test)

longField$Information: Negative

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8351, p-value = 1.313e-10

------------------------------------------------------
longField$Information: Positive

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.5736, p-value < 2.2e-16

------------------------------------------------------
longField$Information: None

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.6098, p-value < 2.2e-16
```

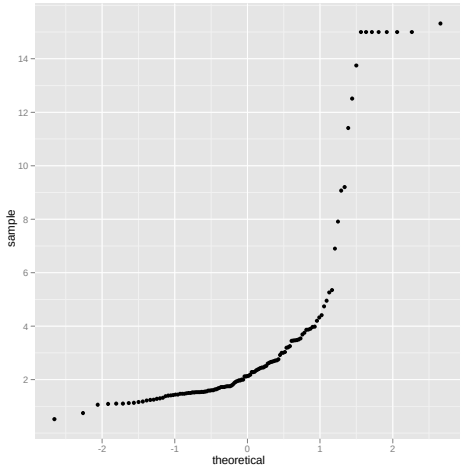
The resulting output above indicates that the data are very heavily non-normal as the *p*-values are all highly significant.

We could also look at some Q-Q plots for each of the information groups (negative, positive and none). To do this we need to use the original dataframe *fieldData* because it was entered in the wide format, which is the correct format for plotting separate plots for the different information groups.

To plot the Q-Q plots we would execute the following commands one at a time (remember that the variable names in *fieldData* were *bhvneg*, *bhvpos* and *bhvnone*, rather than *positive*, *negative* and *none*):

```
qplot(sample = fieldData$bhvneg, stat="qq")
qplot(sample = fieldData$bhvpos, stat="qq")
qplot(sample = fieldData$bhvnone, stat="qq")
```





The resulting Q-Q plots suggest that the data are very heavily skewed. This will be, in part, because if a child didn't put their hand in the box after 15 seconds we gave them a score of 15 and asked them to move on to the next box (this was for ethical reasons: if a child hadn't put their hand in the box after 15 s we assumed that they did not want to do the task).

To log-transform the scores we need to execute the following commands one at a time:

```
fieldData$logneg <- log(fieldData$bhvneg + 1)
fieldData$logpos <- log(fieldData$bhvpos + 1)
fieldData$lognone <- log(fieldData$bhvnone + 1)
```

As you can see, in the above code we used the *fieldData* dataframe because to be able to log-transform each of the types of information variables we needed the data to be in the wide format. If you now execute *fieldData*, you will see that three variables have been added to the dataframe *logneg*, *logpos* and *lognone.*

We can then rerun the Shapiro–Wilk test on these transformed scores by executing:

```
shapiro.test(fieldData$logneg)
shapiro.test(fieldData$logpos)
shapiro.test(fieldData$lognone)
```

```
> shapiro.test(fieldData$logneg)

        Shapiro-Wilk normality test

data:  fieldData$logneg
W = 0.9669, p-value = 0.003371

> shapiro.test(fieldData$logpos)

        Shapiro-Wilk normality test

data:  fieldData$logpos
W = 0.8675, p-value = 2.801e-09

> shapiro.test(fieldData$lognone)

        Shapiro-Wilk normality test

data:  fieldData$lognone
W = 0.8296, p-value = 8.109e-11
```

These are all still significant, suggesting that the data are still not normal. If you look at the Field (2006) paper, you will see that once the data were log-transformed, the data were found to be normal. However, in the Field (2006) paper I used the Kolmogorov–Smirnov (KS) test for normality whereas I have used the Shapiro–Wilk test here. Various studies have found that, even in this corrected form, the KS test is less powerful for testing normality than the Shapiro–Wilk test (Stephens, 1974).

What we can do is conduct a robust ANOVA on these data. Remember that to conduct a robust ANOVA we need the data to be in the wide format, therefore we will need to use the original dataframe *fieldData*.

First, we need to get rid of the **code** variable in the *fieldData* dataframe because we want only the scores. To do this we could create a new data frame (*fieldData2)* that excludes this first column by executing:

```
fieldData2<-fieldData[, -c(1)]
```

We can now do a one-way repeated measures ANOVA based on trimmed means by executing:

```
rmanova(fieldData2)
```

```
[1] "The number of groups to be compared is"
[1] 3
$test
[1] 78.15207

$df
[1]  1.241041 94.319122

$siglevel
[1] 6.661338e-16

$tmeans
[1] 5.154675 2.164935 2.296753

$ehat
[1] 0.6174746

$etil
[1] 0.6205205
```

In the output above we are given a test statistic, *F*, for the effect of information (*$test*), the degrees of freedom (*$df*), the *p*-value (*$siglevel*), the group means (*$tmeans*). Given that the significance level is much less than .05, we can say that there were significant differences in approach times after hearing different types of information, $F(1.24, 94.32) = 78.15$, $p < .001$.

Assuming we leave the default options, to run *post hoc* tests based on a 20% trimmed mean, we execute:

```
rmmcp(fieldData2)
```

```
$test
     Group Group      test      p.value p.crit        se
[1,]     1     2  8.464641 1.417089e-12 0.0169 0.2853735
[2,]     1     3  7.068277 6.528487e-10 0.0250 0.2928762
[3,]     2     3 -2.528902 1.351455e-02 0.0500 0.0814480

$psihat
     Group Group     psihat   ci.lower     ci.upper
[1,]     1     2  2.4155844  1.7169548  3.114213982
[2,]     1     3  2.0701299  1.3531329  2.787126862
[3,]     2     3 -0.2059740 -0.4053687 -0.006579304

$con
     [,1]
[1,]    0

$num.sig
[1] 3
```

The output shows the *post hoc* tests based on trimmed means. If the value of *p.value* is less than the critical value (*p.crit*) and the confidence interval does not cross zero then the comparison is significant. The columns labelled *group* tells you which groups are being compared (the numbers relate to the columns in the dataframe.

- ✓ [1,] tests the difference between negative information and positive information. This contrast is significant because *p.value* (.00) is less than *p.crit* (.02) and the confidence interval does not cross zero.
- ✓ [2,] tests the difference between negative information and no information. This contrast is significant because *p.value* (.00) is less than *p.crit* (.03) and the confidence interval does not cross zero.

- ✓ [3,] tests the difference between negative information and positive information. This contrast is significant because *p.value* (.01) is less than *p.crit* (.05) and the confidence interval does not cross zero.

We could report that a child took longer to place their hand in the box that they believed contained an animal about which they had heard bad things compared to the boxes that they believed contained animals that they had heard positive information about, $\Psi = 2.42$ (1.72, 3.11), $p < .001$, or no information $\Psi = 2.07$ (1.35, 2.79), $p < .001$. There was also a significant difference between the approach times for the 'positive information' and 'no information' boxes in that children took longer to place their hand in the box that they believed contained an animal about which they had heard no information compared to a box containing an animal about which they had heard positive information, $\Psi = -0.21$ (–0.41, –0.01), $p < .05$.

Smart Alex's solutions

Task 1

- Students often worry about the consistency of marking between lecturers. Lecturers obtain reputations for being 'hard' or 'light' markers (or to use the students' terminology, 'evil manifestations from Beelzebub's bowels' and 'nice people'), but there is often little to substantiate these reputations. A group of students investigated the consistency of marking by submitting the same essays to four different lecturers. The mark given by each lecturer was recorded for each of the eight essays. This design is repeated measures because every lecturer marked every essay. The independent variable was the lecturer who marked the report and the dependent variable was the percentage mark given. The data are in the file **TutorMarks.dat**. Conduct a one-way ANOVA on these data by hand.

Data for essay marks example:

Essay	Tutor 1 (Professor Field)	Tutor 2 (Professor Smith)	Tutor 3 (Professor Scrote)	Tutor 4 (Professor Death)	Mean	s^2
1	62	58	63	64	**61.75**	**6.92**
2	63	60	68	65	**64.00**	**11.33**
3	65	61	72	65	**65.75**	**20.92**
4	68	64	58	61	**62.75**	**18.25**
5	69	65	54	59	**61.75**	**43.58**
6	71	67	65	50	**63.25**	**84.25**
7	78	66	67	50	**65.25**	**132.92**
8	75	73	75	45	**67.00**	**216.00**
Mean	**68.875**	**64.25**	**65.25**	**57.375**		

There were eight essays, each marked by four different lecturers. Their marks are shown in the table. In addition, the mean mark given by each lecturer is shown in the table, and also the mean mark that each essay received and the variance of marks for a particular essay. Now, the total variance within essays will in part be caused by the fact that different lecturers are harder or softer markers (the manipulation), and will, in part, be caused by the fact that the essays themselves will differ in quality (individual differences).

***The total sum of squares* (SS_T)**

Remember from one-way independent ANOVA that SS_T is calculated using the following equation:

$$SS_T = s^2_{grand}(N-1)$$

Well, in repeated-measures designs the total sum of squares is calculated in exactly the same way. The grand variance in the equation is simply the variance of all scores when we ignore the group to which they belong. So if we treated the data as one big group it would look as follows:

62	58	63	64
63	60	68	65
65	61	72	65
68	64	58	61
69	65	54	59
71	67	65	50
78	66	67	50
75	73	75	45

Grand Mean = 63.9375

The variance of these scores is 55.028 (try this on your calculator). We used 32 scores to generate this value, and so *N* is 32. As such the equation becomes:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 55.028(32-1) \\ &= 1705.868 \end{aligned}$$

The degrees of freedom for this sum of squares, as with the independent ANOVA will be *N*–1, or 31.

The within-participant sum of squares **(SS_W)**

The crucial variation in this design is that there is a variance component called the within-participant variance (this arises because we've manipulated our independent variable within each participant). This is calculated using a sum of squares. Generally speaking, when we calculate any sum of squares we look at the squared difference between the mean and individual scores. This can be expressed in terms of the variance across a number of scores and the number of scores on which the variance is based. For example, when we calculated the residual sum of squares in independent ANOVA (SS_R) we used the following equation:

$$\begin{aligned} SS_R &= \sum (x_i - \bar{x}_i)^2 \\ &= s^2(n-1) \end{aligned}$$

This equation gave us the variance between individuals within a particular group, and so is an estimate of individual differences within a particular group. Therefore, to get the total value of individual differences we have to calculate the sum of squares within each group and then add them up:

$$SS_R = s^2_{group1}(n_1 - 1) + s^2_{group2}(n_2 - 1) + s^2_{group3}(n_3 - 1)$$

This is all well and good when we have different people in each group, but in repeated-measures designs we've subjected people to more than one experimental condition, and therefore we're interested in the variation not within a group of people (as in independent ANOVA) but within an actual person. That is, how much variability is there within an individual? To find this out we actually use the same equation but we adapt it to look at people rather than groups. So, if we call this sum of squares SS_W (for within-participant SS) we could write it as:

$$SS_W = s^2_{person1}(n_1 - 1) + s^2_{person2}(n_2 - 1) + s^2_{person3}(n_3 - 1) \ldots + s^2_{person\,n}(n_n - 1)$$

This equation simply means that were looking at the variation in an individual's scores and then adding these variances for all the people in the study. Some of you may have noticed that, in our example, we're using essays rather than people, and so to be pedantic we'd write this as:

$$SS_W = s^2_{essay1}(n_1 - 1) + s^2_{essay2}(n_2 - 1) + s^2_{essay3}(n_3 - 1) \ldots + s^2_{essay\,n}(n_n - 1)$$

The *n*s simply represent the number of scores on which the variances are based (i.e. the number of experimental conditions, or in this case the number of lecturers). All of the variances we need are in the table, so we can calculate SS_W as:

$$\begin{aligned}SS_W &= s^2_{essay1}(n_1 - 1) + s^2_{essay2}(n_2 - 1) + s^2_{essay3}(n_3 - 1) \ldots + s^2_{essay\,n}(n_n - 1)\\ &= (6.92)(4-1) + (11.33)(4-1) + (20.92)(4-1) + (18.25)(4-1)\\ &\quad + (43.58)(4-1) + (84.25)(4-1) + (132.92)(4-1) + (216)(4-1)\\ &= 20.76 + 34 + 62.75 + 54.75 + 130.75 + 252.75 + 398.75 + 648\\ &= 1602.5\end{aligned}$$

The degrees of freedom for each person are $n - 1$ (i.e. the number of conditions minus 1). To get the total degrees of freedom we add the *df* for all participants. So, with eight participants (essays) and four conditions (i.e. $n = 4$) we get $8 \times 3 = 24$ degrees of freedom.

***The model sum of squares* (SS_M)**

So far, we know that the total amount of variation within the data is 1705.868 units. We also know that 1602.5 of those units are explained by the variance created by individuals' (essays') performances under different conditions. Now some of this variation is the result of our experimental manipulation and some of this variation is simply random fluctuation. The next step is to work out how much variance is explained by our manipulation and how much is not.

In independent ANOVA, we worked out how much variation could be explained by our experiment (the model sum of squares) by looking at the means for each group and comparing these to the overall mean. So, we measured the variance resulting from the differences between group means and the overall mean. We do exactly the same thing with a repeated-measures design. First we calculate the mean for each level of the independent variable (in this case the mean mark given by each lecturer) and compare these values to the overall mean of all marks. So, we calculate this sum of squares in the same way as for independent ANOVA:

1. Calculate the difference between the mean of each group and the grand mean.
2. Square each of these differences.
3. Multiply each result by the number of subjects within that group (n_i).
4. Add the values for each group together:

$$SS_M = \sum n_i\left(\bar{x}_i - \bar{x}_{grand}\right)^2$$

Using the means from the essay data, we can calculate SS_M as follows:

$$\begin{aligned}SS_M &= 8(68.875 - 63.9375)^2 + 8(64.25 - 63.9375)^2 + 8(65.25 - 63.9375)^2 + \ldots\\ &\quad + 8(57.375 - 63.9375)^2\\ &= 8(4.9375)^2 + 8(0.3125)^2 + 8(1.3125)^2 + 8(-6.5625)^2\\ &= 554.125\end{aligned}$$

For SS_M, the degrees of freedom (df_M) are again one less than the number of things used to calculate the sum of squares. For the model sums of squares we calculated the sum of squared errors between the four means and the grand mean. Hence, we used four things to calculate these sums of squares. So, the degrees of freedom will be 3. So, as with independent ANOVA, the model degrees of freedom is always the number of *groups* (*k*) minus 1:

$$df_M = k - 1 = 3$$

***The residual sum of squares* (SS_R)**

We now know that there are 1706 units of variation to be explained in our data, and that the variation across our conditions accounts for 1602 units. Of these 1602 units, our experimental manipulation can explain 554 units. The final sum of squares is the residual sum of squares (SS_R), which tells us how much of the variation cannot be explained by the model. This value is the amount of variation caused by extraneous factors outside of experimental control (such as natural variation in

the quality of the essays). Knowing SS_W and SS_M already, the simplest way to calculate SS_R is to subtract SS_M from SS_W:

$$\begin{aligned} SS_R &= SS_W - SS_M \\ &= 1602.5 - 554.125 \\ &= 1048.375 \end{aligned}$$

The degrees of freedom are calculated in a similar way:

$$\begin{aligned} df_R &= df_W - df_M \\ &= 24 - 3 \\ &= 21 \end{aligned}$$

The mean squares

SS_M tells us how much variation the model (e.g. the experimental manipulation) explains and SS_R tells us how much variation is due to extraneous factors. However, because both of these values are summed values, the number of scores that were summed influences them. As with independent ANOVA, we eliminate this bias by calculating the average sum of squares (known as the *mean squares*, MS), which is simply the sum of squares divided by the degrees of freedom:

$$MS_M = \frac{SS_M}{df_M} = \frac{554.125}{3} = 184.708$$

$$MS_R = \frac{SS_R}{df_R} = \frac{1048.375}{21} = 49.923$$

MS_M represents the average amount of variation explained by the model (the systematic variation), whereas MS_R is a gauge of the average amount of variation explained by extraneous variables (the unsystematic variation).

The F-ratio

The *F*-ratio is a measure of the ratio of the variation explained by the model and the variation explained by unsystematic factors. It can be calculated by dividing the model mean squares by the residual mean squares. You should recall that this is exactly the same as for independent ANOVA:

$$F = \frac{MS_M}{MS_R}$$

So, as with the independent ANOVA, the *F*-ratio is still the ratio of systematic variation to unsystematic variation. As such, it is the ratio of the experimental effect to the effect on performance of unexplained factors. For the marking data, the *F*-ratio is:

$$F = \frac{MS_M}{MS_R} = \frac{184.708}{49.923} = 3.70$$

This value is greater than 1, which indicates that the experimental manipulation had some effect above and beyond the effect of extraneous factors. As with independent ANOVA this value can be compared against a critical value based on its degrees of freedom (which are df_M and df_R, which are 3 and 21 in this case).

Task 2

- Repeat the analysis above in **R** and interpret the results.

First of all we need to load the data:

```
tutorData<-read.delim("TutorMarks.dat", header = TRUE)
```

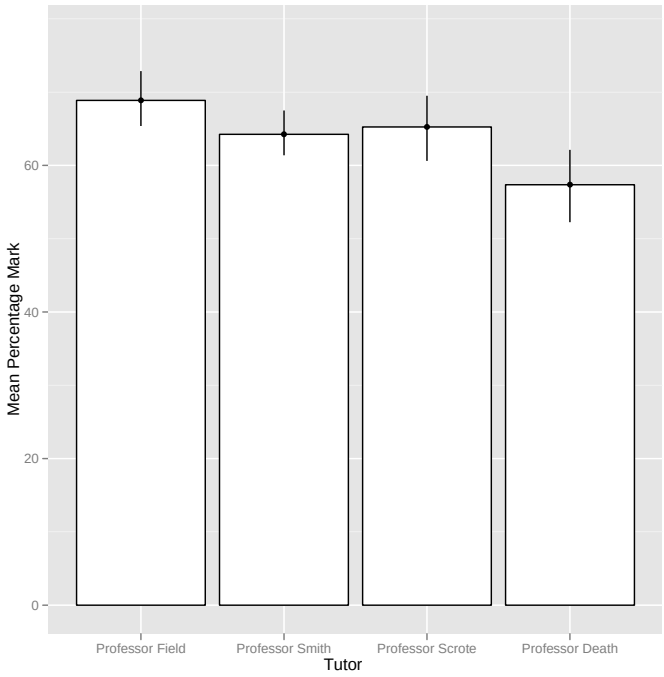
The data were originally entered into **R** in the wide format, but we need them to be in the long format for these analyses. To convert the data into the long format we could execute:

```
longTutor<-melt(tutorData, id = "Essay", measured = c("tutor1", "tutor2", "tutor3",
"tutor4"))
names(longTutor)<-c("Essay", "Tutor", "Mark")

longTutor$Tutor<-factor(longTutor$Tutor, labels = c("Professor Field", "Professor
Smith", "Professor Scrote", "Professor Death"))
longTutor<-longTutor[order(longTutor$Essay),]
```

We can now plot an error bar graph by executing:

```
tutorBar <- ggplot(longTutor, aes(Tutor, Mark))
 tutorBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black")
+ stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Tutor", y =
"Mean Percentage Mark")
```

The resulting graph should look like this:



We could have a look at the descriptive statistics by executing:

```
by(longTutor$Mark, longTutor$Tutor, stat.desc)
```

```
longTutor$Tutor: Professor Field
     nbr.val     nbr.null       nbr.na          min          max        range
  8.00000000   0.00000000   0.00000000  62.00000000  78.00000000  16.00000000
         sum       median         mean      SE.mean CI.mean.0.95          var
551.00000000  68.50000000  68.87500000   1.99497136   4.71735765  31.83928571
     std.dev     coef.var
  5.64263110   0.08192568
------------------------------------------------------------
longTutor$Tutor: Professor Smith
     nbr.val     nbr.null       nbr.na          min          max        range
  8.00000000   0.00000000   0.00000000  58.00000000  73.00000000  15.00000000
         sum       median         mean      SE.mean CI.mean.0.95          var
514.00000000  64.50000000  64.25000000   1.66636902   3.94033660  22.21428571
     std.dev     coef.var
  4.71320334   0.07335725
------------------------------------------------------------
longTutor$Tutor: Professor Scrote
     nbr.val     nbr.null       nbr.na          min          max        range
   8.0000000    0.0000000    0.0000000   54.0000000   75.0000000   21.0000000
         sum       median         mean      SE.mean CI.mean.0.95          var
 522.0000000   66.0000000   65.2500000    2.4476665    5.7878116   47.9285714
     std.dev     coef.var
   6.9230464    0.1061003
```

```
------------------------------------------------------------
longTutor$Tutor: Professor Death
     nbr.val     nbr.null       nbr.na          min          max        range
   8.0000000    0.0000000    0.0000000   45.0000000   65.0000000   20.0000000
         sum       median         mean      SE.mean CI.mean.0.95          var
 459.0000000   60.0000000   57.3750000    2.7962826    6.6121577   62.5535714
     std.dev     coef.var
   7.9090816    0.1378489
```

From the descriptive statistics and the error bar graph we can see that, on average, Professor Field gave the highest marks to the essays (that's because I'm so nice, you see … or it could be because I'm stupid and so have low academic standards?). Professor Death, on the other hand, gave very low grades. These mean values are useful for interpreting any effects that may emerge from the main analysis.

Next we need to set some orthogonal contrasts so that we can look at Type III sums of squares. Let's imagine we had reason to believe that Professor Death was a particularly harsh marker compared to all the other professors. Therefore, our first contrast might compare Professors Field, Smith and Scrote (combined) to Professor Death. We might also predict that Professor Field would be more generous in his marking than Professors Scrote and Smith and so our next contrast could compare Professor Field to Professor Scrote and Professor Smith (combined). We then need a third contrast to separate Professor Scrote from Professor Smith.

To set these orthogonal contrasts we can first create variables representing each contrast and then bind these variables together and set then as the contrast for **Tutor**:

```
NicevsNasty<-c(1, 1, 1, -3)
FieldvsScroteSmith<-c(2, -1, -1, 0)
ScrotevsSmith<-c(0, 1, -1, 0)
contrasts(longTutor$Tutor)<-cbind(NicevsNasty, FieldvsScroteSmith, ScrotevsSmith)
```

Next we can conduct the *ezANOVA* by executing:

```
tutorModel<-ezANOVA(data = longTutor, dv = .(Mark), wid = .(Essay), within =
.(Tutor), type = 3, detailed = TRUE)
```

```
$ANOVA
Effect      DFn DFd    SSn      SSd        F            p           p<.05
(Intercept) 1   7   130816.125 103.375  8858.165659  4.027213e-12    *
Tutor       3   21  554.125    1048.375 3.699893     2.784621e-02    *

  ges
0.9912725
0.3248333

$`Mauchly's Test for Sphericity`

   Effect     W          p       p<.05
2  Tutor   0.1310624  0.04305676   *

$`Sphericity Corrections`

   Effect  GGe       p[GG]      p[GG]<.05    HFe        p[HF]       p[HF]<.05
2  Tutor  0.5576185  0.06287878             0.7122543   0.04712856      *
```

The output above shows the results from the *ezANOVA()*. We'll begin with the sphericity information. Mauchly's test for sphericity should be non-significant if we are to assume that the condition of sphericity has been met. The important column is the one containing the significance value (*p*) and in this case the value, .043, is less than the critical value of .05, so we reject the assumption that the variances of the differences between levels are equal. In other words, the assumption of sphericity has been violated, $W = .13$, $p = .043$.

R produces two corrections based upon the estimates of sphericity advocated by Greenhouse and Geisser (1959) and Huynh and Feldt (1976). Both of these estimates give rise to a correction factor that is applied to the degrees of freedom used to assess the observed *F*-ratio. The *Greenhouse–Geisser correction* varies between 1/(*k*–1) (where *k* is the number of repeated-measures conditions) and 1. The closer that $\hat{\varepsilon}$ is to 1.00, the more homogeneous the variances of differences, and hence the closer the data are to being spherical. In a situation in which there are four conditions (as with our data) the lower limit of $\hat{\varepsilon}$ will be 1/(4–1), or .33 (known as the lower-bound estimate of sphericity). The calculated value of $\hat{\varepsilon}$ in the output is .558. This is closer to the lower limit of .33 than it is to the

upper limit of 1 and it therefore represents a substantial deviation from sphericity. We will see how these values are used in the next section.

The main ANOVA

The output also shows the results of the ANOVA for the within-subjects variable. This table can be read much the same as for one-way between-group ANOVA. There is a sum of squares for the repeated-measures effect of **tutor**, which tells us how much of the total variability is explained by the experimental effect. Note the value is 554.125, which is model sum of squares (SS_M) that we calculated in the previous task. There is also an error term (*ssd* in the output), which is the amount of unexplained variation across the conditions of the repeated-measures variable. This is the residual sum of squares (SS_R) that was calculated earlier, and note the value is 1048.375 (which is the same value as calculated). As I explained earlier, these sums of squares are converted into mean squares by dividing by the degrees of freedom. As we saw before, the *df* for the effect of **tutor** (*DFn* in the output) is simply $k-1$, where k is the number of levels of the independent variable. The error *df* (*DFd* in the output) is $(n-1)(k-1)$, where n is the number of participants (or in this case, the number of essays) and k is as before. The *F*-ratio is obtained by dividing the mean squares for the experimental effect (184.708) by the error mean squares (49.923). As with between-group ANOVA, this test statistic represents the ratio of systematic variance to unsystematic variance. The value of $F = 3.70$ (the same as we calculated earlier) is then compared against a critical value for 3 and 21 degrees of freedom. **R** displays the exact significance level for the *F*-ratio. The significance of *F* is .028, which is significant because it is less than the criterion value of .05. We can, therefore, conclude that there was a significant difference between the marks awarded by the four lecturers. However, this main test does not tell us which lecturers differed from each other in their marking.

Although this result seems very plausible, we have learnt that the violation of the sphericity assumption makes the *F*-test inaccurate. We know from Mauchly's test that these data were non-spherical and so we need to make allowances for this violation. The **R** output also contains *p*-values that have been corrected using the Greenhouse–Geisser, and Huynh–Feldt; these are labelled *p[GG]* and *p[HF]*, respectively. For these data the corrections result in the observed *F* being non-significant when using the Greenhouse–Geisser correction (because $p = .06$, which is greater than .05). However, it was noted earlier that this correction is quite conservative, and so can miss effects that genuinely exist. It is, therefore, useful to consult the Huynh–Feldt-corrected *F*-statistic. Using this correction, the *F*-value is still significant because the probability value of .047 is just below the criterion value of .05. So, by this correction we would accept the hypothesis that the lecturers differed in their marking. However, it was also noted earlier that this correction is quite liberal and so tends to accept values as significant when, in reality, they are not significant. This leaves us with the puzzling dilemma of whether or not to accept this *F*-statistic as significant. I mentioned earlier that Stevens (2002) recommends taking an average of the two estimates, and certainly when the two corrections give different results (as is the case here) this is wise advice. If the two corrections give rise to the same conclusion it makes little difference which you choose to report (although if you accept the *F*-statistic as significant it is best to report the conservative Greenhouse–Geisser estimate to avoid criticism!). Although it is easy to calculate the average of the two correction factors and to correct the degrees of freedom accordingly, it is not so easy to then calculate an exact probability for those degrees of freedom. Therefore, should you ever be faced with this perplexing situation (though, to be honest, that's fairly unlikely) I recommend taking an average of the two significance values to give you a rough idea of which correction is giving the most accurate answer. In this case, the average of the two *p*-values is $(.063 + .047)/2 = .055$. Therefore, we should probably go with the Greenhouse–Geisser correction and conclude that the *F*-ratio is non-significant.

These data illustrate how important it is to use a valid critical value of *F*: it can mean the difference between a statistically significant result and a non-significant result. More important, it can mean the difference between making a Type I error and not. Had we not used the corrections for sphericity we would have concluded erroneously that the markers gave significantly different marks. However, I should quantify this statement by saying that this example also highlights how arbitrary it is that we use a .05 level of significance. These two corrections produce significance values only marginally less than or more than .05, and yet they lead to completely opposite conclusions! So, we might be well advised to look at an effect size to see whether the effect is substantive regardless of its significance.

We also saw earlier that a final option, when you have data that violate sphericity, is to use multivariate test statistics (MANOVA) because they do not make this assumption (see O'Brien & Kaiser, 1985).

The interpretation of these results should stop now because the main effect is non-significant. However, we will look at the output for p*ost hoc* tests and contrasts to illustrate how these are displayed in **R**.

Post hoc tests

For *post hoc* tests we can use the *pairwise.t.test()* function. To get a *post hoc* test for the current data, execute:

```
pairwise.t.test(longTutor$Mark, longTutor$Tutor, paired = TRUE, p.adjust.method =
"bonferroni")

Pairwise comparisons using paired t tests

data:  longTutor$Mark and longTutor$Tutor

                Professor Field Professor Smith Professor Scrote
Professor Smith  0.022           -               -
Professor Scrote 1.000           1.000           -
Professor Death  0.261           0.961           0.637

P value adjustment method: bonferroni
```

By looking at the significance values we can see that the only difference between group means is between Professor Field and Professor Smith (p = .022). Looking at the means of these groups, we can see that I give significantly higher marks than Professor Smith. However, there is a rather anomalous result in that there is no significant difference between the marks given by Professor Death and myself (p = .261) even though the mean difference between our marks is higher (11.5) than the mean difference between myself and Professor Smith (4.6). The reason for this result is the sphericity in the data. The interested reader might like to run some correlations between the four tutors' grades. You will find that there is a very high positive correlation between the marks given by Professor Smith and myself (indicating a low level of variability in our data). However, there is a very low correlation between the marks given by Professor Death and myself (indicating a high level of variability between our marks). It is this large variability between Professor Death and myself that has produced the non-significant result despite the average marks being very different (this observation is also evident from the standard errors).

However, the significant contrast should be ignored because of the non-significant main effect (remember that the data did not obey sphericity). The important point to note is that the sphericity in our data has led to some important issues being raised about correction factors, and about applying discretion to your data (it's comforting to know that the computer does not have all of the answers, but it's slightly alarming to realize that this means we have to actually know some of the answers ourselves). In this example we would have to conclude that no significant differences existed between the marks given by different lecturers. However, the ambiguity of our data might make us consider running a similar study with a greater number of essays being marked.

Unfortunately, when using *ezANOVA()* you cannot view the contrasts that you specified, which is why it is better to use *lme()*.

Reporting one-way repeated-measures ANOVA

We could report the main finding as follows:

- ✓ The results show that the mark of an essay was not significantly affected by the lecturer who marked it, $F(1.67, 11.71) = 3.70$, $p > .05$.

If you choose to report the sphericity test as well:

- ✓ Mauchly's test indicated that the assumption of sphericity had been violated, $W = .13$, $p = .043$, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity ($\varepsilon = .56$). The results show that the mark of an essay was not significantly affected by the lecturer who marked it, $F(1.67, 11.71) = 3.70$, $p > .05$.

Remember that because the main ANOVA was not significant we shouldn't report any further analysis.

Using *lme()*

If we want to look at the overall main effect then we need to compare the model containing the predictor from a baseline that includes no predictors other than the intercept. We can specify the baseline model as we did in the chapter:

```
baseline<-lme(Mark ~ 1, random = ~1|Essay/Tutor, data = longTutor, method = "ML")
```

To see the overall effect of **Tutor** we need to add it to the model. To do this we would execute:

```
tutorModel<-lme(Mark ~ Tutor, random = ~1|Essay/Tutor, data = longTutor, method =
 "ML")
```

However, it is quicker to use the *update()* function by executing:

```
 tutorModel<-update(baseline, .~. + Tutor)
```

This command takes the model called *baseline* (which we have already created), and the .~. means keep the outcome and predictors the same as the baseline model. The *+ Tutor* means 'add **Tutor** as a predictor'. Executing this command creates a model called *tutorModel* that includes **Tutor** as a predictor. By comparing these two models (*baseline* and *tutorModel*) we can see whether adding the variable **Tutor** as a predictor significantly improves the model (in other words, by using group means to predict the essay mark – does the model fit the data better than when we don't include this predictor?). To compare the models execute:

```
anova(baseline, tutorModel)
```

```
           Model df    AIC      BIC      logLik    Test    L.Ratio P-Value
baseline       1  4  226.0472  231.9101 -109.0236
tutorModel     2  7  219.4777  229.7379 -102.7389 1 vs 2  12.56946 .0057
```

The output above shows the comparison of the baseline model and the model that includes **Tutor** as a predictor (*tutorModel*). The degrees of freedom between the models change from 4 to 7, which is a difference of 3. This is because **Tutor** has been coded with three contrasts, which means that three parameters (one for each contrast) have been added to the model. The AIC and BIC tell us about the fit of the model (smaller values mean a better fit). The fact that these values are smaller in the final model than the baseline tells us that the fit of the model has improved. The likelihood ratio (*L.Ratio* in the output) tells us whether this improvement in fit is significant, which it is because the *p*-value of .0057 is less than .05. Therefore, **Tutor** is a significant predictor of **Mark**. We can conclude then that the tutor marking the essay had a significant effect on the mark that was awarded, $\chi^2(3) = 12.57$, $p = .006$.

We can further explore the model by executing:

```
summary(tutorModel)
```

```
Formula: ~1 | Tutor %in% Essay
        (Intercept)  Residual
StdDev:    5.999271 0.0305077

Fixed effects: Mark ~ Tutor
                          Value Std.Error DF  t-value p-value
(Intercept)             63.9375 1.1337704 21 56.39369  0.0000
TutorNicevsNasty         2.1875 0.6545826 21  3.34182  0.0031
TutorFieldvsScroteSmith  1.3750 0.9257196 21  1.48533  0.1523
TutorScrotevsSmith      -0.5000 1.6033934 21 -0.31184  0.7582
 Correlation:
                        (Intr) TtrNcN TtrFSS
TutorNicevsNasty        0
TutorFieldvsScroteSmith 0      0
TutorScrotevsSmith      0      0      0
```

The output above shows the parameter estimates for the model. Most important, these include the parameters for the three contrasts that we set. First, when we compare tutors who are considered to be nice (Professors Field, Scrote and Smith) to tutors who are considered to be nasty (Professor Death), essay marks were significantly different, $b = 2.19$, $t(21) = 3.34$, $p = .003$. From the descriptive statistics that we obtained earlier we can see that Professors Field, Scrote and Smith ((68.88 + 65.25 + 64.25)/3 = 66.13) gave significantly higher marks than Professor Death ($M = 57.38$). The second contrast tells us that there was no significant difference between the marks given by Professor Field and the marks given by Professors Smith and Scrote (combined), $b = 1.38$, $t(21) = 1.49$, $p = .152$. The final contrast tells us that there was no significant difference between the marks given by Professor Scrote and those given by Professor Smith, $b = -0.5$, $t(21) = -0.31$, $p = .758$.

Although the contrasts are informative, if there had been no logical set of contrasts to do we might have done *post hoc* tests. We can do this using the *glht()* function. To get *post hoc* tests for the current data, we would execute:

```
postHocs<-glht(tutorModel, linfct = mcp(Tutor = "Tukey"))

summary(postHocs)

confint(postHocs)
```

```
Linear Hypotheses:
                                       Estimate Std. Error z value
Professor Smith - Professor Field == 0   -4.625      3.000  -1.542
Professor Scrote - Professor Field == 0  -3.625      3.000  -1.208
Professor Death - Professor Field == 0  -11.500      3.000  -3.834
Professor Scrote - Professor Smith == 0   1.000      3.000   0.333
Professor Death - Professor Smith == 0   -6.875      3.000  -2.292
Professor Death - Professor Scrote == 0  -7.875      3.000  -2.625
                                       Pr(>|z|)
Professor Smith - Professor Field == 0   0.4124
Professor Scrote - Professor Field == 0  0.6214
Professor Death - Professor Field == 0   <0.001 ***
Professor Scrote - Professor Smith == 0  0.9872
Professor Death - Professor Smith == 0   0.1001
Professor Death - Professor Scrote == 0  0.0432 *

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
                                       Estimate lwr      upr
Professor Smith - Professor Field == 0  -4.6250 -12.3305   3.0805
Professor Scrote - Professor Field == 0 -3.6250 -11.3305   4.0805
Professor Death - Professor Field == 0 -11.5000 -19.2055  -3.7945
Professor Scrote - Professor Smith == 0  1.0000  -6.7055   8.7055
Professor Death - Professor Smith == 0  -6.8750 -14.5805   0.8305
Professor Death - Professor Scrote == 0 -7.8750 -15.5805  -0.1695
```

The output above shows the results of the *post hoc* tests. We can see that Professor Death gave significantly lower essay marks than Professors Field ($p < .001$) and Scrote ($p = .043$). However, none of the other comparisons between tutors were significant.

Effect sizes

If we make sure that we have executed the *rcontrast* function from the book:

```
rcontrast<-function(t, df)
{r<-sqrt(t^2/(t^2 + df))
 print(paste("r = ", r))
 }
```

we can use it to calculate *r* for the contrasts we did by executing these commands (the values of *t* and *df* come from the output for *summary(tutorModel)*:

```
rcontrast(3.34182, 21)
rcontrast(1.48533, 21)
rcontrast(-0.31184,21)
```

```
[1] "r =  0.589213430732642"
[1] "r =  0.308333624869346"
[1] "r =  0.0678920554095802"
```

These show that the difference between the nice tutors and Professor Death was a large effect (r = .59), between Professors Smith and Scrote (combined) and Professor Field was a medium effect (r = .31), but between Professors Smith and Scrote a small effect (r = .07).

Reporting the results

As we have made a multilevel model, we need to report our results differently than when we use *ezANOVA*. Marks were significantly affected by the tutor who marked the essay, $\chi^2(3) = 12.57$, p = .006. Orthogonal contrasts revealed that tutors who were considered to be generous markers (Professors Field, Scrote and Smith) gave significantly higher essay marks than Professor Death, b = 2.19, $t(21) = 3.34$, p = .003; there was no significant differences between the marks given by Professor Field and Professors Smith and Scrote (combined), b = 1.38, $t(21) = 1.49$, p = .152 or between the marks given by Professor Scrote and those given by Professor Smith, b = –0.5, $t(21)$ = –0.31, p = .758.

Robust ANOVA

Just for practice, let's also run a robust ANOVA on these data. The functions for the robust methods need the data to be in wide format rather than long. However, the data we originally loaded in were in this format so we can simply reuse these (remember they are stored in an object called *tutorData*).

We want only the scores, so we need to get rid of the **Essay** variable. The **Essay** variable is in the first column, so we could create a new data frame (*tutordata2*) that excludes this first column by executing:

```
tutorData2<-tutorData[, -c(1)]
```

Assuming we are happy with the default level of trimming, we can do one-way repeated measures ANOVA based on trimmed means by executing:

```
rmanova(tutorData2)
```

If we wanted to do a one-way repeated measured ANOVA based on 2000 bootstrap samples, then we could execute:

```
rmanovab(tutorData2, nboot = 2000)
```

rmanova()	Rmanovab()
[1] "The number of groups to be compared is"	[1] "The number of groups to be compared is"
[1] 4	[1] 4
$test	$teststat
[1] 2.348734	[1] 2.348734
$df	$crit
[1] 1.994226 9.971132	[1] 5.625
$siglevel	
[1] 0.1460211	
$tmeans	
[1] 68.50000 63.83333 65.50000 58.16667	
$ehat	

```
[1] 0.5322935


$etil
[1] 0.6647422
```

The table above shows the output of both these commands. For *rmanova()* (left-hand side of the table) we are given a test statistic, *F*, for the effect of tutor (*$test*), the degrees of freedom (*$df*), the *p*-value (*$siglevel*) and the group means(*$tmeans*). Given that the significance level (0.146) is greater than .05, we can say that there was no significant differences in marks when the essays were marked by the different tutors, $F(1.99, 9.97) = 2.35$, $p = .146$. (Note that I have reported the test statistic, its degrees of freedom and the *p*-value, which you can find in the output.)

The output of *rmanovab()* (right-hand side of the table) tells us much the same things but we get only a test statistic (*$teststat*) and the critical value for this statistic at a .05 level of significance (*$crit*). If the test statistic is significant then the test statistic should be greater than the critical value. In this case, the test statistic (2.35) is less than the critical value (5.63), indicating no significant differences in marks between tutors, $F = 2.35$, $F_{crit} = 5.63$, $p > .05$. Both of these robust methods yield non-significant results (unlike the original ANOVA).

The *post hoc* tests for each analysis are conducted using the same command structure. Assuming you leave the default options, to run *post hoc* tests based on a 20% trimmed mean, we execute:

```
rmmcp(tutorData2)
```

```
$test
     Group Group       test      p.value  p.crit        se
[1,]     1     2 14.0532321 3.282054e-05 0.00851 0.2727724
[2,]     1     3  0.8549393 4.316329e-01 0.02500 4.0938579
[3,]     1     4  1.4101496 2.175644e-01 0.01020 7.5642093
[4,]     2     3 -0.4022510 7.041190e-01 0.05000 3.3146800
[5,]     2     4  0.9388455 3.909130e-01 0.01690 5.8582588
[6,]     3     4  1.2214564 2.763582e-01 0.01270 5.1850671

$psihat
     Group Group    psihat   ci.lower  ci.upper
[1,]     1     2  3.833333   2.682422  4.984244
[2,]     1     3  3.500000 -13.773252 20.773252
[3,]     1     4 10.666667 -21.249070 42.582404
[4,]     2     3 -1.333333 -15.318993 12.652326
[5,]     2     4  5.500000 -19.217805 30.217805
[6,]     3     4  6.333333 -15.544067 28.210734

$con
     [,1]
[1,]    0

$num.sig
[1] 1
```

In the output above, if the value of *p.value* is less than the critical value (*p.crit*) and the confidence interval does not cross zero then the comparison is significant. The columns labeled *group* tells you which groups are being compared (the numbers relate to columns in the dataframe).

- ✓ [1,] tests the difference between Professor Field and Professor Smith. This contrast is significant because *p.value* (.000) is less than *p.crit* (.008) and the confidence interval does not cross zero.
- ✓ [2] tests the difference between Professor Field and Professor Scrote. This contrast is not significant because *p.value* (.432) is greater than *p.crit* (.025) and the confidence interval crosses zero.
- ✓ [3] tests the difference between Professor Field and Professor Death. This contrast is not significant because *p.value* (.218) is greater than *p.crit* (.010) and the confidence interval crosses zero.

- ✓ [4] tests the difference between Professor Smith and Professor Scrote. This contrast is not significant because *p.value* (.704) is greater than *p.crit* (.050) and the confidence interval crosses zero.
- ✓ [5] tests the difference between Professor Smith and Professor Death. This contrast is not significant because *p.value* (.391) is greater than *p.crit* (.017) and the confidence interval crosses zero.
- ✓ [6] tests the difference between Professor Scrote and Professor Death. This contrast is not significant because *p.value* (.276) is greater than *p.crit* (.013) and the confidence interval crosses zero.

We could report that there was a significant difference between the marks given by Professor Field and Professor Smith, Ψ = 3.83 (2.68, 4.98), $p < .05$. However, there was no significant difference between the essay marks given by Professor Field and Professor Scrote, Ψ = 3.50 (–13.77, 20.77), $p > .05$ or Professor Death, Ψ = 10.67 (–21.25, 42.58), $p > .05$. Similarly, there was no significant difference between the marks given by Professor Smith and Professors Scrote, Ψ = -1.33 (-15.32, 12.65), $p > .05$, or Death, Ψ = 5.50 (–19.22, 30.22), $p > .05$, or between Professor Scrote and Professor Death, Ψ = 6.33 (–15.54, 28.21), $p < .05$. Note that in each case I have reported *psihat* and its confidence interval.

To conduct *post hoc* tests based on trimmed means and a bootstrap, execute:

```
pairdepb(tutorData2, nboot = 2000)
```

```
$test
     Group Group       test       se
[1,]     1     2  4.2092178 1.108678
[2,]     1     3  0.9192771 3.263434
[3,]     1     4  1.8441848 5.603198
[4,]     2     3 -0.5380311 3.097714
[5,]     2     4  1.1977475 4.731103
[6,]     3     4  1.5978885 4.589390

$psihat
     Group Group    psihat ci.lower ci.upper
[1,]     1     2  4.666667     -Inf      Inf
[2,]     1     3  3.000000     -Inf      Inf
[3,]     1     4 10.333333     -Inf      Inf
[4,]     2     3 -1.666667     -Inf      Inf
[5,]     2     4  5.666667     -Inf      Inf
[6,]     3     4  7.333333     -Inf      Inf

$crit
[1] Inf
```

The output above shows the *post hoc* tests based on trimmed means and a bootstrap (*pairdepb*). The interpretation of these results is similar to that for the trimmed means. If the value of *test* is greater than the critical value (*$crit*) and the confidence interval does not cross zero then the contrast is significant. Therefore, we're comparing each value of test against Inf (which means infinite); as you can see, all values of test are smaller than this value and all their confidence intervals cross zero so we can conclude that none of the groups differ significantly.

We could again report that (note that the values and confidence intervals for *psihat* have changed): there was no significant difference between the marks given by Professor Field and Professor Smith, Ψ = 4.67 (–Inf, Inf), $p > .05$, Professor Scrote, Ψ = 3.00 (–Inf, Inf), $p > .05$, or Professor Death, Ψ = 10.33 (–Inf, Inf), $p > .05$. Similarly, there was no significant difference between the marks given by Professor Smith and Professor Scrote Ψ = –1.67 (–Inf, Inf), $p > .05$, or Death, Ψ = 5.66 (–Inf, Inf), $p > .05$, or between Professor Scrote and Professor Death Ψ = 7.33 (–Inf, Inf), $p < .05$. Note that in each case I have reported *psihat* and its confidence interval.

Task 3

Imagine I wanted to look at the effect alcohol has on the roving eye. The 'roving eye' effect is the propensity of people in relationships to 'eye up' members of the opposite sex. I took 20 men and fitted them with incredibly sophisticated glasses that could track their eye movements and record

both the movement and the object being observed (this is the point at which it should be apparent that I'm making it up as I go along). Over four different nights I plied these poor souls with 1, 2, 3 or 4 pints of strong lager in a nightclub. Each night I measured how many different women they eyed up (a woman was categorized as having been eyed up if the man's eye moved from her head to her toe and back up again). To validate this measure we also collected the amount of dribble on the man's chin while looking at a woman. The data are in the file **RovingEye.dat**. Analyse them with a one-way ANOVA.

First of all we need to load the data:

```
rovingData<-read.delim("RovingEye.dat", header = TRUE)
```

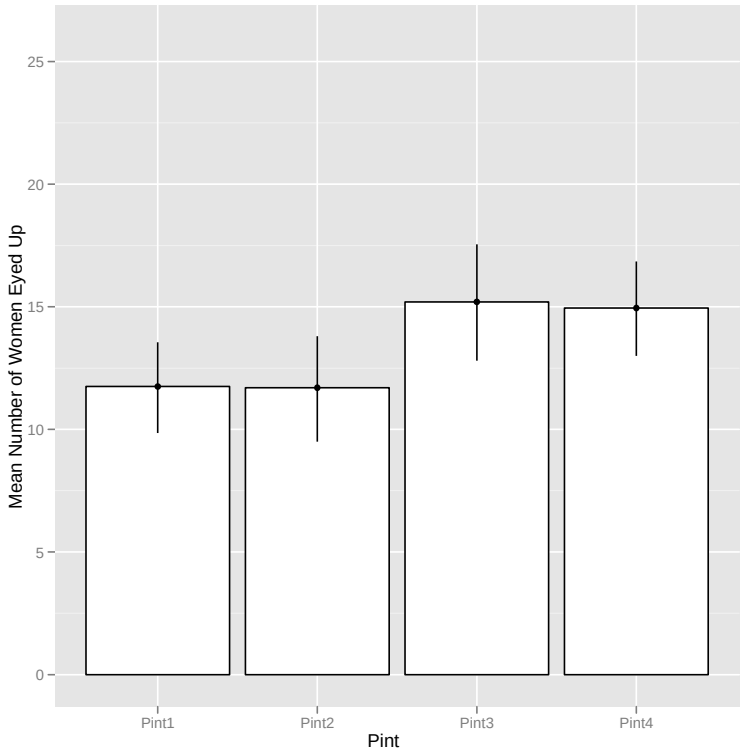
The data were originally entered into **R** in the wide format, but we need them to be in the long format for these analyses. To convert the data into the long format we could execute:

```
longRoving<-melt(rovingData, id = "Participant", measured = c("pint1", "pint2",
 "pint3", "pint4"))
names(longRoving)<-c("Participant", "Pint", "Number_of_Women")

longRoving$Pint<-factor(longRoving$Pint, labels = c("Pint1", "Pint2", "Pint3",
 "Pint4"))
longRoving<-longRoving[order(longRoving$Participant),]
```

We can now plot an error bar graph by executing:

```
rovingBar <- ggplot(longRoving, aes(Pint, Number_of_Women))
rovingBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black")
 + stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Pint", y =
 "Mean Number of Women Eyed Up")
```

The resulting graph should look like this:



This error bar chart of the roving eye data shows the mean number of women that were eyed up after different doses of alcohol. It's clear from this chart that the mean number of women is pretty similar between 1 and 2 pints, and for 3 and 4 pints, but there is a jump after 2 pints.

Next we need to set some orthogonal contrasts so that we can look at Type III sums of squares. To set orthogonal contrasts we can first create variables representing each contrast and then bind these variables together and set them as the contrast for **Pint**:

```
Pint1vsMore<-c(3, -1, -1, -1)
Pint4vsPint2andPint3<-c(0, -1, -1, 2)
Pint2vsPint3<-c(0, 1, -1, 0)
contrasts(longRoving$Pint)<-cbind(Pint1vsMore, Pint4vsPint2andPint3, Pint2vsPint3)
```

Next we can run an *ezANOVA* by executing:

```
rovingModel<-ezANOVA(data = longRoving, dv = .(Number_of_Women), wid = .(Participant),
 within = .(Pint), type = 3, detailed = TRUE)

rovingModel

$ANOVA
       Effect DFn DFd     SSn   SSd          F            p  p<.05       ges
1 (Intercept)   1  19 14364.8 915.7 298.057442 4.532206e-13     *   0.8875433
2        Pint   3  57   225.1 904.4   4.728992 5.144516e-03     *   0.1100626

$`Mauchly's Test for Sphericity`
  Effect         W          p        p<.05
2   Pint    0.4769123     0.02246469     *

$`Sphericity Corrections`
   Effect       GGe      p[GG]    p[GG]<.05       HFe       p[HF]       p[HF]<.05
2   Pint    0.7450699 0.01143403          *   0.849085  0.008241601          *
```

If we look at the part of the output above that contains Mauchly's test, we hope to find that it's non-significant if we are to assume that the condition of sphericity has been met. However, the significance value (.022) is less than the critical value of .05, so we accept that the assumption of sphericity has been violated.

The main ANOVA

The output above also shows the main result of the ANOVA. The significance of *F* is .005, which is significant because it is less than the criterion value of .05. We can, therefore, conclude that alcohol had a significant effect on the average number of women who were eyed up. However, this main test does not tell us which quantities of alcohol made a difference to the number of women eyed up.

This result is all very nice, but as yet we haven't done anything about our violation of the sphericity assumption. This table contains an additional row giving the corrected values of *F* for two different types of adjustment (Greenhouse–Geisser and Huynh–Feldt). First we decide which correction to apply and to do this we need to look at the estimates of sphericity: if the Greenhouse–Geisser and Huynh–Feldt estimates are less than 0.75 we should use Greenhouse–Geisser, and if they are above 0.75 we use Huynh–Feldt. We discovered in the book that, based on these criteria, we should use Huynh–Feldt here. Using this corrected value we still find a significant result because the observed *p* (.008) is still less than the criterion of .05.

The main effect of alcohol doesn't tell us anything about which doses of alcohol produced different results to other doses. So, we might do some *post hoc* tests as well. We can do some *post hoc* tests by executing:

```
 pairwise.t.test(longRoving$Number_of_Women,    longRoving$Pint,    paired    =    TRUE,
p.adjust.method = "bonferroni")

Pairwise comparisons using paired t tests

data:  longRoving$Number_of_Women and longRoving$Pint

      Pint1 Pint2 Pint3
Pint2 1.000 -     -
Pint3 0.136 0.038 -
Pint4 0.242 0.202 1.000

P value adjustment method: bonferroni
```

The output above shows the table from **R** that contains these tests. We read down each column and look for values less than .05. By looking at the significance values we can see that the only difference between condition means is between 2 and 3 pints of alcohol.

We can view the means and standard deviations by executing:

```
ezStats(data = longRoving, dv = .(Number_of_Women), wid = .(Participant), within =
.(Pint), type = 3)
```

```
Warning: Converting "Participant" to factor for ANOVA.
Note: model has only an intercept; equivalent type-III tests substituted.
   Pint  N  Mean       SD     FLSD
1 Pint1 20 11.75 4.314907 2.522365
2 Pint2 20 11.70 4.657761 2.522365
3 Pint3 20 15.20 5.800181 2.522365
4 Pint4 20 14.95 4.673272 2.522365
```

Interpreting and writing the result

We could report the main finding as follows:

- ✓ Mauchly's test indicated that the assumption of sphericity had been violated, W = .48, p = .022, therefore degrees of freedom were corrected using Huynh–Feldt estimates of sphericity (ε = .85). The results show that the number of women eyed up was significantly affected by the amount of alcohol drunk, $F(2.55, 48.40) = 4.73$, $p < .05$, r = .40. Bonferroni *post hoc* tests revealed a significant difference in the number of women eyed up only between 2 and 3 pints, $p < .05$). No other comparisons were significant (all ps > .05).

Using lme()

If we want to look at the overall main effect then we need to compare the model containing the predictor from a baseline that includes no predictors other than the intercept. We can specify the baseline model as we did in the chapter:

```
baseline<-lme(Number_of_Women ~ 1, random = ~1|Participant/Pint, data =
 longRoving, method = "ML")
```

To see the overall effect of **Pint** we need to add it to the model. To do this we would execute:

```
 rovingModel<-lme(Number_of_Women ~ Pint, random = ~1|Participant/Pint, data =
longRoving, method = "ML")
```

However, it is quicker to use the *update()* function by executing:

```
 rovingModel<-update(baseline, .~. + Pint)
```

This command takes the model called *baseline* (which we have already created), and the .~. means keep the outcome and predictors the same as the baseline model. The *+ Pint* means 'add **Pint** as a predictor'. Executing this command creates a model called *rovingModel* that includes **Pint** as a predictor. By comparing these two models (*baseline* and *rovingModel*) we can see whether adding the variable **Pint** as a predictor significantly improves the model (in other words, by using group means to predict the number of women eyed up – does the model fit the data better than when we don't include this predictor?). To compare the models execute:

```
anova(baseline, rovingModel)
```

```
           Model df       AIC      BIC     logLik    Test     L.Ratio  p-value
baseline       1  4     487.6204 497.1485  -239.8102
rovingModel    2  7     480.2849 496.9591  -233.1425 1 vs 2  13.33552  0.004
```

The output above shows the comparison of the baseline model and the model that includes **Pint** as a predictor (*rovingModel*). The degrees of freedom between the models change from 4 to 7, which is a difference of 3. This is because **Pint** has been coded with three contrasts, which means that three parameters (one for each contrast) have been added to the model. The AIC and BIC tell us about the fit of the model (smaller values mean a better fit). The fact that these values are smaller in the final model than the baseline tells us that the fit of the model has got better. The likelihood ratio (*L.Ratio* in the output) tells us whether this improvement in fit is significant, which it is because the p-value of .004 is less than .05. Therefore, the amount drunk is a significant predictor of the number of women

eyed up. We can conclude, then, that the number of pints consumed had a significant effect on the number of women that the men eyed up, $\chi^2(3) = 13.34$, $p = .004$.

We can further explore the model by executing:

```
summary(rovingModel)
```

```
Formula: ~1 | Pint %in% Participant
        (Intercept) Residual
StdDev:    3.596012 1.463573

Fixed effects: Number_of_Women ~ Pint
                          Value Std.Error DF   t-value p-value
(Intercept)              13.40 0.7761653 57 17.264363  0.0000
PintPint1vsMore          -0.55 0.2571209 57 -2.139071  0.0367
PintPint4vsPint2andPint3  0.50 0.3636239 57  1.375047  0.1745
PintPint2vsPint3         -1.75 0.6298151 57 -2.778593  0.0074
 Correlation:
                          (Intr) PntP1M PP4P2P
PintPint1vsMore           0
PintPint4vsPint2andPint3 0      0
PintPint2vsPint3         0      0      0
```

The output above shows the parameter estimates for the model. Most important, these include the parameters for the three contrasts that we set. First, when we compare 1 pint to more than 1 pint (2, 3 and 4 pints) the number of women eyed up were significantly different, $b = -0.55$, $t(57) = -2.14$, $p = .04$. From the descriptive statistics that we obtained earlier we can see that 2 pints, 3 pints and 4 pints ((11.70 + 15.20 + 14.95)/3 = 13.95) resulted in significantly more women being eyed up than 1 pint ($M = 11.75$). The second contrast tells us that there was no significant difference between drinking 4 pints and drinking 2 or 3 pints (combined), $b = 0.50$, $t(57) = 1.38$, $p = .17$. The final contrast tells us that there was a significant difference between drinking 2 pints and drinking 3 pints, $b = -1.75$, $t(57) = -2.78$, $p = .00$; looking at the means, we can see that after drinking 3 pints men eyed up significantly more women than after drinking 2 pints.

Although the contrasts are informative, if there had been no logical set of contrasts to do we might have done *post hoc* tests. We can, do this using the *glht()* function. To get *post hoc* tests for the current data, we would execute:

```
postHocs<-glht(rovingModel, linfct = mcp(Pint = "Tukey"))
summary(postHocs)
confint(postHocs)
```

```
 Linear Hypotheses:
                    Estimate Std. Error z value Pr(>|z|)
 Pint2 - Pint1 == 0   -0.050      1.228  -0.041   1.0000
 Pint3 - Pint1 == 0    3.450      1.228   2.810   0.0257 *
 Pint4 - Pint1 == 0    3.200      1.228   2.606   0.0452 *
 Pint3 - Pint2 == 0    3.500      1.228   2.851   0.0226 *
 Pint4 - Pint2 == 0    3.250      1.228   2.647   0.0409 *
 Pint4 - Pint3 == 0   -0.250      1.228  -0.204   0.9970

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

 Linear Hypotheses:
                    Estimate   lwr      upr
 Pint2 - Pint1 == 0 -0.05000 -3.20301  3.10301
 Pint3 - Pint1 == 0  3.45000  0.29699  6.60301
 Pint4 - Pint1 == 0  3.20000  0.04699  6.35301
 Pint3 - Pint2 == 0  3.50000  0.34699  6.65301
 Pint4 - Pint2 == 0  3.25000  0.09699  6.40301
 Pint4 - Pint3 == 0 -0.25000 -3.40301  2.90301
```

The output above shows the results of the *post hoc* tests. We can see that there was a significant difference between: (1) 1 pint and 3 pints ($p < .05$); (2) 4 pints and 1 pint ($p < .05$); (3) 2 pints and 3 pints ($p < .05$); and (4) 4 pints and 2 pints ($p < .05$). However, there was no significant difference between drinking 1 pint and 2 pints ($p = 1.00$), or between drinking 3 pints and 4 pints ($p = 1.00$).

Effect sizes

If we make sure that we have executed the *rcontrast* function from the book, we can use it to calculate *r* for the contrasts we did by executing these commands (the values of *t* and *df* come from the output for *summary(rovingModel)*:

```
rcontrast(-2.139071, 57)
rcontrast(1.375047, 57)
rcontrast(-2.778593, 57)
  [1] "r =  0.272596820779416"
  [1] "r =  0.179181836930639"
  [1] "r =  0.345385212987429"
```

which show that the difference between drinking 1 pint and more than 1 pint (2, 3 or 4 pints) was a medium effect (r = .27), between drinking 4 pints and less than 4 pints (1, 2 or 3 pints) was a small to medium effect (r = .18), and between 2 and 3 pints was a medium to large effect (r = .35).

Reporting the Results

As we have made a multilevel model, we need to report our results differently than when we use ezANOVA. The results show that the number of women eyed up was significantly affected by the amount of alcohol drunk, $\chi^2(3) = 13.34$, $p = .004$. Orthogonal contrasts revealed that drinking more than 1 pint resulted in significantly more women being eyed up, $b = -0.55$, $t(57) = -2.14$, $p = .04$; there was no significant difference between drinking 4 pints and drinking 2 and 3 pints (combined), $b = 0.50$, $t(57) = 1.38$, $p = .17$. However, significantly more women were eyed up after 3 pints than after 2 pints, $b = -1.75$, $t(57) = -2.78$, $p < .01$.

Task 4

In the previous chapter we came across the beer-goggles effect, a severe perceptual distortion after imbibing alcohol that makes previously unattractive people suddenly become the hottest thing since Spicy Gonzalez's extra-hot Tabasco-marinated chillies. Imagine we followed up the fabricated example from the previous chapter to look at whether the beer-goggles effect is made worse by the fact that it usually occurs in clubs that have dim lighting. We took a sample of 26 men (because the effect is stronger in men) and gave them various doses of alcohol over four different weeks (0 pints, 2 pints, 4 pints and 6 pints of lager). This is our first independent variable. Each week (and, therefore, in each state of drunkenness) participants were asked to select a mate in a normal club (that had dim lighting) and then select a second mate in a specially designed club that had bright lighting. As such, the second independent variable was whether the club had dim or bright lighting. The outcome measure was the attractiveness of each mate as assessed by a panel of independent judges. To recap, all participants took part in all levels of the alcohol consumption variable, and selected mates in both brightly and dimly lit clubs. The data are in the file **BeerGogglesLighting.dat**. Analyse them with a two-way repeated-measures ANOVA.

First of all, remember to read in the data and give it a sensible name:

```
gogglesData<-read.delim("BeerGogglesLighting.dat", header = TRUE)
```

We then need to reshape the data as we did in the book chapter, we can do this by executing:

```
longGoggles<-melt(gogglesData, id = "Participant", measured = c( "dim0", "bright0",
"dim2", "bright2", "dim4", "bright4", "dim6", "bright6"))
names(longGoggles)<-c("Participant", "Groups", "Attractiveness")
```

We then need to create separate columns for the two variables **Lighting** and **Pints**. We can do this by executing:

```
longGoggles$Lighting<-gl(2, 26, 208, labels = c("Dim", "Bright"))
longGoggles$Pints<-gl(4,52, 208, labels = c("0", "2", "4", "6"))
```

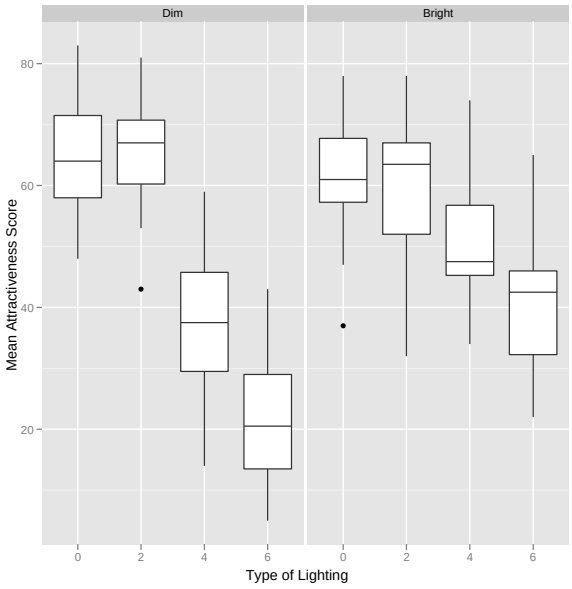
```
longGoggles<-longGoggles[order(longGoggles$Participant),]
```

The data now look like this (I have only put in a small section to save space):

```
Participant  Groups Attractiveness Lighting  Pints
1            dim0             58      Dim     0
1            bright0          65   Bright     0
1            dim2             65      Dim     2
1            bright2          65   Bright     2
1            dim4             44      Dim     4
1            bright4          50   Bright     4
```

Let's create a boxplot of the data. We can do this using the reshaped data (*longGoggles*):

```
gogglesBoxplot <- ggplot(longGoggles, aes(Pints, Attractiveness))
gogglesBoxplot + geom_boxplot() + facet_wrap(~Lighting, nrow = 1) + labs(x = "Type of
Lighting", y = "Mean Attractiveness Score")
```

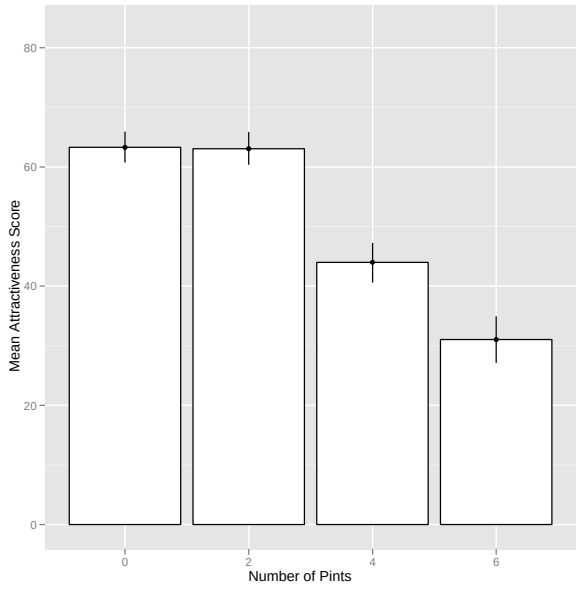


This box plot displays the mean attractiveness of the partner selected (with error bars) in dim and brightly lit clubs after the different doses of alcohol. The chart shows that in both dim and brightly lit clubs there is a tendency for men to select less attractive mates as they consume more and more alcohol.

We could draw a bar graph of the effect of the number of pints consumed by executing:

```
pintsBar <- ggplot(longGoggles, aes(Pints, Attractiveness))
pintsBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Black")
+ stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Number of
Pints", y = "Mean Attractiveness Score")
```

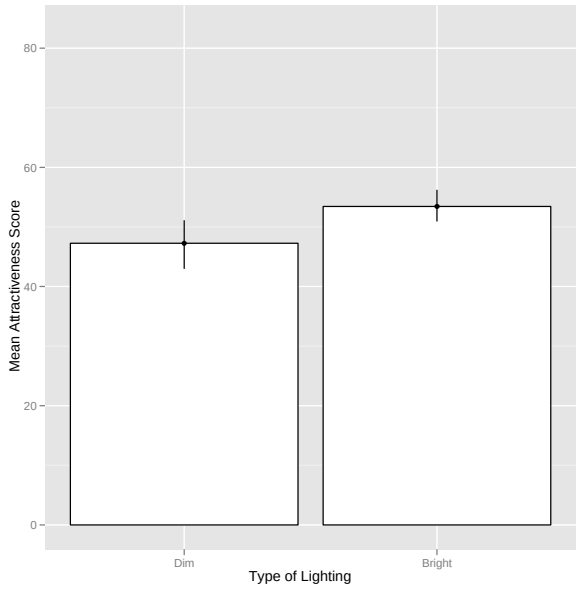
The resulting graph should look like this:



To plot an error bar graph of the effect of lighting on the attractiveness of the date chosen we would execute:

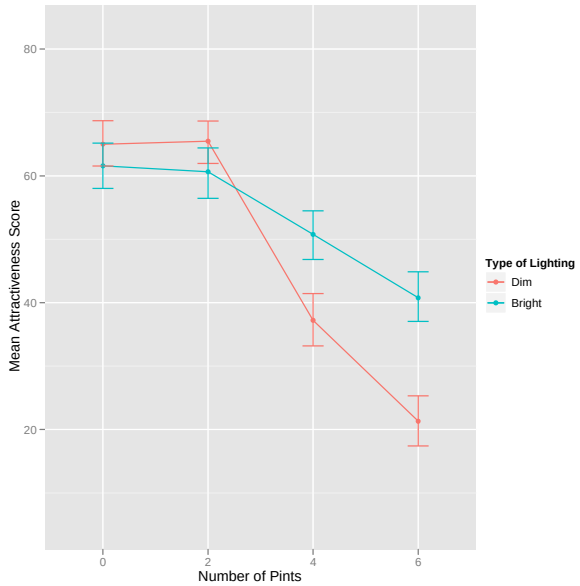
```
lightingBar <- ggplot(longGoggles, aes(Lighting, Attractiveness))
 lightingBar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour =
"Black") + stat_summary(fun.data = mean_cl_boot, geom = "pointrange") + labs(x = "Type
of Lighting", y = "Mean Attractiveness Score")
```

The resulting graph should look like this:



To plot an interaction graph, to look at the interaction between the number of pints drunk and the type of lighting on the attractiveness of the woman chosen, we would execute:

```
gogglesInt <- ggplot(longGoggles, aes(Pints, Attractiveness, colour = Lighting))
gogglesInt + stat_summary(fun.y = mean, geom = "point") + stat_summary(fun.y = mean,
geom = "line", aes(group= Lighting)) + stat_summary(fun.data = mean_cl_boot, geom =
"errorbar", width = 0.2) + labs(x = "Type of Drink", y = "Mean Attractiveness Score",
colour = "Type of Lighting")
```



This graph is the same as the boxplot above; it displays the mean attractiveness of the partner selected (with error bars) in dim and brightly lit clubs after the different doses of alcohol. The chart shows that in both dim and brightly lit clubs there is a tendency for men to select less attractive mates as they consume more and more alcohol.

We can request some descriptive statistics by executing:

```
options(digits = 3)
by(longGoggles$Attractiveness, list(longGoggles$Pints, longGoggles$Lighting),
stat.desc, basic = FALSE)
by(longGoggles$Attractiveness, longGoggles$Pints, stat.desc, basic = FALSE)
by(longGoggles$Attractiveness, longGoggles$Lighting, stat.desc, basic = FALSE)
 options(digits = 7)
```

The resulting output below shows the means for all conditions in a table. These means correspond to those plotted in the graphs.

```
> by(longGoggles$Attractiveness, list(longGoggles$Pints, longGoggles$Lighting), stat.desc, basic = FALSE)
: 0
: Dim
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      64.000       65.000        2.021        4.163      106.240       10.307        0.159
------------------------------------------------------------------------------
: 2
: Dim
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      67.000       65.462        1.718        3.538       76.738        8.760        0.134
------------------------------------------------------------------------------
: 4
: Dim
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      37.500       37.231        2.131        4.388      118.025       10.864        0.292
------------------------------------------------------------------------------
: 6
: Dim
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      20.500       21.308        2.093        4.311      113.902       10.672        0.501
------------------------------------------------------------------------------
: 0
: Bright
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      61.000       61.577        1.903        3.920       94.174        9.704        0.158
------------------------------------------------------------------------------
: 2
: Bright
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      63.500       60.654        2.089        4.302      113.435       10.651        0.176
------------------------------------------------------------------------------
: 4
: Bright
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      47.500       50.769        2.028        4.178      106.985       10.343        0.204
------------------------------------------------------------------------------
: 6
: Bright
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
      42.500       40.769        2.113        4.352      116.105       10.775        0.264

>by(longGoggles$Attractiveness, longGoggles$Pints, stat.desc, basic = FALSE)
longGoggles$Pints: 0
```

```
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      64.000        63.288         1.395        2.801      101.229        10.061         0.159 
------------------------------------------------------------------------------------------------ 
longGoggles$Pints: 2
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      65.000        63.058         1.381        2.772       99.114         9.956         0.158 
------------------------------------------------------------------------------------------------ 
longGoggles$Pints: 4
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      45.500        44.000         1.738        3.489      157.020        12.531         0.285 
------------------------------------------------------------------------------------------------ 
longGoggles$Pints: 6
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      31.000        31.038         2.006        4.028      209.293        14.467         0.466 


> by(longGoggles$Attractiveness, longGoggles$Lighting, stat.desc, basic = FALSE)
longGoggles$Lighting: Dim
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      49.000        47.250         2.101        4.167      459.180        21.428         0.454 
------------------------------------------------------------------------------------------------ 
longGoggles$Lighting: Bright
      median          mean       SE.mean CI.mean.0.95          var       std.dev      coef.var 
      53.500        53.442         1.304        2.585      176.715        13.293         0.249 
> options(digits = 7)
```

Before we conduct the ANOVA, we need to set some orthogonal contrasts so that we can use type III sums of squares. We could set some such contrasts as follows:

```
NonevsAlcohol<-c(4, -1, -1, -1)
MaxvsLess<-c(0, -1, -1, 2)
TwovsFour<-c(0, 1, -1,0)

DimvsLight<-c(1, -1)

contrasts(longGoggles$Pints)<-cbind(NonevsAlcohol, MaxvsLess, TwovsFour)
```

Next we can run the two-way ANOVA using *ezANOVA* by executing:

```
gogglesModel<-ezANOVA(data = longGoggles, dv = .(Attractiveness), wid =
.(Participant),  within = .(Pints, Lighting), type = 3, detailed = TRUE)
options(digits = 3)
gogglesModel
```

```
$ANOVA
          Effect DFn DFd    SSn  SSd      F        p p<.05    ges
1    (Intercept)   1  25 527225 3282 4016.2 3.91e-29     * 0.9614
2          Pints   3  75  38592 9243  104.4 1.06e-26     * 0.6461
3       Lighting   1  25   1994 2128   23.4 5.65e-05     * 0.0862
4 Pints:Lighting   3  75   5765 6487   22.2 2.14e-10     * 0.2143

$`Mauchly's Test for Sphericity`
          Effect     W     p p<.05
2          Pints 0.820 0.454
4 Pints:Lighting 0.898 0.768

$`Sphericity Corrections`
          Effect   GGe    p[GG] p[GG]<.05   HFe    p[HF] p[HF]<.05
2          Pints 0.873 1.32e-23         * 0.984 2.56e-26         *
4 Pints:Lighting 0.936 7.18e-10         * 1.067 2.14e-10         *
```

The lighting variable had only two levels (dim or bright) and so the assumption of sphericity doesn't apply and **R** doesn't produce a significance value. However, for the effects of alcohol consumption and the interaction of alcohol consumption and lighting, we do have to look at Mauchly's test. The significance values are both above .05 (they are 0.454 and 0.768, respectively) and so we know that the assumption of sphericity has been met for both alcohol consumption and the interaction of alcohol consumption and lighting.

The output above also shows the main ANOVA summary table. The main effect of lighting is shown by the *F*-ratio in the row labelled **Lighting**. The significance of this value is well below the usual cut-off point of .05. We can conclude that average attractiveness ratings were significantly affected by whether mates were selected in a dim or well-lit club. We can easily interpret this result further because there were only two levels: attractiveness ratings were higher in the well-lit clubs (look back at the error bar graph that we plotted earlier), so we could conclude that when we ignore how much alcohol was consumed, the mates selected in well-lit clubs were significantly more attractive than those chosen in dim clubs.

The main effect of alcohol consumption is shown by the *F*-ratio in the row labelled *Pints*. The probability associated with this *F*-ratio is reported as .000 (i.e. $p < .001$), which is well below the critical value of .05. We can conclude that there was a significant main effect of the amount of alcohol consumed on the attractiveness of the mate selected. We know that generally there was an effect, but without further tests (e.g. *post hoc* comparisons) we can't say exactly which doses of alcohol had the most effect. If we look back at the error bar graph that we plotted earlier, we can see that when you ignore the lighting in the club, the attractiveness of mates is similar after no alcohol and 2 pints of lager but starts to rapidly decline at 4 pints and continues to decline after 6 pints.

We can look at some *post hoc* tests for the main affect of alcohol (**Pints**) by executing:

```
pairwise.t.test(longGoggles$Attractiveness, longGoggles$Pints, paired = TRUE,
p.adjust.method = "bonferroni")
 options(digits = 7)
```

```
  0       2       4
2 1       -       -
4 4.8e-09 4.6e-10 -
6 < 2e-16 < 2e-16 3.2e-06

P value adjustment method: bonferroni
```

The above output shows the resulting *post hoc* tests. In this example I've chosen a Bonferroni correction. The mean attractiveness was significantly higher after no pints than it was after 4 pints and 6 pints (both *p*s are less than .001). We can also see that the mean attractiveness after 2 pints was significantly higher than after 4 pints and 6 pints (again, both *p*s are less than .001). Finally, the mean attractiveness after 4 pints was significantly higher than after 6 pints ($p < .001$). So, we can conclude that the beer goggles effect doesn't kick in until after 2 pints, and that it has an ever-increasing effect (well, up to 6 pints at any rate!).

The interaction effect is shown by the *F*-ratio in the row labelled *Pints:Lightening*. The resulting *F*-ratio is 22.22 (1921.81/86.50), which has an associated probability value of less than .001. As such, there is a significant interaction between the amount of alcohol consumed and the lighting in the club on the attractiveness of the mate selected.

We could look at some p*ost hoc* tests for the interaction between **Lighting** and **Pints** by executing:

```
pairwise.t.test(longGoggles$Attractiveness, longGoggles$Groups, paired = TRUE,
p.adjust.method = "bonferroni")
 options(digits = 7)
```

```
        Pairwise comparisons using paired t tests

data:  longGoggles$Attractiveness and longGoggles$Groups

        dim0    bright0 dim2    bright2 dim4    bright4 dim6
bright0 1.0000  -       -       -       -       -       -
dim2    1.0000  1.0000  -       -       -       -       -
bright2 1.0000  1.0000  1.0000  -       -       -       -
dim4    9.6e-07 2.3e-07 4.8e-12 1.3e-08 -       -       -
bright4 0.0100  0.0286  0.0006  0.2057  0.0006  -       -
dim6    1.1e-13 6.0e-12 9.3e-13 1.3e-11 0.0024 3.3e-08 -
bright6 2.4e-07 2.1e-07 1.0e-08 2.7e-08 1.0000 0.0617  2.3e-06

P value adjustment method: bonferroni
```

The resulting output above shows that, when in dim lighting, attractiveness of the partner chosen is significantly reduced after 4 pints and 6 pints compared with no pints.

Writing the result

We can report the three effects from this analysis as follows:

- ✓ The results show that the attractiveness of the mates selected was significantly lower when the lighting in the club was dim compared to when the lighting was bright, $F(1, 25) = 23.4$, $p < .001$.
- ✓ The main effect of alcohol on the attractiveness of mates selected was significant, $F(3, 75) = 104.4$, $p < .001$. This indicated that when the lighting in the club was ignored, the attractiveness of the mates selected differed according to how much alcohol was drunk before the selection was made. Specifically, *post hoc* tests revealed that, compared to a baseline of when no alcohol had been consumed, the attractiveness of selected mates was not different after 2 pints ($p > .05$), but was significantly lower after 4 and 6 pints (both ps $< .001$). The mean attractiveness after 2 pints was also significantly higher than after 4 pints and 6 pints (both ps $< .001$), and the mean attractiveness after 4 pints was significantly higher than after 6 pints ($p < .001$). To sum up, the beer-goggles effect seems to take effect after 2 pints have been consumed and has an increasing impact until 6 pints are consumed.
- ✓ The lighting × alcohol interaction was significant, $F(3, 75) = 22.2$, $p < .001$, indicating that the effect of alcohol on the attractiveness of the mates selected differed when lighting was dim compared to when it was bright.

Using lme()

If we want to look at the overall main effect then we need to compare the model containing the predictors from a baseline that includes no predictors other than the intercept. We can specify the baseline model as we did in the chapter and then add the predictors to the model one at a time:

```
baseline<-lme(Attractiveness ~ 1, random = ~1|Participant/Pints/Lighting, data =
longGoggles, method = "ML")
PintsModel<-update(baseline, .~. + Pints)
LightingModel<-update(PintsModel, .~. + Lighting)
gogglesModel<-update(LightingModel, .~. + Pints:Lighting)
```

By comparing these models (*baseline, PintsModel* and *LightingModel* and *gogglesModel*) we can see whether adding the variables **Pints** and **Lighting** and their interaction as predictors significantly improves the model (in other words, by using group means to predict the attractiveness of the women chosen – does the model fit the data better than when we don't include these predictors?). To compare the models execute:

```
anova(baseline, PintsModel, LightingModel, gogglesModel)
```

Executing the above command produces the output below, which first compares the effect of **Pints** to the baseline (i.e., no predictors). By adding **Pints** as a predictor we increase the degrees of freedom by 3 (the three contrasts that we used to code this variable) and significantly improve the model. In other words, the number of pints drunk had a significant effect on attractiveness, $\chi^2(3) = 144.40$, $p < .0001$. Next, we see the effect of adding the main effect of **Lighting** into the model (compared to the previous model that contained only the effect of **Pints**). The degrees of freedom are increased by 1 (the one contrast used to code this variable) and the fit of the model is significantly improved; the type of lighting used in the club had a significant effect on attractiveness, $\chi^2(1) = 14.87$, $p < .001$. The final model (which includes both main effects and the interaction between them) is then compared to the previous model (which includes only the two main effects). The interaction term significantly improves the model fit; therefore, attractiveness of the woman chosen was significantly affected by the combined effect of the number of pints drunk and type of lighting, $\chi^2(3) = 53.82$, $p < .0001$. These results confirm the overall effects that we looked at with *ezANOVA()* in the previous section, and you should look back at that section to remind yourself of how we interpreted these effects.

```
               Model df      AIC      BIC    logLik     Test   L.Ratio   p-value
baseline           1  5 1770.963 1787.651 -880.4816
PintsModel         2  8 1632.559 1659.259 -808.2793 1 vs 2 144.40462  <.0001
LightingModel      3  9 1619.688 1649.726 -800.8442 2 vs 3  14.87011   1e-04
gogglesModel       4 12 1571.870 1611.920 -773.9350 3 vs 4  53.81834   <.0001
```

We can further explore the model by executing:

```
summary(gogglesModel)
```

The output below shows the parameter estimates for the model (I've edited some of the names to save space). Most important, these include the parameters for the contrasts that we set for each variable. First, we get the three contrasts for **Pints**, which show a significant effect on attractiveness when comparing no alcohol to some alcohol, b = 4.73, $t(75)$ = 10.21, p = .00, when comparing 6 pints to 2 and 4 pints (combined), b = –10.01, $t(75)$ = -12.22, p = .00, and when comparing 2 pints to 4 pints, b = 14.12, $t(75)$ = 9.95, p = .00.

Next, we get the contrast for **Lighting**, which shows a significant effect on attractiveness when comparing dim lighting to bright lighting, b = 6.74, $t(114)$ = 17.26, p < .001, and when comparing positive to neutral imagery, b = 6.83, $t(100)$ = 5.27, p < .001. The next three effects are the contrasts for the interaction term which were all found to be significant (all ps < .0001).

```
Formula: ~1 | Lighting %in% Pints %in% Participant
        (Intercept) Residual
StdDev:     8.78435 2.382856

Fixed effects: Attractiveness ~ Pints + Lighting + Pints:Lighting
                                    Value Std.Error  DF   t-value p-value
(Intercept)                      46.06667 1.0289595 100  44.77015       0
NonevsAlcohol                     4.73333 0.4634222  75  10.21387       0
MaxvsLess                       -10.01282 0.8192225  75 -12.22235       0
TwovsFour                        14.11538 1.4189349  75   9.94787       0
LightingBright                    6.83333 1.2957421 100   5.27368       0
NonevsAlcohol:LightingBright     -2.56410 0.5945273 100  -4.31284       0
MaxvsLess:LightingBright          5.03205 1.0509856 100   4.78794       0
TwovsFour:LightingBright         -9.17308 1.8203605 100  -5.03915       0
 Correlation:
                                 (Intr) PntsNA PntsML PntsTF LghtnB PNA:LB PML:LB
PintsNonevsAlcohol               -0.113
PintsMaxvsLess                    0.000  0.000
PintsTwovsFour                    0.000  0.000  0.000
LightingBright                   -0.630  0.074  0.000  0.000
PintsNonevsAlcohol:LightingBright  0.072 -0.641  0.000  0.000 -0.115
PintsMaxvsLess:LightingBright     0.000  0.000 -0.641  0.000  0.000  0.000
PintsTwovsFour:LightingBright     0.000  0.000  0.000 -0.641  0.000  0.000  0.000
```

Effect sizes

If we make sure that we have executed the *rcontrast* function from the book, we can use it to calculate r for the contrasts we did by executing these commands (the values of t and df come from the output for *summary(gogglesModel)*:

```
rcontrast( 10.21387, 75)
rcontrast(-12.22235, 75)
rcontrast(9.94787, 75)
rcontrast(5.27368, 100)
rcontrast(-4.31284, 100)
rcontrast(4.78794, 100)
rcontrast(-5.03915, 100)
```

```
[1] "r =  0.762732334903438"
[1] "r =  0.81593768690933"
[1] "r =  0.754232522082961"
[1] "r =  0.466475100455845"
[1] "r =  0.396022557480187"
[1] "r =  0.431846799011667"
[1] "r =  0.450008361154144"
```

Writing the result

- ✓ The number of pints drunk had a significant effect on attractiveness, $\chi^2(3) = 144.40$, p < .0001, as did the type of lighting used in the room, $\chi^2(1) = 14.87$, p < .001. Most important, the pints × lighting interaction was significant, $\chi^2(3) = 53.82$, p < .0001. Contrasts on this interaction term revealed that: (1) drinking some alcohol compared to drinking no alcohol resulted in significantly less attractive women being chosen, and this effect was significantly

greater when the room lighting was dim compared to when it was bright, $t(100) = -4.31$, $p = .00$, $r = .40$; (2) drinking 6 pints compared with drinking 2 and 4 pints (combined) resulted in significantly less attractive women being chosen, and again this effect was significantly greater when in dim lighting than when in bright lighting, $t(100) = 4.79$, $p < .001$, $r = .43$; (3) drinking 4 pints compared to drinking 2 pints resulted in significantly less attractive women being chosen, and this effect was once again significantly greater when dim lighting was used than when bright lighting was used, $t(100) = -5.04$, $r = .45$. To sum up, there was a significant interaction between the amount of alcohol drunk and the lighting in the club, an interaction graph revealed that the decline in the attractiveness of the selected mate seen after 2 pints (compared to after 4) was significantly more pronounced when the lights were dim.