10 Comparing several means: ANOVA (GLM 1)

FIGURE 10.1
My brother Paul (left) and I (right) in our very fetching school uniforms.

10.1. What will this chapter tell me? ①

There are pivotal moments in everyone's life, and one of mine was at the age of 11. Where I grew up in England there were three choices when leaving primary school and moving onto secondary school: (1) state school (where most people go); (2) grammar school (where clever people who pass an exam called the 11+ go); and (3) private school (where rich people go). My parents were not rich and I am not clever and consequently I failed my 11+, so private school and grammar school (where my clever older brother had gone) were out.

This left me to join all of my friends at the local state school. I could not have been happier. Imagine everyone's shock when my parents received a letter saying that some extra spaces had become available at the grammar school; although the local authority could scarcely believe it and had checked the 11+ papers several million times to confirm their findings, I was next on their list. I could not have been unhappier. So, I waved goodbye to all of my friends and trundled off to join my brother at Ilford County High School for Boys (a school that still hit students with a cane if they were particularly bad and that, for some considerable time and with good reason, had 'H.M. Prison' painted in huge white letters on its roof). It was goodbye to normality, and hello to 6 years of learning how not to function in society. I often wonder how my life would have turned out had I not gone to this school; in the parallel universes where the letter didn't arrive and Andy went to state school, or where my parents were rich and Andy went to private school, what became of him? If we wanted to compare these three situations we couldn't use a *t*-test because there are more than two conditions.[1] However, this chapter tells us all about the statistical models that we use to analyse situations in which we want to compare more than two conditions: **analysis of variance** (or **ANOVA** to its friends). This chapter will begin by explaining the theory of ANOVA when different participants are used (*independent ANOVA*). We'll then look at how to carry out the analysis in **R** and interpret the results.

10.2. The theory behind ANOVA ②

10.2.1 Inflated error rates ②

Before explaining how ANOVA works, it is worth mentioning why we don't simply carry out several *t*-tests to compare all combinations of groups that have been tested. Imagine a situation in which there were three experimental conditions and we were interested in differences between these three groups. If we were to carry out *t*-tests on every pair of groups, then that would involve doing three separate tests: one to compare groups 1 and 2, one to compare groups 1 and 3, and one to compare groups 2 and 3. If each of these *t*-tests uses a .05 level of significance then for each test the probability of falsely rejecting the null hypothesis (known as a Type I error) is only 5%. Therefore, the probability of no Type I errors is .95 (95%) for each test. If we assume that each test is independent (hence, we can multiply the probabilities) then the overall probability of no Type I errors is $.95^3 = .95 \times .95 \times .95 = .857$, because the probability of no Type I errors is .95 for each test and there are three tests. Given that the probability of no Type I errors is .857, then we can calculate the probability of making at least one Type I error by subtracting this number from 1 (remember that the maximum probability of any event occurring is 1). So, the probability of at least one Type I error is $1 - .857 = .143$, or 14.3%. Therefore, across this group of tests, the probability of making a Type I error has increased from 5% to 14.3%, a value greater than the criterion accepted by scientists. This error rate across statistical tests conducted on the same experimental data is known as the **familywise** or **experimentwise error rate**. An experiment with three conditions is a relatively simple design, and so the effect of carrying out several tests is not severe. If you imagine that we now increase the number of experimental conditions from three to five (which is only two more groups) then the

[1] Really, this is the least of our problems: there's the small issue of needing access to parallel universes.

number of *t*-tests that would need to done increases to 10.[2] The familywise error rate can be calculated using the following general equation:

$$\text{familywise error} = 1 - (0.95)^n \tag{10.1}$$

in which n is the number of tests carried out on the data. With 10 tests carried out, the familywise error rate is $1 - .95^{10} = .40$, which means that there is a 40% chance of having made at least one Type I error. For this reason we use ANOVA rather than conducting lots of *t*-tests.

10.2.2. Interpreting *F* ②

When we perform a *t*-test, we test the hypothesis that the two samples have the same mean. Similarly, ANOVA tells us whether three or more means are the same, so it tests the null hypothesis that all group means are equal. An ANOVA produces an *F*-statistic or *F*-ratio, which is similar to the *t*-statistic in that it compares the amount of systematic variance in the data to the amount of unsystematic variance. In other words, *F* is the ratio of the model to its error.

ANOVA is an *omnibus* test, which means that it tests for an overall effect: so, it does not provide specific information about which groups were affected. Suppose an experiment was conducted with three different groups, and the *F*-ratio tells us that the means of these three samples are not equal (i.e., that $\overline{X}_1 = \overline{X}_2 = \overline{X}_3$ is *not* true). There are several ways in which the means can differ. The first possibility is that all three sample means are significantly different ($\overline{X}_1 \neq \overline{X}_2 \neq \overline{X}_3$). A second possibility is that the means of groups 1 and 2 are the same but group 3 has a significantly different mean from both of the other groups ($\overline{X}_1 = \overline{X}_2 \neq \overline{X}_3$). Another possibility is that groups 2 and 3 have similar means but group 1 has a significantly different mean ($\overline{X}_1 \neq \overline{X}_2 = \overline{X}_3$). Finally, groups 1 and 3 could have similar means but group 2 has a significantly different mean from both ($\overline{X}_1 = \overline{X}_3 \neq \overline{X}_2$). So, in an experiment, the *F*-ratio tells us only that the experimental manipulation has had some effect, but it doesn't tell us specifically what the effect was.

10.2.3. ANOVA as regression ②

I've hinted several times that all statistical tests boil down to variants on regression. In fact, ANOVA is just a special case of regression. This surprises many scientists because ANOVA and regression are usually used in different situations. The reason is largely historical in that

[2] These comparisons are group 1 vs. 2, 1 vs. 3, 1 vs. 4, 1 vs. 5, 2 vs. 3, 2 vs. 4, 2 vs. 5, 3 vs. 4, 3 vs. 5 and 4 vs. 5. The number of tests required – let's call it C – is calculated using this equation:

$$C = \frac{k!}{2(k-2)!}$$

in which k is the number of experimental conditions. The ! symbol stands for *factorial*, which means that you multiply the value preceding the symbol by all of the whole numbers between zero and that value (so $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$). Thus, with five conditions we find that:

$$C = \frac{5!}{2(5-2)!} = \frac{120}{2 \times 6} = 10$$

two distinct branches of methodology developed in the sciences: correlational research and experimental research. Researchers interested in controlled experiments adopted ANOVA as their procedure of choice, whereas those looking for real-world relationships adopted multiple regression. As we all know, scientists are intelligent, mature and rational people, and so neither group was tempted to slag off the other and claim that their own choice of methodology was far superior to the other (yeah, right!). With the divide in methodologies came a chasm between the statistical methods adopted by the two opposing camps (Cronbach, 1957, documents this divide in a lovely article). This divide has lasted many decades, to the extent that now students are generally taught regression and ANOVA in very different contexts and many textbooks teach ANOVA and regression in entirely different ways.

Although many considerably more intelligent people than me have attempted to redress the balance (notably the great Jacob Cohen, 1968), I am passionate about making my own small, feeble-minded attempt to enlighten you (and I set the ball rolling in sections 7.12 and 9.4.2). There are several good reasons why I think ANOVA should be taught within the context of regression. First, it provides a familiar context: I wasted many trees trying to explain regression, so why not use this base of knowledge to explain a new concept? (It should make it easier to understand.) Second, the traditional method of teaching ANOVA (known as the variance-ratio method) is fine for simple designs, but becomes impossibly cumbersome in more complex situations (such as analysis of covariance). The regression model extends very logically to these more complex designs without anyone needing to get bogged down in mathematics. Finally, the variance-ratio method becomes extremely unmanageable in unusual circumstances such as when you have unequal sample sizes.[3] The regression method makes these situations considerably simpler. Although these reasons are good enough, it is also the case that **R** very much deals with ANOVA in a regression-y sort of way (known as the general linear model, or GLM).

I have mentioned that ANOVA is a way of comparing the ratio of systematic variance to unsystematic variance in an experimental study. The ratio of these variances is known as the *F*-ratio. However, any of you who have read Chapter 7 should recognize the *F*-ratio (see section 7.2.3) as a way to assess how well a regression model can predict an outcome compared to the error within that model. If you haven't read Chapter 7 (surely not!), have a look before you carry on (it should only take you a couple of weeks to read). How can the *F*-ratio be used to test differences between means *and* whether a regression model fits the data? The answer is that when we test differences between means we *are* fitting a regression model and using *F* to see how well it fits the data, but the regression model contains only categorical predictors (i.e., grouping variables). So, just as the *t*-test could be represented by the linear regression equation (see section 9.4.2), ANOVA can be represented by the multiple regression equation in which the number of predictors is one less than the number of categories of the independent variable.

Let's take an example. There was a lot of controversy, when I wrote the first edition of my SPSS book, surrounding the drug Viagra. Admittedly there's less controversy now, but the controversy has been replaced by an alarming number of spam emails on the subject (for which I'll no doubt be grateful in 20 years' time), so I'm going to stick with the example. Viagra is a sexual stimulant (used to treat impotence) that broke into the black market under the belief that it will make someone a better lover (oddly enough, there was a glut of journalists taking the stuff at the time in the name of 'investigative journalism'… hmmm!). In the psychology literature, sexual performance issues have been linked to a loss of libido (Hawton, 1989). Suppose we tested this hypothesis by taking three groups of participants and administering one group with a placebo (such as a sugar pill), one group with a low dose of Viagra and one with a high dose. The dependent variable was an objective measure

[3] Having said this, it is well worth the effort in trying to obtain equal sample sizes in your different conditions because unbalanced designs do cause statistical complications (see section 10.3).

Table 10.1 Data in **Viagra.dat**

	Placebo	Low Dose	High Dose
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
$\bar{X}$	**2.20**	**3.20**	**5.00**
s	**1.30**	**1.30**	**1.58**
s^2	**1.70**	**1.70**	**2.50**
Grand mean = **3.467** Grand SD = **1.767** Grand variance = **3.124**			

of libido (I will tell you only that it was measured over the course of a week – the rest I will leave to your own imagination). The data can be found in the file **Viagra.dat** (which is described in detail later in this chapter) and are in Table 10.1.

If we want to predict levels of libido from the different levels of Viagra then we can use the general equation that keeps popping up:

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

If we want to use a linear model, then we saw in section 9.4.2 that when there are only two groups we could replace the 'model' in this equation with a linear regression equation with one dummy variable to describe two groups. This dummy variable was a categorical variable with two numeric codes (0 for one group and 1 for the other). With three groups, however, we can extend this idea and use a multiple regression model with two dummy variables. In fact, as a general rule we can extend the model to any number of groups and the number of dummy variables needed will be one less than the number of categories of the independent variable. In the two-group case, we assigned one category as a base category (remember that in section 9.4.2 we chose the picture condition to act as a base) and this category was coded 0. When there are three categories we also need a base category and you should choose the condition to which you intend to compare the other groups. Usually this category will be the control group. In most well-designed science experiments there will be a group of participants who act as a baseline for other categories. This baseline group should act as the reference or base category, although the group you choose will depend upon the particular hypotheses that you want to test. In unbalanced designs (in which the group sizes are unequal) it is important that the base category contains a fairly large number of cases to ensure that the estimates of the regression coefficients are reliable. In the Viagra example, we can take the placebo group as the base category because this group was a placebo control. We are interested in comparing both the high- and low-dose groups to the group that received no Viagra at all. If the placebo group is the base category then the two dummy variables that we have to create represent the other two conditions: so, we should have one dummy variable called **high** and the other one called **low**). The resulting equation is described as:

$$\text{libido}_i = b_0 + b_2\text{high}_i + b_1\text{low}_i + \varepsilon_i \tag{10.2}$$

In equation (10.2), a person's libido can be predicted from knowing their group code (i.e., the code for the **high** and **low** dummy variables) and the intercept (b_0) of the model. The dummy variables in equation (10.2) can be coded in several ways, but the simplest way

is to use a similar technique to that of the *t*-test. The base category is always coded 0. If a participant was given a high dose of Viagra then they are coded 1 for the **high** dummy variable and 0 for all other variables. If a participant was given a low dose of Viagra then they are coded 1 for the **low** dummy variable and 0 for all other variables (this is the same type of scheme we used in section 7.12). Using this coding scheme we can express each group by combining the codes of the two dummy variables (see Table 10.2).

Table 10.2 Dummy coding for the three-group experimental design

Group	Dummy Variable 1 (high)	Dummy Variable 2 (low)
Placebo	0	0
Low Dose Viagra	0	1
High Dose Viagra	1	0

Placebo group: Let's examine the model for the placebo group. In the placebo group both the **high** and **low** dummy variables are coded 0. Therefore, if we ignore the error term (ε_i), the regression equation becomes:

$$\text{libido}_i = b_0 + (b_1 \quad 0) + (b_2 \quad 0) = b_0$$
$$\overline{X}_{\text{placebo}} = b_0$$

This is a situation in which the high- and low-dose groups have both been excluded (because they are coded with 0). We are looking at predicting the level of libido when both doses of Viagra are ignored, and so the predicted value will be the mean of the placebo group (because this group is the only one included in the model). Hence, the intercept of the regression model, b_0, is always the mean of the base category (in this case the mean of the placebo group).

High-dose group: If we examine the high-dose group, the dummy variable **high** will be coded 1 and the dummy variable **low** will be coded 0. If we replace the values of these codes into equation (10.2) the model becomes:

$$\text{libido}_i = b_0 + (b_1 \quad 0) + (b_2 \quad 1) = b_0 + b_2$$

We know already that b_0 is the mean of the placebo group. If we are interested in only the high-dose group then the model should predict that the value of libido for a given participant equals the mean of the high-dose group. Given this information, the equation becomes:

$$\text{libido}_i = b_0 + b_2$$
$$\overline{X}_{\text{high}} = \overline{X}_{\text{placebo}} + b_2$$
$$b_2 = \overline{X}_{\text{high}} - \overline{X}_{\text{placebo}}$$

Hence, b_2 represents the difference between the means of the high-dose group and the placebo group.

Low-dose group: Finally, if we look at the model when a low dose of Viagra has been taken, the dummy variable **low** is coded 1 (and hence **high** is coded as 0). Therefore, the regression equation becomes:

$$\text{libido}_i = b_0 + (b_1 \quad 1) + (b_2 \quad 0) = b_0 + b_1$$

We know that the intercept is equal to the mean of the base category and that for the low-dose group the predicted value should be the mean libido for a low dose. Therefore the model can be reduced to:

$$\begin{aligned} \text{libido}_i &= b_0 + b_1 \\ \overline{X}_{\text{low}} &= \overline{X}_{\text{placebo}} + b_1 \\ b_1 &= \overline{X}_{\text{low}} - \overline{X}_{\text{placebo}} \end{aligned}$$

Hence, b_1 represents the difference between the means of the low-dose group and the placebo group. This form of dummy variable coding is the simplest form, but as we will see later, there are other ways in which variables can be coded to test specific hypotheses. These alternative coding schemes are known as *contrasts* (see section 10.4.2). The idea behind contrasts is that you code the dummy variables in such a way that the *b*-values represent differences between groups that you are interested in testing.

SELF-TEST

✓ To illustrate exactly what is going on I have created a file called **dummy.dat**. This file contains the Viagra data but with two additional variables (**dummy1** and **dummy2**) that specify to which group a data point belongs (as in Table 10.2). Access this file and run multiple regression analysis using libido as the outcome and **dummy1** and **dummy2** as the predictors. If you're stuck on how to run the regression then read Chapter 7 again (these chapters are ordered for a reason).

The resulting analysis is shown in Output 10.1. It might be a good idea to remind yourself of the group means from Table 10.1. The first thing to notice is that, just as in the regression chapter, an ANOVA has been used to test the overall fit of the model. This test is significant, $F(2, 12) = 5.12$, $p < .05$. Given that our model represents the group differences, this ANOVA tells us that using group means to predict scores is significantly better than using the overall mean: in other words, the group means are significantly different.

In terms of the regression coefficients, *b*s, the constant is equal to the mean of the base category (the placebo group). The regression coefficient for the first dummy variable (b_2) is equal to the difference between the means of the high-dose group and the placebo group ($5.0 - 2.2 = 2.8$). Finally, the regression coefficient for the second dummy variable (b_1) is equal to the difference between the means of the low-dose group and the placebo group ($3.2 - 2.2 = 1$). This analysis demonstrates how the regression model represents the three-group situation. We can see from the significance values of the *t*-tests that the difference between the high-dose group and the placebo group (b_2) is significant because $p < .05$. The difference between the low-dose and the placebo group is not, however, significant ($p = .282$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2000     0.6272   3.508  0.00432 **
dummy1        2.8000     0.8869   3.157  0.00827 **
dummy2        1.0000     0.8869   1.127  0.28158
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.402 on 12 degrees of freedom
Multiple R-squared: 0.4604,       Adjusted R-squared: 0.3704
F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Output 10.1

A four-group experiment can be described by extending the three-group scenario. I mentioned earlier that you will always need one less dummy variable than the number of groups in the experiment: therefore, this model requires three dummy variables. As before, we need to specify one category as a base category (a control group). This base category should have a code of 0 for all three dummy variables. The remaining three conditions will have a code of 1 for the dummy variable that described that condition and a code of 0 for the other two dummy variables. Table 10.3 illustrates how the coding scheme would work.

Table 10.3 Dummy coding for the four-group experimental design

	Dummy variable 1	*Dummy variable 2*	*Dummy variable 3*
Group 1	1	0	0
Group 2	0	1	0
Group 3	0	0	1
Group 4 (base)	0	0	0

10.2.4. Logic of the *F*-ratio ②

In Chapter 7 we learnt a little about the *F*-ratio and its calculation. To recap, we learnt that the *F*-ratio is used to test the overall fit of a regression model to a set of observed data. In other words, it is the ratio of how good the model is compared to how bad it is (its error). I have just explained how ANOVA can be represented as a regression equation, and this should help you to understand what the *F*-ratio tells you about your data. Figure 10.2 shows the Viagra data in graphical form (including the group means, the overall mean and the difference between each case and the group mean). In this example, there were three groups; therefore, we want to test the hypothesis that the means of three groups are different (so the null hypothesis is that the group means are the same). If the group means were all the same, then we would not expect the placebo group to differ from the low-dose group or the high-dose group, and we would not expect the low-dose group to differ from the high-dose group. Therefore, on the diagram, the three shaded blue lines would be in the same vertical position (the exact position would be the grand mean – the solid horizontal line in the figure). We can see from the diagram that the group means are actually different because the horizontal blue lines (the group means) are in different vertical positions. We have just found out that in the regression model, b_2 represents the difference between the means of the placebo and the high-dose group, and b_1 represents the difference in means between the placebo and the low-dose groups. These two distances are represented in Figure 10.2 by the vertical arrows. If the null hypothesis is true and all the groups have the same means, then these *b* coefficients should be zero (because if the group means are equal then the difference between them will be zero).

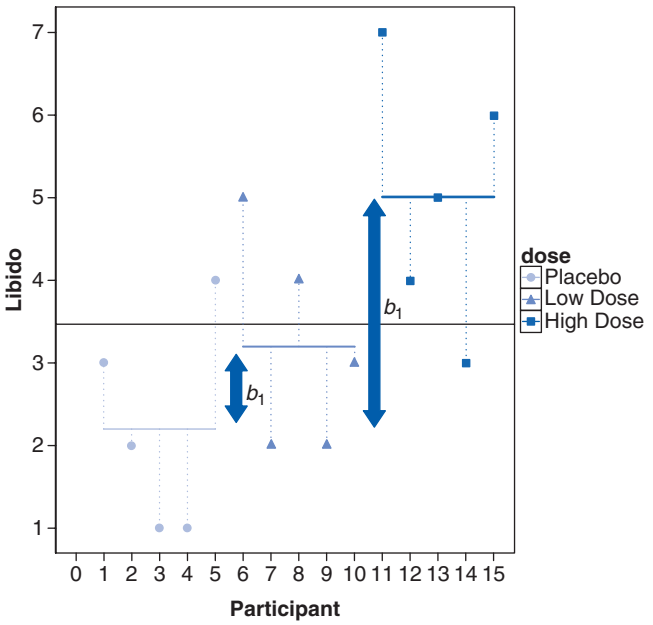


FIGURE 10.2
The Viagra data in graphical form. The shaded blue horizontal lines represent the mean libido of each group. The shapes represent the libido of individual participants (different shapes indicate different experimental groups). The black horizontal line is the average libido of all participants

The logic of ANOVA follows from what we understand about regression:

- The simplest model we can fit to a set of data is the grand mean (the mean of the outcome variable). This basic model represents 'no effect' or 'no relationship between the predictor variable and the outcome'.
- We can fit a different model to the data collected that represents our hypotheses. If this model fits the data well then it must be better than using the grand mean. Sometimes we fit a linear model (the line of best fit), but in experimental research we often fit a model based on the means of different conditions.
- The intercept and one or more regression coefficients can describe the chosen model.
- The regression coefficients determine the shape of the model that we have fitted; therefore, the bigger the coefficients, the greater the deviation between the line and the grand mean.
- In correlational research, the regression coefficients represent the slope of the line, but in experimental research they represent the differences between group means.
- The bigger the differences between group means, the greater the difference between the model and the grand mean.
- If the differences between group means are large enough, then the resulting model will be a better fit of the data than the grand mean.
- If this is the case we can infer that our model (i.e., predicting scores from the group means) is better than not using a model (i.e., predicting scores from the grand mean). Put another way, our group means are significantly different.

Just like when we used ANOVA to test a regression model, we can compare the improvement in fit due to using the model (rather than the grand mean) to the error that still remains. Another way of saying this is that when the grand mean is used as a model, there will be a certain amount of variation between the data and the grand mean. When a model is fitted it will explain some of this variation, but some will be left unexplained. The *F*-ratio

is the ratio of the explained to the unexplained variation. Look back at section 7.2.3 to refresh you memory on these concepts before reading on. This may all sound quite complicated, but actually most of it boils down to variations on one simple equation (see Jane Superbrain Box 10.1).

JANE SUPERBRAIN 10.1

You might be surprised to know that ANOVA boils down to one equation (well, sort of) ②

At every stage of the ANOVA we're assessing variation (or deviance) from a particular model (be that the most basic model, or the most sophisticated model). We saw back in section 2.4.1 that the extent to which a model deviates from the observed data can be expressed, in general, in the form of equation (10.3). So, in ANOVA, as in regression, we use equation (10.3) to calculate the fit of the most basic model, and then the fit of the best model (the line of best fit). If the best model is any good then it should fit the data significantly better than our basic model:

$$\text{deviation} = \sum(\text{observed} - \text{model})^2 \qquad (10.3)$$

The interesting point is that all of the sums of squares in ANOVA are variations on this one basic equation. All that changes is what we use as the model, and what the corresponding observed data are. Look through the various sections on the sums of squares and compare the resulting equations to equation (10.3); hopefully, you can see that they are all basically variations on this general form of the equation!

10.2.5. Total sum of squares (SS_T) ②

To find the total amount of variation within our data we calculate the difference between each observed data point and the grand mean. We then square these differences and add them together to give us the total sum of squares (SS_T):

$$SS_T = \sum_{i=1}^{N}(x_i - \bar{x}_{grand})^2 \qquad (10.4)$$

We also saw in section 2.4.1 that the variance and the sums of squares are related such that variance, $s^2 = SS/(N-1)$, where N is the number of observations. Therefore, we can calculate the total sums of squares from the variance of all observations (the **grand variance**) by rearranging the relationship ($SS = s^2(N-1)$). The grand variance is the variation between all scores, regardless of the experimental condition from which the scores come. Figure 10.3 shows the different sums of squares graphically (note the similarity to Figure 7.4 which we looked at when we learnt about regression). The top left panel shows the total sum of squares: it is the sum of the squared distances between each point and the solid horizontal line (which represents the mean of all scores).

The grand variance for the Viagra data is given in Table 10.1, and if we count the number of observations we find that there were 15 in all. Therefore, SS_T is calculated as follows:

$$\begin{aligned} SS_T &= s^2_{grand}(n-1) \\ &= 3.124(15-1) = 3.124 \times 14 = 43.74 \end{aligned}$$

Before we move on, it is important to understand degrees of freedom, so have a look back at Jane Superbrain Box 2.2 to refresh your memory. We saw before that when we estimate population values, the degrees of freedom are typically one less than the number of scores used to calculate the population value. This is because to get these estimates we have to hold something constant in the population (in this case the mean), which leaves all but one of the scores free to vary (see Jane Superbrain Box 2.2). For SS_T, we used the entire sample (i.e., 15 scores) to calculate the sums of squares and so the total degrees of freedom (df_T) are one less than the total sample size ($N - 1$). For the Viagra data, this value is 14.

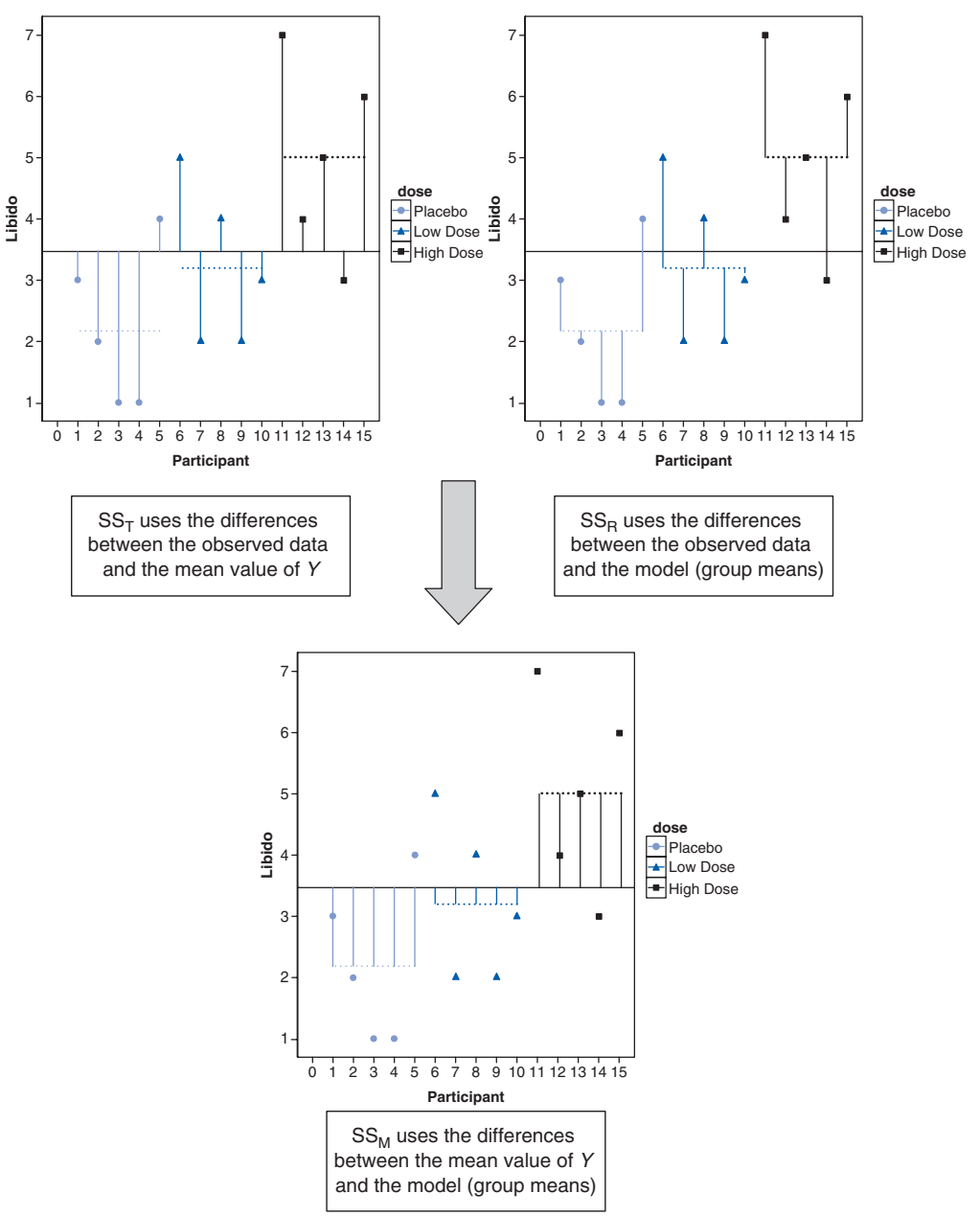


FIGURE 10.3 Graphical representation of the different sums of squares in ANOVA designs

10.2.6. Model sum of squares (SS_M) ②

So far, we know that the total amount of variation within the data is 43.74 units. We now need to know how much of this variation the regression model can explain. In the ANOVA scenario, the model is based upon differences between group means, and so the model sums of squares tell us how much of the total variation can be explained by the fact that different data points come from different groups.

In section 7.2.3 we saw that the model sum of squares is calculated by taking the difference between the values predicted by the model and the grand mean (see Figure 7.4). In ANOVA, the values predicted by the model are the group means (the dashed horizontal lines in Figure 10.3). The bottom panel in Figure 10.3 shows the model sum of squared error: it is the sum of the squared distances between what the model predicts for each data point (i.e., the dotted horizontal line for the group to which the data point belongs) and the overall mean of the data (the solid horizontal line).

For each participant the value predicted by the model is the mean for the group to which the participant belongs. In the Viagra example, the predicted value for the five participants in the placebo group will be 2.2, for the five participants in the low-dose condition it will be 3.2, and for the five participants in the high-dose condition it will be 5. The model sum of squares requires us to calculate the differences between each participant's predicted value and the grand mean. These differences are then squared and added together (for reasons that should be clear in your mind by now). We know that the predicted value for participants in a particular group is the mean of that group. Therefore, the easiest way to calculate SS_M is to do the following:

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 participants within that group (n_k).
4. Add the values for each group together.

The mathematical expression for this process is:

$$SS_M = \sum_{n=1}^{k} n_k(\bar{x}_k - \bar{x}_{grand})^2 \tag{10.5}$$

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

$$\begin{aligned} SS_M &= 5(2.200 - 3.467)^2 + 5(3.200 - 3.467)^2 + 5(5.00 - 3.467)^2 \\ &= 5(-1.267)^2 + 5(-0.267)^2 + 5(1.533)^2 \\ &= 8.025 + 0.335 + 11.755 \\ &= 20.135 \end{aligned}$$

For SS_M, the degrees of freedom (df_M) will always be one less than the number of parameters estimated. In short, this value will be the number of groups minus one (which you'll see denoted as $k - 1$). So, in the three-group case the degrees of freedom will always be 2 (because the calculation of the sums of squares is based on the group means, two of which will be free to vary in the population if the third is held constant).

10.2.7. Residual sum of squares (SS_R) ②

We now know that there are 43.74 units of variation to be explained in our data, and that our model can explain 20.14 of these units (nearly half). 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 such as individual differences in weight, testosterone or whatever. Knowing SS_T and SS_M already, the simplest way to calculate SS_R is to subtract SS_M from SS_T ($SS_R = SS_T - SS_M$); however, telling you to do this provides little insight into what is being calculated and, of course, if you've messed up the calculations of either SS_M or SS_T (or indeed both!) then SS_R will be incorrect also.

We saw in section 7.2.3 that the residual sum of squares is the difference between what the model predicts and what was actually observed. In ANOVA, the values predicted by the model are the group means (the dashed horizontal lines in Figure 10.3). The top left panel shows the residual sum of squared error: it is the sum of the squared distances between each point and the dotted horizontal line for the group to which the data point belongs.

We already know that for a given participant, the model predicts the mean of the group to which that person belongs. Therefore, SS_R is calculated by looking at the difference between the score obtained by a person and the mean of the group to which the person belongs. In graphical terms the vertical lines in Figure 10.3 represent this sum of squares. These distances between each data point and the group mean are squared and then added together to give the residual sum of squares, SS_R, thus:

$$SS_R = \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2 \tag{10.6}$$

Now, the sum of squares for each group represents the sum of squared differences between each participant's score in that group and the group mean. Therefore, we can express SS_R as $SS_R = SS_{group1} + SS_{group2} + SS_{group3} + \ldots$. Given that we know the relationship between the variance and the sums of squares, we can use the variances for each group in the Viagra data to create an equation like we did for the total sum of squares. As such, SS_R can be expressed as:

$$SS_R = \sum s_k^2 (n_k - 1) \tag{10.7}$$

This just means take the variance from each group (s_k^2) and multiply it by one less than the number of people in that group ($n_k - 1$). When you've done this for each group, add them all up. For the Viagra data, this gives us:

$$\begin{aligned} SS_R &= s_{group1}^2 (n_1 - 1) + s_{group2}^2 (n_2 - 1) + s_{group3}^2 (n_3 - 1) \\ &= (1.70)(5 - 1) + (1.70)(5 - 1) + (2.50)(5 - 1) \\ &= (1.70 \times 4) + (1.70 \times 4) + (2.50 \times 4) \\ &= 6.8 + 6.8 + 10 \\ &= 23.60 \end{aligned}$$

The degrees of freedom for SS_R (df_R) are the total degrees of freedom minus the degrees of freedom for the model ($df_R = df_T - df_M = 14 - 2 = 12$). Put another way, it's $N - k$: the total sample size, N, minus the number of groups, k.

10.2.8. Mean squares ②

SS_M tells us the *total* variation that the regression model (e.g., the experimental manipulation) explains, and SS_R tells us the *total* variation that is due to extraneous factors. However, because both of these values are summed values they will be influenced by the number of scores that were summed; for example, SS_M used the sum of only 3 different values (the group means) compared to SS_R and SS_T, which used the sum of 12 and 15 values, respectively. To eliminate this bias we can calculate the average sum of squares (known as the *mean squares*, MS), which is simply the sum of squares divided by the degrees of freedom. The reason why we divide by the degrees of freedom rather than the number of parameters used to calculate the SS is because we are trying to extrapolate to a population and so some parameters within that populations will be held constant (this is the same reason why we divide by $N - 1$ when calculating the variance; see Jane Superbrain Box 2.2). So, for the Viagra data we find the following mean squares:

$$MS_M = \frac{SS_M}{df_M} = \frac{20.135}{2} = 10.067$$

$$MS_R = \frac{SS_R}{df_R} = \frac{23.60}{12} = 1.967$$

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

10.2.9. The *F*-ratio ②

The *F*-ratio is a measure of the ratio of the variation explained by the model and the variation explained by unsystematic factors. In other words, it is the ratio of how good the model is against how bad it is (how much error there is). It can be calculated by dividing the model mean squares by the residual mean squares.

$$F = \frac{MS_M}{MS_R} \tag{10.8}$$

As with the independent *t*-test, the *F*-ratio is, therefore, a measure of the ratio of systematic variation to unsystematic variation. In experimental research, it is the ratio of the experimental effect to the individual differences in performance. An interesting point about the *F*-ratio is that because it is the ratio of systematic variance to unsystematic variance, if its value is less than 1 then it must, by definition, represent a non-significant effect. The reason why this statement is true is because if the *F*-ratio is less than 1 it means that MS_R is greater than MS_M, which in real terms means that there is more unsystematic than systematic variance. You can think of this in terms of the effect of natural differences in ability being greater than differences brought about by the experiment. In this scenario, we can, therefore, be sure that our experimental manipulation has been unsuccessful (because it has brought about less change than if we left our participants alone!). For the Viagra data, the *F*-ratio is:

$$F = \frac{MS_M}{MS_R} = \frac{10.067}{1.967} = 5.12$$

This value is greater than 1, which indicates that the experimental manipulation had some effect above and beyond the effect of individual differences in performance. However, it doesn't yet tell us whether the *F*-ratio is large enough to not be a chance result. To discover this we can compare the obtained value of *F* against the maximum value we would expect to get by chance if the group means were equal in an *F*-distribution with the same degrees of freedom (these values can be found in the Appendix); if the value we obtain exceeds this critical value we can be confident that this reflects an effect of our independent variable (because this value would be very unlikely if there were no effect in the population). In this case, with 2 and 12 degrees of freedom the critical values are 3.89 ($p = .05$) and 6.93 ($p = .01$). The observed value, 5.12, is, therefore, significant at a .05 level of significance but not significant at a .01 level. The exact significance produced by **R** should, therefore, fall somewhere between .05 and .01 (which, incidentally, it does).

10.3. Assumptions of ANOVA ③

The assumptions under which the *F*-statistic is reliable are the same as for all parametric tests based on the normal distribution (see section 5.2). That is, the variances in each experimental condition need to be fairly similar (*homogeneity of variance*), observations should be independent and the dependent variable should be measured on at least an interval scale. In terms of normality, what matters is that distributions *within groups* are normally distributed.

10.3.1. Homogeneity of variance ②

As with the *t*-test, there is an assumption that the variances of the groups are equal. This assumption can be tested using Levene's test, which tests the null hypothesis that the variances of the groups are the same (see section 5.7.1). Basically, it is an ANOVA test conducted on the absolute differences between the observed data and the mean or median from which the data came (see Oliver Twisted). If Levene's test is significant (i.e., the *p*-value is less than .05) then we can say that the variances are significantly different. This would mean that we had violated one of the assumptions of ANOVA and we would have to take steps to rectify this matter.

OLIVER TWISTED

Please Sir, can I have some more … Levene's test?

'Liar! Liar! Pants on fire!' screams Oliver, his cheeks red and his eyes about to explode, 'You promised in Chapter 5 to explain Levene's test properly and you haven't, you spatula head'. True enough, Oliver, I do have a spatula for a head. I also have a very nifty little demonstration of Levene's test in the additional material for this chapter on the companion website. It will tell you more than you could possibly want to know. Let's go fry an egg …

10.3.2. Is ANOVA robust? ③

You often hear people say 'ANOVA is a robust test', which means that it doesn't matter much if we break the assumptions of the test: the *F*-ratio will still be accurate. There is

some truth to this statement, but it is also an oversimplification of the situation. For one thing, the term *ANOVA* covers many different situations and the performance of *F* has been investigated in only some of those situations. There are two issues to consider. First, does *F* control the Type I error rate or is it significant even when there are no differences between means? Second, does *F* have enough power (i.e., is it able to detect differences when they are there)? Let's have a look at the evidence.

Looking at normality first, Glass et al. (1972) reviewed a lot of evidence that suggests that *F* controls the Type I error rate well under conditions of skew, kurtosis and non-normality. Skewed distributions seem to have little effect on the error rate and power for two-tailed tests (but can have serious consequences for one-tailed tests). However, some of this evidence has been questioned (see Jane Superbrain Box 5.1). In terms of kurtosis, leptokurtic distributions make the Type I error rate too low (too many null effects are significant) and consequently the power is too high; platykurtic distributions have the opposite effect. The effects of kurtosis seem unaffected by whether sample sizes are equal or not. One study that is worth mentioning in a bit of detail is by Lunney (1970) who investigated the use of ANOVA in just about the most non-normal situation you could imagine: when the dependent variable is binary (it could have values of only 0 or 1). The results showed that when the group sizes were equal, ANOVA was accurate when there were at least 20 degrees of freedom and the smallest response category contained at least 20% of all responses. If the smaller response category contained less than 20% of all responses then ANOVA performed accurately only when there were 40 or more degrees of freedom. The power of *F* also appears to be relatively unaffected by non-normality (Donaldson, 1968). This evidence suggests that *when group sizes are equal* the *F*-statistic can be quite robust to violations of normality.

However, when group sizes are not equal the accuracy of *F* is affected by skew, and non-normality also affects the power of *F* in quite unpredictable ways (Wilcox, 2005). One situation that Wilcox describes shows that when means are equal the error rate (which should be 5%) can be as high as 18%. If you make the differences between means bigger you should find that power increases, but actually he found that initially power *decreased* (although it increased when he made the group differences bigger still). As such *F* can be biased when normality is violated.

Turning to violations of the assumption of homogeneity of variance, ANOVA is fairly robust in terms of the error rate when sample sizes are equal. However, when sample sizes are unequal, ANOVA is not robust to violations of homogeneity of variance (this is why earlier on I said it's worth trying to collect equal-sized samples of data across conditions!). When groups with larger sample sizes have larger variances than the groups with smaller sample sizes, the resulting *F*-ratio tends to be conservative. That is, it's more likely to produce a non-significant result when a genuine difference does exist in the population. Conversely, when the groups with larger sample sizes have smaller variances than the groups with smaller samples sizes, the resulting *F*-ratio tends to be liberal. That is, it is more likely to produce a significant result when there is no difference between groups in the population (put another way, the Type I error rate is not controlled) – see Glass et al. (1972) for a review. When variances are proportional to the means then the power of *F* seems to be unaffected by the heterogeneity of variance and trying to stabilize variances does not substantially improve power (Budescu, 1982; Budescu & Appelbaum, 1981).

Violations of the assumption of independence are very serious indeed. Scariano and Davenport (1987) showed that when this assumption is broken (i.e., observations across groups are correlated) then the Type I error rate is substantially inflated. For example, using the conventional .05 Type I error rate when observations are independent, if these observations are made to correlate moderately (say, with a Pearson coefficient of .5), when comparing three groups, each of 10 observations, the actual Type I error rate is .74 (a substantial inflation!). Therefore, if observations are correlated you might think that you

are working with the accepted .05 error rate (i.e., you'll incorrectly find a significant result only 5% of the time) when in fact your error rate is closer to .75 (i.e., you'll find a significant result on 75% of occasions when, in reality, there is no effect in the population).

There are various things that can be done to combat the litany of woe that you have just read. To find out more see Jane Superbrain Box 10.2.

JANE SUPERBRAIN 10.2

What do I do in ANOVA when assumptions are broken? ③

As we saw in Chapter 5, one common way to rectify problems with assumptions is to transform all of the data and then reanalyse these transformed values (see Chapter 5). When homogeneity of variance is the problem there are versions of the *F*-ratio that have been derived to be robust when homogeneity of variance has been violated. One that can be implemented in **R** is **Welch's *F*** (1951) – see Oliver Twisted.

If you have distributional problems, then there are robust (see section 5.8.4) variants of ANOVA that have been implemented in **R** by Wilcox (2005). These methods are based on bootstrapping or trimmed means and M-estimators (both of which can also include a bootstrap). We'll cover these methods later in the chapter.

On balance, if you have the stomach for it, Wilcox's robust methods are probably the best approach to dealing with violations of assumptions. If you don't have the stomach for it, there are a group of tests (often called assumption-free, distribution-free or non-parametric tests, none of which are particularly accurate names) that you can use instead. The one-way independent ANOVA has a non-parametric counterpart called the Kruskal–Wallis test. If you have non-normally distributed data, or have violated some other assumption, then this test can be a useful way around the problem. This test is described in Chapter 15.

OLIVER TWISTED

Please Sir, can I have some more … Welch's F?

'You don't understand Welch's *F*,' taunts Oliver, 'Andy, Andy, brains all sandy ….' Whatever, Oliver. Welch's *F* adjusts *F* and the residual degrees of freedom to combat problems arising from violations of the homogeneity of variance assumption. There is a lengthy explanation about Welch's *F* in the additional material available on the companion website. Oh, and Oliver, microchips are made of sand.

10.4. Planned contrasts ②

The *F*-ratio tells us only whether the model fitted to the data accounts for more variation than extraneous factors, but it doesn't tell us where the differences between groups lie. So, if the *F*-ratio is large enough to be statistically significant, then we know only that one or more of the differences between means are statistically significant (e.g., either b_2 or b_1 is statistically significant). It is, therefore, necessary after conducting an ANOVA to

carry out further analysis to find out which groups differ. In multiple regression, each *b* coefficient is tested individually using a *t*-test and we could do the same for ANOVA. However, we would need to carry out two *t*-tests, which would inflate the familywise error rate (see section 10.2). Therefore, we need a way to contrast the different groups without inflating the Type I error rate. There are two ways in which to achieve this goal: the first is to break down the variance accounted for by the model into component parts: the second is to compare every group (as if conducting several *t*-tests) but to use a stricter acceptance criterion such that the familywise error rate does not rise above .05. The first option can be done using planned comparisons (also known as **planned contrasts**),[4] whereas the latter option is done using *post hoc* comparisons (see next section). The difference between planned comparisons and ***post hoc* tests** can be likened to the difference between one- and two-tailed tests in that planned comparisons are done when you have specific hypotheses that you want to test, whereas *post hoc* tests are done when you have no specific hypotheses. Let's first look at planned contrasts.

10.4.1. Choosing which contrasts to do ②

In the Viagra example we could have had very specific hypotheses. For one thing, we would expect any dose of Viagra to change libido compared to the placebo group. As a second hypothesis we might believe that a high dose should increase libido more than a low dose. To do planned comparisons, these hypotheses must be derived *before* the data are collected. It is fairly standard in science to want to compare experimental conditions to the control conditions as the first contrast, and then to see where the differences lie between the experimental groups. ANOVA is based upon splitting the total variation into two component parts: the variation due to the experimental manipulation (SS_M) and the variation due to unsystematic factors (SS_R) (see Figure 10.4).

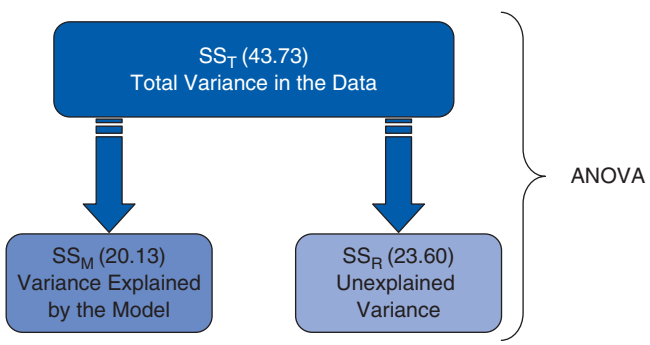


FIGURE 10.4 Partitioning variance for ANOVA

Planned comparisons take this logic a step further by breaking down the variation due to the experiment into component parts (see Figure 10.5). The exact comparisons that are carried out depend upon the hypotheses you want to test. Figure 10.5 shows a situation in which the experimental variance is broken down to look at how much variation is created by the two drug conditions compared to the placebo condition (contrast 1). Then the variation explained by taking Viagra is broken down to see how much is explained by taking a high dose relative to a low dose (contrast 2).

[4] The terms *comparison and contrast* are used interchangeably.

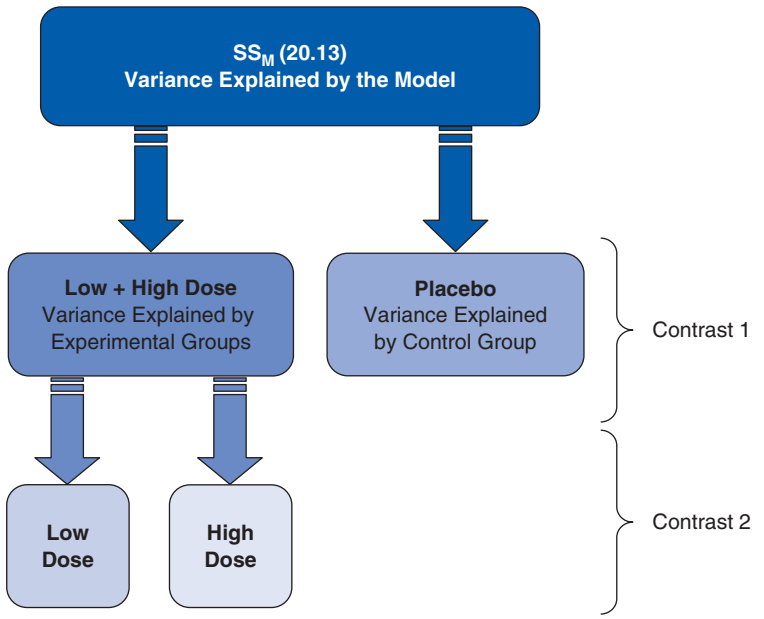


FIGURE 10.5
Partitioning of experimental variance into component comparisons

Typically, students struggle with the notion of planned comparisons, but there are three rules that can help you to work out what to do:

1. If we have a control group, this is usually because we want to compare it against the other groups.
2. Each contrast must compare only two 'chunks' of variation.
3. Once a group has been singled out in a contrast it can't be used in another contrast.

Let's look at these rules in detail. First, if a group is singled out in one comparison, then it should not reappear in another comparison. The important thing to remember is that we are breaking down one chunk of variation into smaller independent chunks. So, in Figure 10.5 contrast 1 involved comparing the placebo group to the experimental groups; because the placebo group is singled out, it should not be incorporated into any other contrasts. You can think of partitioning variance as being similar to slicing up a cake. You begin with a cake (the total sum of squares) and you then cut this cake into two pieces (SS_M and SS_R). You then take the piece of cake that represents SS_M and divide this up into smaller pieces. Once you have cut off a piece of cake you cannot stick that piece back onto the original slice, and you cannot stick it onto other pieces of cake, but you can divide it into smaller pieces of cake. Likewise, once a slice of variance has been split from a larger chunk, it cannot be attached to any other pieces of variance, it can only be subdivided into smaller chunks of variance. Now, all of this talk of cake is making me hungry, but hopefully it illustrates a point.

If you follow the independence of contrasts rule that I've just explained (the cake slicing!), and always compare only two pieces of variance, then you should always end up with one less contrast than the number of groups; there will be $k - 1$ contrasts (where k is the number of conditions you're comparing).

Second, each contrast must compare only two chunks of variance. This rule is so that we can draw firm conclusions about what the contrast tells us. The F-ratio tells us that some of our means differ, but not which ones, and if we were to perform a contrast on more than two chunks of variance we would have the same problem. By comparing only two chunks of variance we can be sure that a significant result represents a difference between these two portions of experimental variation.

Finally, in most social science research we use at least one control condition, and in the vast majority of experimental designs we predict that the experimental conditions will differ from the control condition (or conditions). As such, the biggest hint that I can give you is that when planning comparisons the chances are that your first contrast should be one that compares all of the experimental groups with the control group (or groups). Once you have done this first comparison, any remaining comparisons will depend upon which of the experimental groups you predict will differ.

To illustrate these principles, Figures 10.6 and 10.7 show the contrasts that might be done in a four-group experiment. The first thing to notice is that in both scenarios there are three possible comparisons (one less than the number of groups). Also, every contrast compares only two chunks of variance. What's more, in both scenarios the first contrast is the same: the experimental groups are compared against the control group or groups. In Figure 10.6 there was only one control condition and so this portion of variance is used only in the first contrast (because it cannot be broken down any further). In Figure 10.7 there were two control groups, and so the portion of variance due to the control conditions (contrast 1) can be broken down again so as to see whether or not the scores in the control groups differ from each other (contrast 3).

In Figure 10.6, the first contrast contains a chunk of variance that is due to the three experimental groups and this chunk of variance is broken down by first looking at whether groups E1 and E2 differ from E3 (contrast 2). It is equally valid to use contrast 2 to compare groups E1 and E3 to E2, or to compare groups E2 and E3 to E1. The exact comparison that you choose depends upon your hypotheses. For contrast 2 in Figure 10.6 to be valid we need to have a good reason to expect group E3 to be different from the other two groups. The third comparison in Figure 10.6 depends on the comparison chosen for

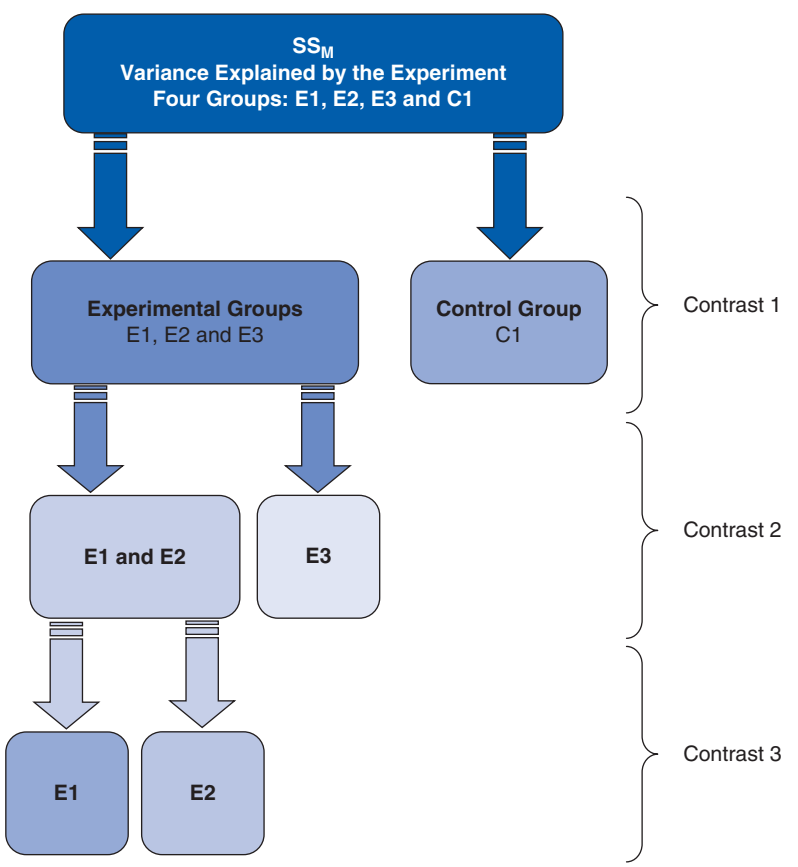


FIGURE 10.6 Partitioning variance for planned comparisons in a four-group experiment using one control group

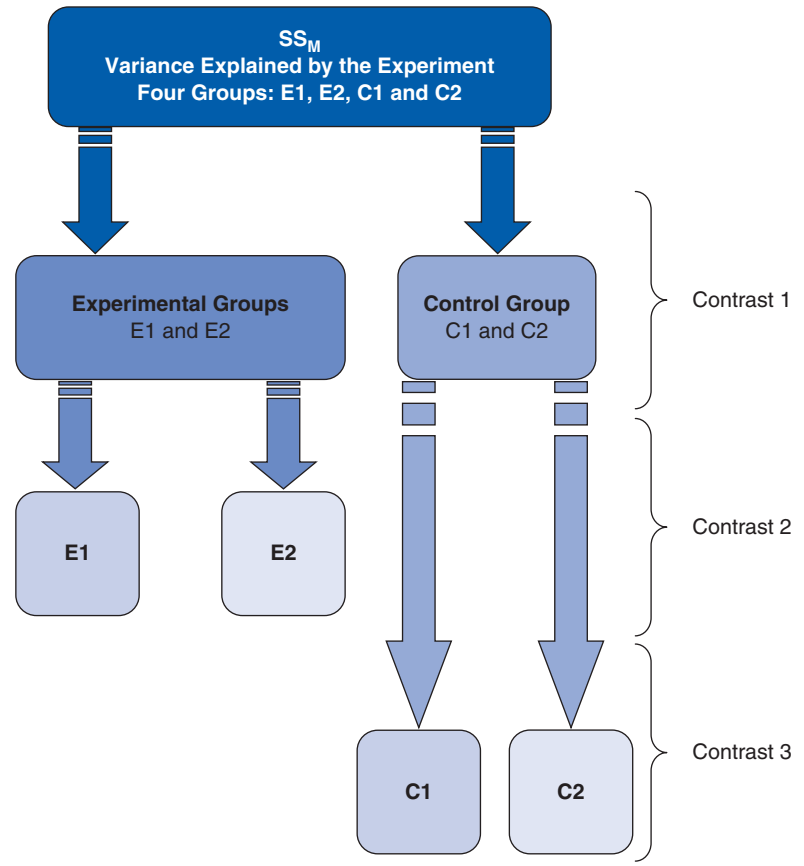


FIGURE 10.7 Partitioning variance for planned comparisons in a four-group experiment using two control groups

contrast 2. Contrast 2 necessarily had to involve comparing two experimental groups against a third, and the experimental groups chosen to be combined must be separated in the final comparison. As a final point, you'll notice that in Figures 10.6 and 10.7, once a group has been singled out in a comparison, it is never used in any subsequent contrasts.

When we carry out a planned contrast we compare 'chunks' of variance, and these chunks often consist of several groups. It is perhaps confusing to understand exactly what these contrasts tell us. Well, when you design a contrast that compares several groups to one other group, you are comparing the means of the groups in one chunk with the mean of the group in the other chunk. As an example, for the Viagra data I suggested that an appropriate first contrast would be to compare the two dose groups with the placebo group. The means of the groups are 2.20 (placebo), 3.20 (low dose) and 5.00 (high dose) and so the first comparison, which compared the two experimental groups to the placebo, is comparing 2.20 (the mean of the placebo group) to the average of the other two groups ((3.20 + 5.00)/2 = 4.10). If this first contrast turns out to be significant, then we can conclude that 4.10 is significantly greater than 2.20, which in terms of the experiment tells us that the average of the experimental groups is significantly different from the average of the controls. You can probably see that logically this means that, if the standard errors are the same, the experimental group with the highest mean (the high-dose group) will be significantly different from the mean of the placebo group. However, the experimental group with the lower mean (the low-dose group) might not necessarily differ from the placebo group;

we have to use the final comparison to make sense of the experimental conditions. For the Viagra data the final comparison looked at whether the two experimental groups differ (i.e., is the mean of the high-dose group significantly different from the mean of the low-dose group?). If this comparison turns out to be significant then we can conclude that having a high dose of Viagra significantly affected libido compared to having a low dose. If the comparison is non-significant then we have to conclude that the dosage of Viagra made no significant difference to libido. In this latter scenario it is likely that both doses affect libido more than placebo, whereas the former case implies that having a low dose may be no different than having a placebo. However, the word *implies* is important here: it is possible that the low-dose group might not differ from the placebo. To be completely sure we must carry out *post hoc* tests.

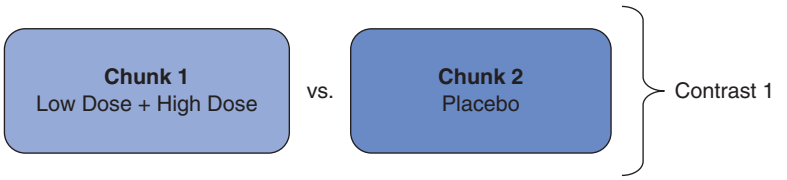
10.4.2. Defining contrasts using weights ②

Hopefully by now you have got some idea of how to plan which comparisons to do (i.e., if your brain hasn't exploded by now). Much as I'd love to tell you that all of the hard work is now over and **R** will magically carry out the comparisons that you've selected, it won't. To get **R** to carry out planned comparisons we need to tell it which groups we would like to compare, and doing this can be quite complex. In fact, when we carry out contrasts we assign values to certain variables in the regression model (sorry, I'm afraid that I have to start talking about regression again) – just as we did when we used dummy coding for the main ANOVA. To carry out contrasts we assign certain values to the dummy variables in the regression model. Whereas before we defined the experimental groups by assigning the dummy variables values of 1 or 0, when we perform contrasts we use different values to specify which groups we would like to compare. The resulting coefficients in the regression model (b_2 and b_1) represent the comparisons in which we are interested. The values assigned to the dummy variables are known as **weights**.

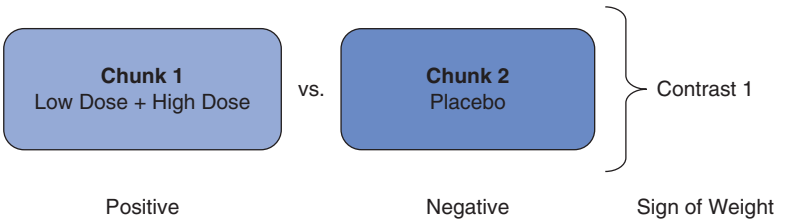
This procedure can seem horribly confusing, but there are a few basic rules for assigning values to the dummy variables to obtain the comparisons you want. I will explain these simple rules before showing how the process actually works. Remember the previous section when you read through these rules, and remind yourself of what I mean by a 'chunk' of variation.

- **Rule 1**: Choose sensible comparisons. Remember that you want to compare only two chunks of variation and that if a group is singled out in one comparison, that group should be excluded from any subsequent contrasts.
- **Rule 2**: Groups coded with positive weights will be compared against groups coded with negative weights. So, assign one chunk of variation positive weights and the opposite chunk negative weights.
- **Rule 3**: The sum of weights for a comparison should be zero. If you add up the weights for a given contrast the result should be zero.
- **Rule 4**: If a group is not involved in a comparison, automatically assign it a weight of 0. If we give a group a weight of 0 then this eliminates that group from all calculations.
- **Rule 5**: For a given contrast, the weights assigned to the group(s) in one chunk of variation should be equal to the number of groups in the opposite chunk of variation.

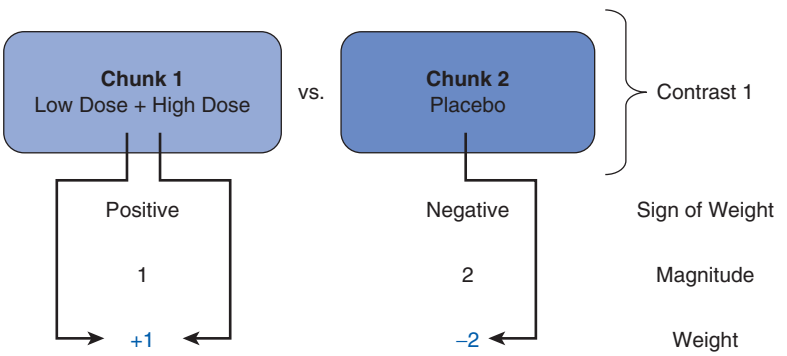
OK, let's follow some of these rules to derive the weights for the Viagra data. The first comparison we chose was to compare the two experimental groups against the control:



Therefore, the first chunk of variation contains the two experimental groups, and the second chunk contains only the placebo group. Rule 2 states that we should assign one chunk positive weights, and the other negative. It doesn't matter which way round we do this, but for convenience let's assign chunk 1 positive weights, and chunk 2 negative weights:



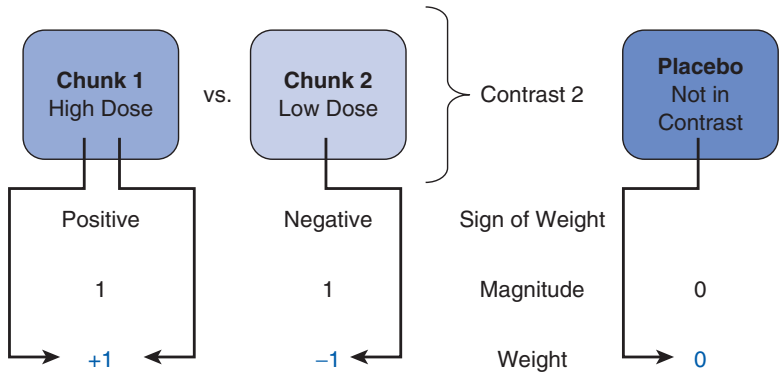
Using rule 5, the weight we assign to the groups in chunk 1 should be equivalent to the number of groups in chunk 2. There is only one group in chunk 2 and so we assign each group in chunk 1 a weight of 1. Likewise, we assign a weight to the group in chunk 2 that is equal to the number of groups in chunk 1. There are two groups in chunk 1 so we give the placebo group a weight of 2. Then we combine the sign of the weights with the magnitude to give us weights of −2 (placebo), 1 (low dose) and 1 (high dose):



Rule 3 states that for a given contrast, the weights should add up to zero, and by following rules 2 and 5 this rule will always be followed (if you haven't followed these rules properly then this will become clear when you add the weights). So, let's check by adding the weights: sum of weights = 1 + 1 − 2 = 0.

The second contrast was to compare the two experimental groups and so we want to ignore the placebo group. Rule 4 tells us that we should automatically assign this group a weight of 0 (because this will eliminate this group from any calculations). We are left with two chunks of variation: chunk 1 contains the low-dose group and chunk 2 contains the high-dose group. By following rules 2 and 5 it should be obvious that one group is assigned

a weight of +1 while the other is assigned a weight of −1. The control group is ignored (and so given a weight of 0). If we add the weights for contrast 2 we should find that they again add up to zero: sum of weights = 1 − 1 + 0 = 0.



The weights for each contrast are codings for the two dummy variables in equation (10.2). Hence, these codings can be used in a multiple regression model in which b_2 represents contrast 1 (comparing the experimental groups to the control), b_1 represents contrast 2 (comparing the high-dose group to the low-dose group), and b_0 is the grand mean:

$$\text{libido}_i = b_0 + b_1\text{contrast}_{1i} + b_2\text{contrast}_{2i} \tag{10.9}$$

Each group is specified now not by the 0 and 1 coding scheme that we initially used, but by the coding scheme for the two contrasts. A code of −2 for contrast 1 and a code of 0 for contrast 2 identifies participants in the placebo group. Likewise, the high-dose group is identified by a code of 1 for both variables, and the low-dose group has a code of 1 for one contrast and a code of −1 for the other (see Table 10.4).

Table 10.4 Orthogonal contrasts for the Viagra data

Group	Dummy variable 1 (contrast$_1$)	Dummy variable 2 (contrast$_2$)	Product contrast$_1$ × contrast$_2$
Placebo	–2	0	0
Low dose	1	–1	–1
High dose	1	1	1
Total	0	0	0

It is important that the weights for a comparison sum to zero because it ensures that you are comparing two unique chunks of variation. Therefore, we can perform a *t*-test. A more important consideration is that when you multiply the weights for a particular group, these products should also add up to zero (see the final column of Table 10.4). If the products add to zero then we can be sure that the contrasts are *independent* or **orthogonal**. It is important

for interpretation that contrasts are orthogonal. When we used dummy variable coding and ran a regression on the Viagra data, I commented that we couldn't look at the individual *t*-tests done on the regression coefficients because the familywise error rate is inflated (see section 10.4). However, if the contrasts are independent then the *t*-tests done on the *b* coefficients are also independent and so the resulting *p*-values are uncorrelated. You might think that it is very difficult to ensure that the weights you choose for your contrasts conform to the requirements for independence but, provided you follow the rules I have laid out, you should always derive a set of *orthogonal* comparisons. You should double-check by looking at the sum of the multiplied weights and if this total is not zero then go back to the rules and see where you have gone wrong (see the final column of Table 10.4).

Earlier on, I mentioned that when you used contrast codings in dummy variables in a regression model the *b*-values represented the differences between the means that the contrasts were designed to test. Although it is reasonable for you to trust me on this issue, for the more advanced students I'd like to take the trouble to show you how the regression model works (this next part is not for the faint-hearted and so those with an equation phobia should move onto the next section!). When we do planned contrasts, the intercept b_0 is equal to the grand mean (i.e., the value predicted by the model when group membership is not known), which when group sizes are equal is:

$$b_0 = \text{grand mean} = \frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}}{3}$$

Placebo group: If we use the contrast codings for the placebo group (see Table 10.4), the predicted value of libido equals the mean of the placebo group. The regression equation can, therefore, be expressed as:

$$\text{libido}_i = b_0 + b_1\text{contrast}_1 + b_2\text{contrast}_2$$
$$\bar{X}_{\text{placebo}} = \left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}}{3}\right) + (-2b_1) + (b_2 \times 0)$$

Now, if we rearrange this equation and then multiply everything by 3 (to get rid of the fraction) we get:

$$2b_1 = \left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}}{3}\right) - \bar{X}_{\text{placebo}}$$
$$6b_1 = \bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}} - 3\bar{X}_{\text{placebo}}$$
$$6b_1 = \bar{X}_{\text{high}} + \bar{X}_{\text{low}} - 2\bar{X}_{\text{placebo}}$$

We can then divide everything by 2 to reduce the equation to its simplest form:

$$3b_1 = \left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}}}{2}\right) - \bar{X}_{\text{placebo}}$$
$$b_1 = \frac{1}{3}\left[\left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}}}{2}\right) - \bar{X}_{\text{placebo}}\right]$$

This equation shows that b_1 represents the difference between the average of the two experimental groups and the control group:

$$\begin{aligned} 3b_1 &= \left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}}}{2}\right) - \bar{X}_{\text{placebo}} \\ &= \frac{5 + 3.2}{2} - 2.2 \\ &= 1.9 \end{aligned}$$

We planned contrast 1 to look at the difference between the average of the experimental groups and the control, and so it should now be clear how b_1 represents this difference. The observant among you will notice that rather than being the true value of the difference between experimental and control groups, b_1 is actually a third of this difference ($b_1 = 1.9/3 = 0.633$). The reason for this division is that the familywise error is controlled by making the regression coefficient equal to the actual difference divided by the number of groups in the contrast (in this case 3).

High-dose group: For the situation in which the codings for the high-dose group (see Table 10.4) are used, the predicted value of libido is the mean for the high-dose group, and so the regression equation becomes:

$$\begin{aligned} \text{libido}_i &= b_0 + b_1\text{contrast}_1 + b_2\text{contrast}_2 \\ \bar{X}_{\text{high}} &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ b_2 &= \bar{X}_{\text{high}} - b_1 - b_0 \end{aligned}$$

We know already what b_1 and b_0 represent, so we place these values into the equation and then multiply by 3 to get rid of some of the fractions:

$$\begin{aligned} b_2 &= \bar{X}_{\text{high}} - b_1 - b_0 \\ b_2 &= \bar{X}_{\text{high}} - \left\{\frac{1}{3}\left[\left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}}}{2}\right) - \bar{X}_{\text{placebo}}\right]\right\} - \left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}}{3}\right) \\ 3b_2 &= 3\bar{X}_{\text{high}} - \left[\left(\frac{\bar{X}_{\text{high}} + \bar{X}_{\text{low}}}{2}\right) - \bar{X}_{\text{placebo}}\right] - (\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}) \end{aligned}$$

If we multiply everything by 2 to get rid of the other fraction, expand all of the brackets and then simplify the equation we get:

$$\begin{aligned} 6b_2 &= 6\bar{X}_{\text{high}} - (\bar{X}_{\text{high}} + \bar{X}_{\text{low}} - 2\bar{X}_{\text{placebo}}) - 2(\bar{X}_{\text{high}} + \bar{X}_{\text{low}} + \bar{X}_{\text{placebo}}) \\ &= 6\bar{X}_{\text{high}} - \bar{X}_{\text{high}} - \bar{X}_{\text{low}} + 2\bar{X}_{\text{placebo}} - 2\bar{X}_{\text{high}} - 2\bar{X}_{\text{low}} - 2\bar{X}_{\text{placebo}} \\ &= 3\bar{X}_{\text{high}} - 3\bar{X}_{\text{low}} \end{aligned}$$

Finally, we can divide the equation by 6 to find out what b_2 represents (remember that 3/6 = 1/2):

$$b_2 = \frac{1}{2}(\bar{X}_{\text{high}} - \bar{X}_{\text{low}})$$

We planned contrast 2 to look at the difference between the experimental groups:

$$\overline{X}_{\text{high}} - \overline{X}_{\text{low}} = 5 - 3.2 = 1.8$$

It should now be clear how b_2 represents this difference. Again, rather than being the absolute value of the difference between the experimental groups, b_2 is actually half of this difference (1.8/2 = 0.9). The familywise error is again controlled, by making the regression coefficient equal to the actual difference divided by the number of groups in the contrast (in this case 2).

SELF-TEST

- ✓ To illustrate these principles, I have created a file called **Contrast.dat** in which the Viagra data are coded using the contrast coding scheme used in this section. Run multiple regression analyses on these data using **libido** as the outcome and using **dummy1** and **dummy2** as the predictor variables (leave all default options).

Output 10.2 shows the result of this regression. The *F*-statistic for the model is the same as when dummy coding was used (compare it to Output 10.1), showing that the model fit is the same (it should be because the model represents the group means and these have not changed); however, the regression coefficients have now changed. The first thing to notice is that the intercept is the grand mean, 3.467 (see, I wasn't telling lies). Second, the regression coefficient for contrast 1 is one-third of the difference between the average of the experimental conditions and the control condition (see above). Finally, the regression coefficient for contrast 2 is half of the difference between the experimental groups (see above). So, when a planned comparison is done in ANOVA a *t*-test is conducted comparing the mean of one chunk of variation with the mean of a different chunk. From the significance values of the *t*-tests we can see that our experimental groups were significantly different from the control ($p < .05$) but that the experimental groups were not significantly different ($p > .05$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4667     0.3621   9.574 5.72e-07 ***
dummy1        0.6333     0.2560   2.474   0.0293 *
dummy2        0.9000     0.4435   2.029   0.0652 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.402 on 12 degrees of freedom
Multiple R-squared: 0.4604,       Adjusted R-squared: 0.3704
F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Output 10.2

CRAMMING SAM'S TIPS Planned contrasts

- After an ANOVA you need more analysis to find out which groups differ.
- When you have generated specific hypotheses before the experiment, use *planned contrasts*.
- Each contrast compares two 'chunks' of variance. (A chunk can contain one or more groups.)
- The first contrast will usually be experimental groups vs. control groups.
- The next contrast will be to take one of the chunks that contained more than one group (if there were any) and divide it in to two chunks.
- You then repeat this process: if there are any chunks in previous contrasts that contained more than one group that haven't already been broken down into smaller chunks, then create a new contrast that breaks it down into smaller chunks.
- Carry on creating contrasts until each group has appeared in a chunk on its own in one of your contrasts.
- You should end up with one less contrast than the number of experimental conditions. If not, you've done it wrong.
- In each contrast assign a 'weight' to each group that is the value of the number of groups in the opposite chunk in that contrast.
- For a given contrast, randomly select one chunk, and for the groups in that chunk change their weights to be negative numbers.
- Breathe a sigh of relief.

10.4.3. Non-orthogonal comparisons ②

I have spent a lot of time labouring how to design appropriate orthogonal comparisons without mentioning the possibilities that non-orthogonal contrasts provide. Non-orthogonal contrasts are comparisons that are in some way related, and the best way to get them is to disobey Rule 1 in the previous section. Using my cake analogy again, non-orthogonal comparisons are where you slice up your cake and then try to stick slices of cake together again. So, for the Viagra data a set of non-orthogonal contrasts might be to have the same initial contrast (comparing experimental groups against the placebo), but then to compare the high-dose group to the placebo. This disobeys rule 1 because the placebo group is singled out in the first contrast but used again in the second contrast. The coding for this set of contrasts is shown in Table 10.5, and by looking at the last column it is clear that when you multiply and add the codings from the two contrasts the sum is not zero. This tells us that the contrasts are not orthogonal.

Table 10.5 Non-orthogonal contrasts for the Viagra data

Group	Dummy variable 1 ($Contrast_1$)	Dummy variable 2 ($Contrast_2$)	Product $Contrast_1 \times Contrast_2$
Placebo	–2	–1	2
Low dose	1	0	0
High dose	1	1	1
Total	0	0	3

There is nothing intrinsically wrong with performing non-orthogonal contrasts. However, if you choose to perform this type of contrast you must be very careful about how you interpret the results. With non-orthogonal contrasts, the comparisons you do are related and so the resulting test statistics and *p*-values will be correlated to some extent. For this reason you should use a more conservative probability level to accept that a given contrast is statistically meaningful (see section 10.5).

10.4.4. Standard contrasts ②

Although under most circumstances you will design your own contrasts, there are special contrasts that have been designed to compare certain situations. Some of these contrasts are orthogonal, whereas others are non-orthogonal.

Table 10.6 shows the contrasts that are available in **R** using the **contrasts()** function. This function is used to code any categorical variable and the resulting codings can be used in pretty much any linear model (ANOVA, regression, logistic regression, etc.). Although the exact codings are not provided in Table 10.6, examples of the comparisons done in a three- and four-group situation are given (where the groups are labelled 1, 2, 3 and 1, 2, 3, 4, respectively). When you code variables **R** will treat the lowest-value code as group 1, the next highest code as group 2, and so on. Therefore, depending on which comparisons you want to make you should code your grouping variable appropriately (and then use Table 10.6 as a guide to which comparisons **R** will carry out). One thing that clever readers might notice about the contrasts in Table 10.6 is that some are orthogonal (i.e., Helmert contrasts) while others are non-orthogonal (e.g., treatment). You might also notice that the comparisons calculated using treatment contrasts are the same as those given by using the dummy variable coding described in Table 10.2).

Table 10.6 Standard contrasts available in **R**

Name	*Definition*	*Contrast*	*Three Groups*	*Four Groups*
Dummy (default)	The default is dummy coding in which each category is compared to the first category	1	1 vs. 2	1 vs. 2
		2	1 vs. 3	1 vs. 3
		3		1 vs. 4
contr. treatment()	Each category is compared to a user-defined baseline category (in this case I chose the second category)	1	2 vs. 1	2 vs. 1
		2	2 vs. 3	2 vs. 3
		3		2 vs. 4
contr. SAS()	Each category is compared to the last category	1	1 vs. 3	1 vs. 4
		2	2 vs. 3	2 vs. 4
		3		3 vs. 4
contr. helmert()	Each category (except the last) is compared to the mean effect of all subsequent categories	1	1 vs. (2, 3)	1 vs. (2, 3, 4)
		2	2 vs. 3	2 vs. (3, 4)
		3		3 vs. 4

10.4.5. Polynomial contrasts: trend analysis ②

One type of contrast deliberately omitted from Table 10.6 is the **polynomial contrast**, which can be obtained using **contr.poly()**. This contrast tests for trends in the data, and in its most basic form it looks for a linear trend (i.e., that the group means increase proportionately). However, there are other trends such as quadratic, cubic and quartic trends that can be examined. Figure 10.8 shows examples of the types of trend that can exist in data sets. The *linear* trend should be familiar to you all by now and represents a simple proportionate change in the value of the dependent variable across ordered categories (the diagram shows a positive linear trend, but of course it could be negative). A **quadratic trend** is where there is one change in the direction of the line (e.g., the line is curved in one place). An example of this might be a situation in which a drug enhances performance on a task at first, but then as the dose increases the performance drops again. To find a quadratic trend you need at least three groups (because in the two-group situation there are not enough categories of the independent variable for the means of the dependent variable to change one way and then another). A **cubic trend** is where there are two changes in the direction of the trend. So, for example, the mean of the dependent variable at first goes up across the first couple of categories of the independent variable, then across the succeeding categories the means go down, but then across the last few categories the means rise again. To have two changes in the direction of the mean you must have at least four categories of the independent variable. The final trend that you are likely to come across is the **quartic trend**, and this trend has three changes of direction (so you need at least five categories of the independent variable).

Polynomial trends should be examined in data sets in which it makes sense to order the categories of the independent variable (so, for example, if you have administered five doses of a drug it makes sense to examine the five doses in order of magnitude). For the Viagra data there are only three groups and so we can expect to find only a linear or quadratic trend (and it would be pointless to test for any higher-order trends).

Each of these trends has a set of codes for the dummy variables in the regression model, so we are doing the same thing that we did for planned contrasts except that the codings have already been devised to represent the type of trend of interest. In fact, the graphs in Figure 10.8

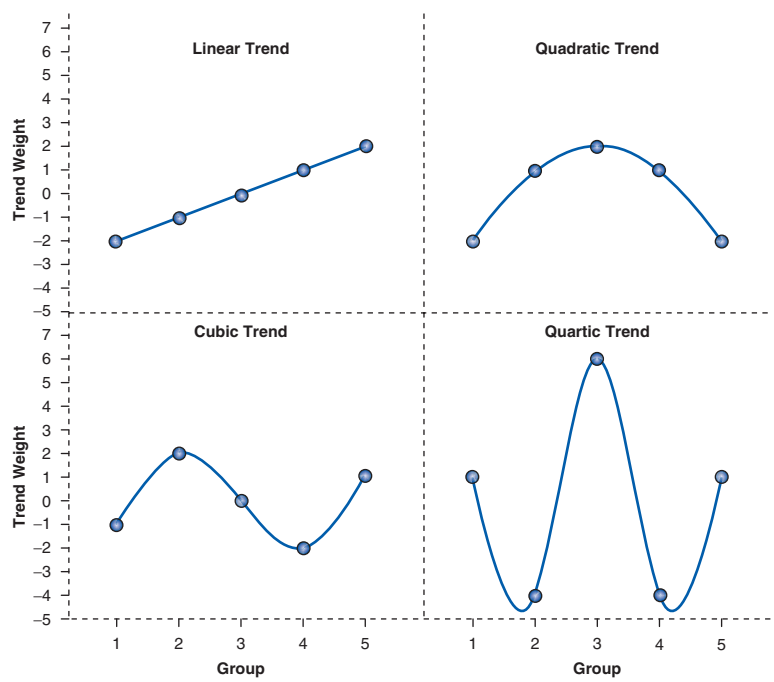


FIGURE 10.8 Linear, quadratic, cubic and quartic trends across five groups

have been constructed by plotting the coding values for the five groups. Also, if you add the codes for a given trend the sum will equal zero and if you multiply the codes you will find that the sum of the products also equals zero. Hence, these contrasts are orthogonal. The great thing about these contrasts is that you don't need to construct your own coding values to do them, because the codings already exist.

10.5. *Post* hoc procedures ②

Often it is the case that you have no specific *a priori* predictions about the data you have collected and instead you are interested in exploring the data for any between-group differences between means that exist. This procedure is sometimes called *data mining* or *exploring data*. Now, personally I have always thought that these two terms have certain 'rigging the data' connotations to them and so I prefer to think of these procedures as 'finding the differences that I should have predicted if only I'd been clever enough'.

Post hoc tests consist of **pairwise comparisons** that are designed to compare all different combinations of the treatment groups. So, it is rather like taking every pair of groups and then performing a *t*-test on each pair of groups. Now, this might seem like a particularly stupid thing to say in the light of what I have already told you about the problems of inflated familywise error rates. However, pairwise comparisons control the familywise error by correcting the level of significance for each test such that the overall Type I error rate (α) across all comparisons remains at .05. There are several ways in which the familywise error rate can be controlled. The most popular (and easiest) way is to divide α by the number of comparisons, k, thus ensuring that the cumulative Type I error is below .05:

$$p_{crit} = \frac{\alpha}{k}$$

Therefore, if we conduct 10 tests, we use .005 as our criterion for significance. This method is known as the **Bonferroni correction** (Figure 10.9). There is a trade-off for controlling the familywise error rate, and that is a loss of statistical power. This means that the probability of rejecting an effect that does actually exist is increased (this is called a Type II error). By being more conservative in the Type I error rate for each comparison, we increase the chance that we will miss a genuine difference in the data.

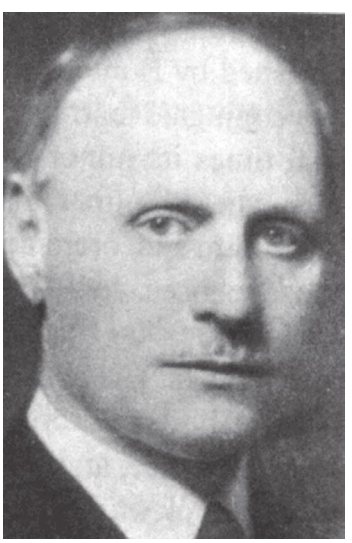


FIGURE 10.9
Carlo Bonferroni before the celebrity of his correction led to drink, drugs and statistics groupies

Let's look at this method, and some variations, using an example. Some research has suggested that children wearing superhero costumes might be more likely to harm themselves because of the unrealistic impression of invincibility that these costumes could create. For example, there are case studies of children reporting to hospital with severe injuries because of falling from windows or trying 'to initiate flight without having planned for landing strategies' (Davies, Surridge, Hole, & Munro-Davies, 2007). Having spent a lot of my childhood dressed in various costumes, I can relate to the imagined power that it bestows upon you; even now, I have been known to dress up as Fisher by donning a beard and glasses and trailing a goat around on a lead in the hope that it might make me more knowledgeable about statistics.

Imagine that we wanted to see whether different *types* of superhero costumes led to more severe injuries. We measured the severity of injury on a scale from 0 to 100 (0 = not at all injured, 100 = dead), and made a note of the type of costume a child was wearing. Let's also entertain the possibility that children fell (probably because of trying to fly) into four groups: Spiderman, Superman, the Hulk, and Teenage Mutant Ninja Turtles (let's face it, who wouldn't want to dress up as a ninja turtle?). These entirely fabricated data are in **Superhero.dat**. There is a task at the end of the chapter to analyse these data, but for now, let's look at comparing all of these groups; we would end up with the six comparisons in Table 10.7. The table shows the unadjusted *p*-value that you get for each comparison. The critical value of p_{crit} based on a Bonferroni correction for each comparison is the Type I error rate divided by the number of comparisons, $\alpha/k = .05/6 = .0083$. If the observed *p* is smaller than the critical value then the comparison is significant (at $\alpha = .05$). In this case, there is a significant difference between ninja turtle and Superman costumes (because .0000 is less than .0083) and between Superman and Hulk costumes (because .0014 is smaller than .0083). In all other cases *p* is bigger than the critical value so the difference is not significant.

Table 10.7 Critical values for *p* based on variations on Bonferroni (* indicates that a comparison is significant)

		Bonferroni			Holm			Benjamini–Hochberg	
	p	$p_{crit}=\frac{\alpha}{k}$		*j*	$p_{crit}=\frac{\alpha}{j}$		*j*	$p_{crit}=\left(\frac{j}{k}\right)\alpha$	
NT–Super	.0000	.0083	*	6	.0083	*	1	.0083	*
Super–Hulk	.0014	.0083	*	5	.0100	*	2	.0167	*
Spider–Super	.0127	.0083		4	.0125		3	.0250	*
NT–Spider	.0252	.0083		3	.0167		4	.0333	*
NT–Hulk	.1704	.0083		2	.0250		5	.0417	
Spider–Hulk	.3431	.0083		1	.0500		6	.0500	

There are various improvements that have been made to the Bonferroni correction over the years and the general principle behind them is easy to understand so it's worth explaining. In an attempt to make the Bonferroni correction less conservative (i.e., to make it better at detecting differences that actually exist), authors such as Hommel, Hochberg and Holm[5] have suggested stepped approaches (Hochberg, 1988; Holm, 1979; Hommel,

[5] Their names all begin with 'Ho', which I find a strange coincidence. If your surname begins with 'Ho' too, beware: a life in multiple comparison research could await you.

1988). Holm's method is very simple to explain. You begin by computing the *p*-value for all of the pairs of groups in your data, you then order them from smallest to largest. We assign each *p* in the list an index (I've labelled it *j*) that tells us where in the list it falls. Table 10.7 shows this process: for the largest *p* we assign an index of 1, the next largest 2, and so on until the smallest one, which will be indexed as the number of comparisons (*k*), in this case 6. The critical value for a given comparison is the Type I error rate divided by the index variable (*j*):

$$p_{\text{crit}} = \frac{\alpha}{j}$$

Starting from the smallest *p*-value, this means that you begin with the normal Bonferroni correction because $j = k$ for this first comparison. However, notice that in subsequent comparisons we do not correct for every comparison made, instead we correct *only for the remaining comparisons*. Unlike the standard Bonferroni correction, the critical value of *p* gets bigger (and less conservative) for each comparison. The key idea behind this method is it is *stepped*. This means that as long as a comparison is significant, we proceed to the next one, but at the point that we encounter a non-significant comparison we *stop and assume that all remaining comparisons are non-significant also*. In Table 10.7, we see a significant difference between Ninja Turtle and Superman costumes (because .0000 is less than .0083); therefore, we move onto the next one down and see a significant difference between Superman and Hulk costumes (because .0014 is smaller than .01); therefore we move down again but find a non-significant difference between Spiderman and Superman costumes (because .0127 is larger than .0125); because of this non-significance we stop and do not consider any further comparisons.

A more modern take on this kind of sequential approach to multiple comparisons is to worry not about the familywise error rate, but to focus on the *false discovery rate (FDR)*. By focusing on the familywise error rate we are obsessing (in some people, literally) about the possibility of making one or more Type I errors. The corresponding belief system can be summed up as 'if I make even one Type I error then my entire set of conclusions is meaningless'. With a belief system like that it's no wonder people look depressed when they're analysing data. Benjamini and Hochberg think about things differently. Their belief system can be summed up as the rather more joyful 'let's try to estimate how many Type I errors (or false discoveries) we have made'. The FDR is simply the proportion of falsely rejected null hypotheses:

$$\text{FDR} = \frac{\text{number of falsely rejected null hypotheses}}{\text{total number of rejected null hypotheses}}$$

As such, the FDR approach to multiple comparisons is less strict than Bonferroni-based methods because it is concerned with keeping the FDR rather than the familywise error rate under control. In Benjamini and Hochberg's method (Benjamini & Hochberg, 1995, 2000) you start by computing the *p*-value for all of the pairs of groups in your data. You then order them and, as with Holm's method, index the order with the letter *j* (notice we order them the opposite way around to Holm's method). For each comparison you deem it significant if the observed *p* is smaller than a critical value defined as:

$$p_{\text{crit}} = \frac{j}{k}\alpha$$

Table 10.7 again shows this process. For the largest *p*-value we again have the normal Bonferroni correction (i.e., α/k), for the other comparisons we use a more liberal criterion.

Like Holm's method this procedure is stepped; however, rather than working down the table we work up (hence it is known as a 'step-up' procedure). So, we begin at the bottom and conclude a non-significant difference between Spiderman and Hulk costumes (because .3431 is greater than .05); given this non-significance we move up the table and see a non-significant difference between Ninja Turtle and Hulk costumes (because .1704 is greater than .0417); given this non-significance we again move up the table and see a significant difference between Ninja Turtle and Spiderman costumes (because .0252 is less than the critical value of .0333); because of this significance we stop and *assume that all other comparisons are also significant*. Procedurally this step-up approach is the opposite of Holm's step-down procedure.

There are many other *post hoc* procedures. I have explained only a few of the main ones that can be implemented in **R**. I could go into all of the other methods in tedious detail but there are some excellent texts already available for those who wish to know (Klockars & Sax, 1986; Toothaker, 1993) and **R** does not implement most of them anyway. (That said, the nice thing about **R** of course is that you could write your own function to do them if you had a few spare hours, a maths degree, and a bottle of gin.) However, it *is* important that you have an idea of which *post hoc* tests perform best. 'Best' is a word that can mean many things. For *post hoc* procedures, deciding on what's 'best' requires us to consider three things: whether the test controls the Type I error rate; whether the test controls the Type II error rate (i.e., has good statistical power); and whether the test is reliable when the test assumptions of ANOVA have been violated.

10.5.1. *Post hoc* procedures and Type I (α) and Type II error rates ②

The Type I error rate and the statistical power of a test are linked. Therefore, there is always a trade-off: if a test is conservative (the probability of a Type I error is small) then it is likely to lack statistical power (the probability of a Type II error will be high). So, it is important that multiple comparison procedures control the Type I error rate but without a substantial loss in power. If a test is too conservative then we are likely to reject differences between means that are, in reality, meaningful.

Bonferroni's and *Tukey's* HSD[6] tests both control the Type I error rate very well but are conservative tests (they lack statistical power). Of the two, Bonferroni has more power when the number of comparisons is small, whereas Tukey is more powerful when testing large numbers of means. Tukey generally has greater power than other tests of which you might have heard such as *Dunn* and *Scheffé*. Holm's method should have more power than Bonferroni, and the Benjamini–Hochberg method should have more power than Holm's procedure. If you are obsessed with controlling the Type I error rate, it is worth remembering that the Benjamini–Hochberg method does not attempt to do this: it controls the FDR.

10.5.2. *Post hoc* procedures and violations of test assumptions ②

Most research on *post hoc* tests has looked at whether the test performs well when the group sizes are different (an unbalanced design), when the population variances are very

[6] HSD stands for 'honest significant difference',which has a slightly dodgy ring to it if you ask me!

different, and when data are not normally distributed. The good news is that most multiple comparison procedures perform relatively well under small deviations from normality. The bad news is that they perform badly when group sizes are unequal and when population variances are different.

There are a variety of tests designed to deal with these situations, none of which are implemented in **R**. *Hochberg's GT2* is one such test and is worth mentioning because it is not implemented in **R** and is completely different than the Hochberg and Benjamini–Hochberg methods that I have already mentioned. Therefore, don't use the Hochberg option in **R** thinking it can cope with unequal variances: it is a different test.

Instead of telling you what can't be done, it might be more helpful to tell you what *can* be done. There are some robust methods that have been implemented in **R** by Wilcox (2005). As with methods for the ANOVA itself, these methods are based on bootstrapping or trimmed means and M-estimators (both of which can also include a bootstrap). All of these methods are very new and so there is very little on which to base advice on what to do for the best. However, all methods have been shown to control the Type I error well when applied to some very extreme distributions. If Type I error control is your main concern then the bootstrap seems to offer a small advantage, and if power is your concern then there are some benefits to methods based on M-estimators (Wilcox, 2003). However, the bottom line is that using any of these methods is undoubtedly better than using a non-robust method.

10.5.3. Summary of *post hoc* procedures ②

The choice of comparison procedure will depend on the exact situation you have and whether it is more important for you to keep strict control over the familywise error rate, the FDR, or to have greater statistical power. However, some general guidelines can be drawn (Toothaker, 1993). When you have equal sample sizes and you are confident that your population variances are similar then Tukey has good power and tight control over the Type I error rate. Bonferroni is generally conservative, but if you want guaranteed control over the Type I error rate then this is the test to use. If there is any doubt over the underlying assumptions (e.g., unequal population variances) then use a robust method based on a bootstrap, trimmed means, or M-estimators.

CRAMMING SAM'S TIPS *Post hoc* tests

- After an ANOVA you need a further analysis to find out which groups differ.
- When you have no specific hypotheses before the experiment, use *post hoc tests*.
- When you have equal sample sizes and group variances are similar, use Tukey.
- If you want guaranteed control over the Type I error rate, then use Bonferroni.
- If there is any doubt that group variances are equal, then use a robust method (e.g., bootstrap or trimmed means).

10.6. One-way ANOVA using R ②

Hopefully you should all have some appreciation for the theory behind ANOVA, so let's put that theory into practice by conducting an ANOVA test on the Viagra data.

10.6.1. Packages for one-way ANOVA in R ①

There are several packages that we will use in this chapter. If you're using R Commander (see the next section) then you don't need to worry: it will load everything it needs automatically. If you're using commands (which we recommend), you will need the packages *car* (for Levene's test), *compute.es* (for effect sizes) *ggplot2* (for graphs), *multcomp* (for *post hoc* tests), *pastecs* (for descriptive statistics), and *WRS* (for robust tests). If you do not have these packages installed (some should be installed from previous chapters), you can install them by executing the following commands:

```
install.packages("compute.es"); install.packages("car"); install.packages
("ggplot2");install.packages("multcomp");install.packages("pastecs");install.
packages("WRS", repos="http://R-Forge.R-project.org")
```

You then need to load these packages by executing these commands:

```
library(compute.es); library(car); library(ggplot2); library(multcomp);
library(pastecs); library(WRS)
```

10.6.2. General procedure for one-way ANOVA ①

To conduct one-way ANOVA you should follow this general procedure:

1. *Enter data*: obviously you need to enter your data.
2. *Explore your data*: as with any analysis, it's a good idea to begin by graphing your data and computing some descriptive statistics. You should also check distributional assumptions and use Levene's test to check for homogeneity of variance (see Chapter 5).
3. *Compute the basic ANOVA*: you can then run the main analysis of variance. Depending on what you found in the previous step, you might need to run a robust version of the test.
4. *Compute contrasts or post hoc tests*: having conducted the main ANOVA you can follow it up with either contrasts or *post hoc* tests. Again, the exact methods you choose will depend upon what you unearth in step 2.

We will work through these steps in turn.

10.6.3. Entering data ①

As with the independent *t*-test, we need to enter the data into **R** using a coding variable to specify to which of the three groups the data belong. So, the data must be entered in two columns (one called **dose** which specifies how much Viagra the participant was given and one called **libido** which indicates the person's libido over the following week). The data are in the file **Viagra.dat**, but I recommend entering them by hand to gain practice in data entry. I have coded the grouping variable so that 1 = placebo, 2 = low dose and 3 = high dose (see section 3.5.4.3).

This data set is small (only 15 cases); therefore, we could enter the data directly into **R** by executing the following code:

```
libido<-c(3,2,1,1,4,5,2,4,2,3,7,4,5,3,6)
dose<-gl(3,5, labels = c("Placebo", "Low Dose", "High Dose"))
viagraData<-data.frame(dose, libido)
```

These commands create a variable called **libido** with the 15 libido scores contained within it, and a variable called **dose**, which uses the *gl()* function to create a factor variable with three groups each containing five participants. These variables are merged into a dataframe called *viagraData*. We can look at the contents of the dataframe by executing:

```
viagraData
```

You will see the following displayed in the console:

```
        dose libido
1    Placebo      3
2    Placebo      2
3    Placebo      1
4    Placebo      1
5    Placebo      4
6   Low Dose      5
7   Low Dose      2
8   Low Dose      4
9   Low Dose      2
10  Low Dose      3
11 High Dose      7
12 High Dose      4
13 High Dose      5
14 High Dose      3
15 High Dose      6
```

10.6.4. One-way ANOVA using R Commander ②

Running ANOVA using commands gives you much more versatility than R Commander. However, you can do a basic one-way ANOVA using R Commander. First load the data from the file **Viagra.dat** by using the **Data⇒Import data⇒from text file, clipboard, or URL…** menu (see section 3.7.3). This data set has two variables: **dose**, which is the grouping variable (1 = placebo, 2 = low dose, 3 = high dose); and **libido**, which is each participant's libido score. Once the data are loaded in a dataframe (I have called the dataframe *viagraData*), you need to convert the variable **dose** into a factor – see section 3.6.2 to remind yourself how to do that.

Once you have done that, you need to explore the data: get some descriptive statistics and test the assumptions. This is explained in Chapter 5. Levene's test looks at whether variances across conditions are equal – in other words, it tests the assumption of homogeneity of variance (see section 10.3.1). Use the **Statistics⇒Variances⇒Levene's test…** menu to run the analysis. The resulting dialog box is fairly self-explanatory (Figure 10.10): select a factor from the list labelled *Groups* (in this case we have only one factor, **dose**) and select the outcome variable from the list labelled *Response Variable* (in this case **libido**). By default, R Commander will base Levene's test on deviations from the median, which is a better measure than using deviations from the mean, but you can change this option if you like. Click on OK to run the analysis. The resulting output is described in section 10.6.5.

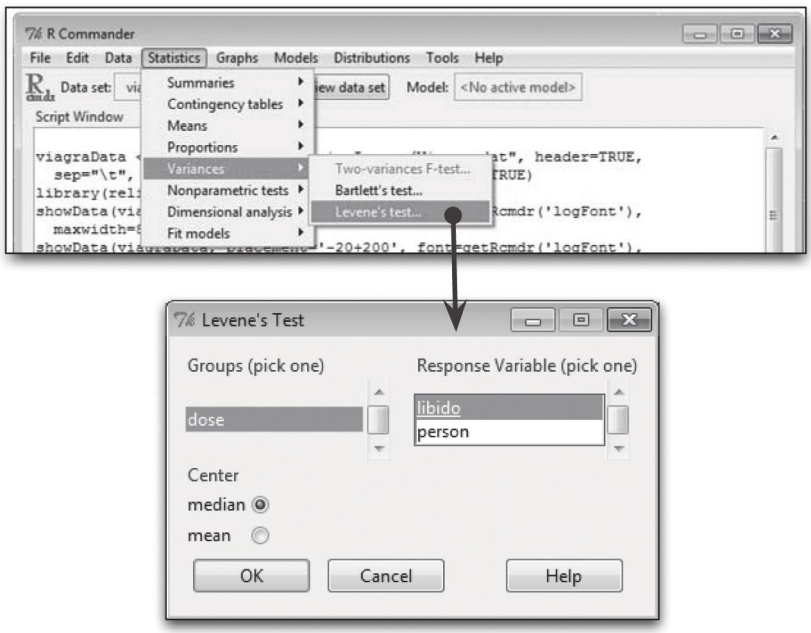


FIGURE 10.10 Levene's test using R Commander

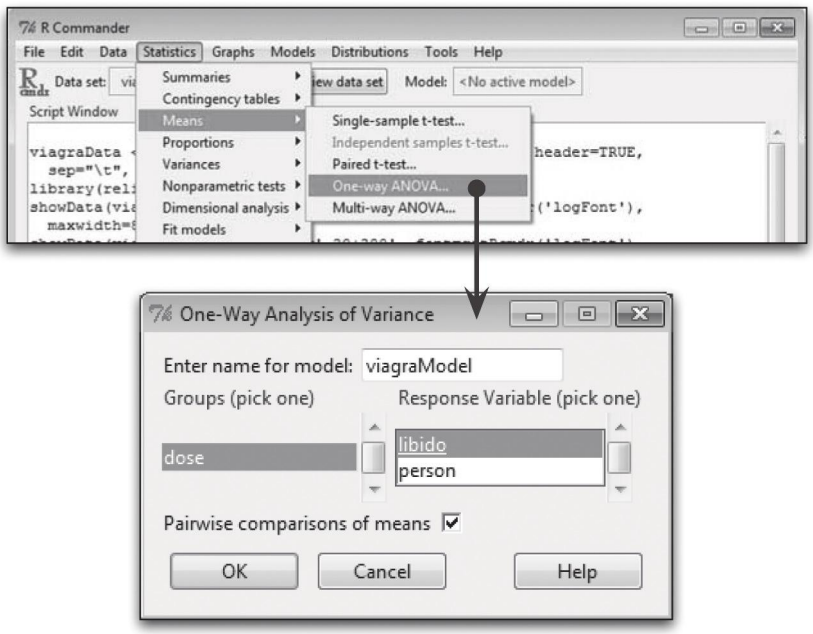


FIGURE 10.11 One-way ANOVA using R Commander

To do the ANOVA, use the **Statistics⇒Means⇒One-way ANOVA…** menu.[7] The resulting dialog box is fairly self-explanatory (Figure 10.11). You need to enter a name for the model that you're going to create (I have chosen *viagraModel*) in the box labelled *Enter name for model:*, select a factor from the list labelled *Groups* (in this case we have only one factor, **dose**) and select the outcome variable (in this case **libido**) from the list labelled

[7] If this menu isn't active it could be because you haven't converted **dose** into a factor. You need to have at least one factor in the dataframe for this menu to be active.

Response Variable. You cannot do planned comparisons using R Commander, but if you want a basic set of *post hoc* tests then select Pairwise comparisons of means ☑. Click on OK to run the analysis. The resulting output is described in sections 10.6.6.1 and 10.6.8.2.

10.6.5. Exploring the data ②

In Chapter 4 we saw that it is always a good idea to look at a graph of your data. In this case we will produce a line graph with error bars.

SELF-TEST

✓ Use *ggplot2* to produce a line chart with error bars showing bootstrapped confidence intervals for the Viagra data.

Figure 10.12 shows a line chart with error bars of the Viagra data. It's clear from this chart that all of the error bars overlap, indicating that, at face value, there are no between-group differences (although this measure is only approximate). The line that joins the means seems to indicate a linear trend in that, as the dose of Viagra increases, so does the mean level of libido.

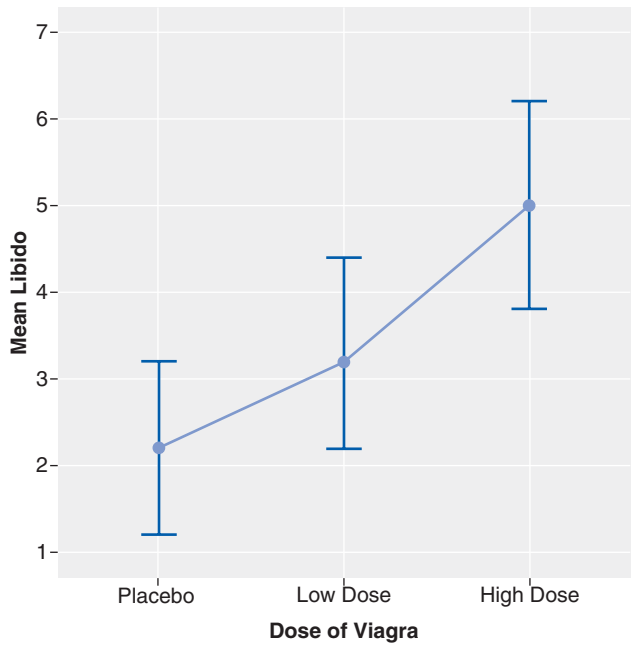


FIGURE 10.12 Error bar chart of the Viagra data (95% bootstrapped confidence intervals)

To get some descriptive statistics for each group we can use the *by()* function that we encountered in Chapter 5. Remember that this function takes the general form:

```
by(variable, group, output)
```

in which *variable* is the thing that you want to summarize (in this case **libido**), *group* is the variable that defines the groups by which you want to organize the output (in this case **dose**), and *output* is a function that tells **R** what output you would like to see (i.e., the mean). If we use the function **stat.desc()** from the package *pastecs* then **R** will output a host of useful descriptive statistics. Therefore, by combining *by()* and *stat.desc()*, we can get a table of descriptives for each group in a single line of code:

```
by(viagraData$libido, viagraData$dose, stat.desc)
```

Output 10.3 shows the resulting descriptive statistics (I have edited the output slightly to fit the page so you will see more decimal places and a few extra variables). Most of the variables are self-explanatory: we have the number of valid cases (*nbr.val*), minimum (*min*) and maximum (*max*) **libido**, the range, median, mean and variance (*var*), standard deviation (*std.dev*), standard error (*SE.mean*) and confidence interval (*CI.mean.0.95*).

```
viagraData$dose: Placebo
nbr.val  min  max  range  sum   median  mean  SE.mean
5.000    1.00 4.00  3.00  11.00 2.00    2.2000  0.5831

CI.mean.0.95          var      std.dev     coef.var
   1.6189318    1.7000000    1.3038405    0.5926548
------------------------------------------------------------------------
viagraData$dose: Low Dose

 nbr.val  min  max  range  sum   median  mean  SE.mean
5.000    2.00 5.00  3.00   16.00 3.00    3.200 0.5831

CI.mean.0.95          var      std.dev     coef.var
   1.6189318    1.7000000    1.3038405    0.4074502
------------------------------------------------------------------------
viagraData$dose: High Dose

nbr.val  min  max  range  sum   median  mean  SE.mean
5.00    3.00 7.00  4.00   25.00 5.0    5.0000  0.7071

CI.mean.0.95          var      std.dev     coef.var
   1.9632432    2.5000000    1.5811388    0.3162278
```

Output 10.3

The first thing to notice from Output 10.3 is that the means and standard deviations correspond to those shown in Table 10.1. In addition, we are told the standard error. You should remember that the standard error is the standard deviation of the sampling distribution of these data (so for the placebo group, if you took lots of samples from the population from which these data come, the means of these samples would have a standard deviation of 0.5831).

We are also given confidence intervals for the mean. By now, you should be familiar with what a confidence interval tells us, and that is that if we took 100 samples from the population from which the placebo group came and constructed confidence intervals for the mean, then 95 of these intervals would contain the true value of the mean. *CI.mean.0.95* doesn't give you the interval itself, but the value to add or subtract from the mean to create the interval. For example, in the placebo group the lower bound of the CI would be the mean minus *CI.mean.0.95* (i.e., 2.2000 – 1.6189 = 0.5811) and the upper bound of the CI would be the mean plus *CI.mean.0.95* (i.e., 2.2000 + 1.6189 = 3.8189). In other words, the true value of the mean is likely to be between 0.5811 and 3.8189. Although these diagnostics are not immediately important, we will refer back to them throughout the analysis.

The final thing before we get to the ANOVA itself is to compute Levene's test (see Chapter 5 and section 10.3.1). We encountered the *levene.Test()* function from the *car* package in Chapter 5, and we can again use it here. Just to remind you, it takes the general form:

```
leveneTest(outcome variable, group, center = median/mean)
```

So, if we want to do a Levene's test to see whether the variance in libido (the outcome) varies across groups that received different doses of the drug (**dose**), we can execute:

```
leveneTest(viagraData$libido, viagraData$dose, center = median)
```

The output (Output 10.4) shows that Levene's test is very non-significant, $F(2, 12) = 0.118$, $p = .89$. This means that for these data the variances are very similar (hence the high probability value); in fact, if you look at Output 10.3 you'll see that the variances of the placebo and low-dose groups are identical. Had this test been significant, we could instead conduct and report the results of Welch's F or a robust version of ANOVA, which we'll cover in the next section.

```
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  2  0.1176   0.89
      12
```

Output 10.4

10.6.6. The main analysis ②

10.6.6.1. When the test assumptions are met ②

There are two functions that can be used for ANOVA: *lm()*, which we used in Chapter 7, and **aov()**. As I explained earlier in the chapter, ANOVA is just a special case of the general linear model; therefore, we can use the linear model function, *lm()*, to run the analysis. For the current example, we are predicting **libido** from group membership (i.e., **dose** of Viagra) so our model is:

$$\text{libido}_i = \text{dose}_i + \text{error}_i$$

Therefore, we can create a model (which I've called *viagraModel*) using *lm()* by executing:

```
viagraModel<-lm(libido~dose, data = viagraData)
```

where *libido~dose* simply creates the model 'libido predicted from dose'.

The other function we can use is *aov()*, which stands for analysis of variance. Actually, *aov()* and *lm()* are exactly the same as each other. However, *aov()* takes the output from *lm()* and returns it to us in a way that is more in keeping with a traditional ANOVA approach. It's what is known as a 'wrapper': it is *lm()* but 'wrapped' up differently. I'm going to stick with the *aov()* function because it yields output that maps onto traditional ANOVA methods, but be clear that underneath we're actually using *lm()* to do the hard work.

The *aov()* function has the following general format:

```
newModel<-aov(outcome ~ predictor(s), data = dataFrame, na.action = an
action))
```

in which:

- *newModel* is an object created that contains information about the model. We can get summary statistics for this model by executing *summary(newModel)* for the main ANOVA summary and *summary.lm(newModel)* for specific parameters of the model.
- *outcome* is the variable that you're trying to predict, also known as the dependent variable. In this example it will be the variable **libido**.
- *predictor(s)* lists the variable or variables from which you're trying to predict the outcome variable, also known as the independent variable(s). In this example it will be the variable **dose**. In more complex designs we can specify several predictors or independent variables, but we'll come to that in subsequent chapters.
- *dataFrame* is the name of the dataframe from which your outcome and predictor variables come.
- *na.action* is an optional command. If you have complete data (as we have here) you can ignore it, but if you have missing values (i.e., NAs in the dataframe) then it can be useful to use *na.action* = *na.exclude*, which will exclude all cases with missing values – see R's Souls' Tip 7.1).

For the current example, then, we could execute the following command:

```
viagraModel<-aov(libido ~ dose, data = viagraData)
```

to generate the model (note that the command is basically identical to when we used *lm()* to run an ANOVA above). We now have an object called *viagraModel* that contains information about how well **dose** predicts **libido**. To see the summary statistics execute:

```
summary(viagraModel)
```

Executing this command generates Output 10.5. The output is divided into effects due to the model (the experimental effect) and residuals (this is the unsystematic variation in the data). The effect labelled *dose* is the overall experimental effect. In this row we are told the sums of squares for the model (SS_M = 20.13) and this value corresponds to the value calculated in section 10.2.6. The degrees of freedom are equal to 2 and the mean squares value for the model corresponds to that calculated in section 10.2.8 (10.067). The sum of squares and mean squares represent the experimental effect. The row labelled *Residuals* gives details of the unsystematic variation within the data (the variation due to natural individual differences in libido and different reactions to Viagra). The table tells us how much unsystematic variation exists (the residual sum of squares, SS_R) and this value (23.60) corresponds to the value calculated in section 10.2.7. The table then gives the average amount of unsystematic variation, the mean squares (MS_R), which corresponds to the value (1.967) calculated in section 10.2.8. The test of whether the group means are the same is represented by the *F*-ratio for the effect of **dose**. The value of this ratio is 5.12, which is the same as was calculated in section 10.2.9. Finally, **R** tells us whether this value is likely to have happened by chance. The final column labelled *Pr(>F)* indicates the likelihood of an *F*-ratio the size of the one obtained occurring if there was no effect in the population (see also R's Souls' Tip 10.1). In this case, there is a probability of .025 that an *F*-ratio of this size would occur if in reality there was no effect (that's only a 2.5% chance!). We have seen in previous chapters that we use a cut-off point of .05 as a criterion for statistical significance. Hence, because the observed significance value is less than .05 we can say that there was a significant effect of Viagra. However, at this stage we still do not know exactly what the effect of Viagra was (we don't know which groups differed). One thing that is

interesting here is that we obtained a significant experimental effect, yet our error bar plot indicated that no significant difference would be found. This contradiction illustrates how the error bar chart can act only as a rough guide to the data.

```
            Df Sum Sq Mean Sq F value  Pr(>F)
dose         2 20.133 10.0667  5.1186 0.02469 *
Residuals   12 23.600  1.9667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output 10.5

R's Souls' Tip 10.1 One- and two-tailed tests in ANOVA ②

A question I get asked a lot by students is 'is the significance of the ANOVA one- or two-tailed, and if it's two-tailed can I divide by 2 to get the one-tailed value?' The answer is that to do a one-tailed test you have to be making a directional hypothesis (i.e., the mean for cats is greater than for dogs). ANOVA is a non-specific test, so it just tells us generally whether there is a difference or not, and because there are several means you can't possibly make a directional hypothesis. As such, it's invalid to halve the significance value.

The *aov()* function also automatically generates some plots that we can use to test the assumptions. We can see these graphs by executing:

```
plot(viagraModel)
```

The results are in Figure 10.13. You will actually see four graphs, but the first two are the most important for ANOVA. The first graph (on the left of the figure) can be used for testing homogeneity of variance. We encountered this kind of plot in Chapter 7: essentially, if it has a funnel shape then we're in trouble. The plot we have shows points that are equally spread for the three groups, which implies that variances are similar across groups (which was also the conclusion reached by Levene's test). The second plot (on the right) is a Q-Q plot (see Chapter 5), which tells us something about the normality of residuals in the model. We want our residuals to be normally distributed, which means that the dots on the graph should cling lovingly to the diagonal line. Ours look like they have had a bit of an argument with the diagonal line, which suggests that we may not be able to assume normality of errors and should perhaps use a robust version of ANOVA instead (which will be explained sooner than you might like).

10.6.6.2. When variances are not equal across groups ②

If Levene's test is significant then it is reasonable to assume that population variances are different across groups.[8] In this case, if our distributions are as they should be, we can apply

[8] It's worth reminding you that any significance test depends on sample size: in small samples there won't be power to detect differences across groups, and in large samples even small differences in variances might be deemed significant. As such, don't place too much weight on Levene's test if it's non-significant in a small sample, or significant in a large sample.

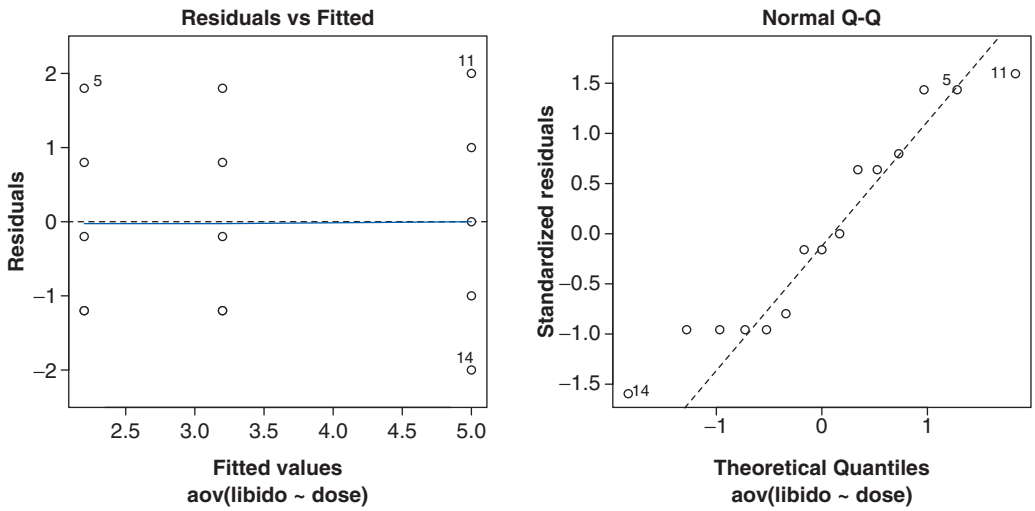


FIGURE 10.13
Plots of an ANOVA model

Welch's *F* to the data, which makes adjustments for differences in group variances. This test is produced by the *oneway.test()* function, which is built into **R**. The format of this test is the same as *aov()*:

```
oneway.test(outcome ~ predictor, data = dataframe)
```

Therefore, we can get the output for Welch's *F* for the current data by executing:

```
oneway.test(libido ~ dose, data = viagraData)
```

Output 10.6 shows Welch's *F*-ratio. For our data we didn't need this test because our Levene's test was not significant, indicating that our population variances were similar. However, when homogeneity of variance has been violated you should look at this *F*-ratio *instead* of the ones in the previous section. If you're interested in how these values are calculated then look at Oliver Twisted, but to be honest it's not that much fun and you'd probably enjoy yourself more if you spent the time sticking jellyfish down your pants. You're much better off just trusting that **R** has done what it was supposed to do. Note that the error degrees of freedom have been adjusted – you should remember this when you report the values. For these data, Welch's $F(2, 7.94) = 4.23$, $p = .054$, which is just about non-significant. If we were using this test it would imply that the mean libido did not differ significantly across different doses of Viagra.

```
        One-way analysis of means (not assuming equal variances)

data:  libido and dose
F = 4.3205, num df = 2.000, denom df = 7.943, p-value = 0.05374
```

Output 10.6

10.6.6.3. Robust ANOVA – it's not for the weak of heart ③

Wilcox (2005) describes a set of robust procedures for conducting one-way ANOVA. Load these functions using the instructions in section 5.8.4. Having done this, we now have access to Wilcox's functions. The first issue with using these functions is that most of them require the data to be in wide format rather than the long format that we have been using

so far in this chapter. We can convert the data to wide format using the *unstack()* command (see section 3.9.4), which has the general form:

```
newDataFrame<-unstack(oldDataFrame, scores ~ columns)
```

In this case our scores are stored in the variable **libido** and we want to make different columns for each group, so our columns variable is **dose**. Therefore, we can reformat the data by executing:

```
viagraWide<-unstack(viagraData, libido ~ dose)
```

This command creates a new dataframe called *viagraWide*, which is our Viagra data but in wide format, so each column represents a different group:

```
  Placebo Low.Dose High.Dose
1       3        5         7
2       2        2         4
3       1        4         5
4       1        2         3
5       4        3         6
```

This is the format that Wilcox's functions expect. The first robust function, **t1way()**, is based on a trimmed mean. It takes the general form:

```
t1way(dataFrame, tr = .2, grp = c(x, y, …, z))
```

in which,

- *dataFrame* is the name of the dataframe to be analysed.
- *tr* is the proportion of trimming to be done. The default is .2 or 20%, and you need to use this option only if you want to specify an amount other than 20%.
- *grp* can be used to specify particular groups by referring to their column in the dataframe; for example, if we wanted to analyse only the placebo and high-dose group, we could do this using *grp = c(1,3)*.

As such, for an ANOVA of the Viagra data based on 20% trimmed means we simply execute:

```
t1way(viagraWide)
```

If we wanted to trim only 10% of the data then we could execute:

```
t1way(viagraWide, tr = .1)
```

If you execute this command you will see Output 10.7, which shows that, based on this robust test, there is not a significant difference in libido scores across the three dose groups, $F_t(2, 7.94) = 4.32$, $p = .054$.

We can also compare medians rather than means using **med1way()**, which takes the general form:

```
med1way(dataFrame, grp = c(x, y, …, z))
```

in which, *dataFrame* is the name of the dataframe to be analysed and *grp* is used in the same way as in *t1way()*. As such, for an ANOVA of the Viagra data based on medians we simply execute:

```
med1way(viagraWide)
```

If you execute this command you will see Output 10.7, which shows that, based on this robust test, there is not a significant difference in median libido scores across the three dose groups, $F_m = 4.78$, $p = .07$.

A final method is to add a bootstrap to the trimmed mean method using **t1waybt()**. This function has the general form:

```
t1waybt(dataFrame, tr = .2, alpha = .05, grp = c(x, y, …, z), nboot = 599)
```

which is the same as *t1way()* except that we have two additional options. The first is *alpha*, which sets the Type I error rate. The default is .05, which is fairly standard, so unless you want something different you don't need to use this option. The second is *nboot*, which specifies the number of bootstrap samples to be used. The default is 599, which, if anything, you might want to increase (but it's probably not necessary to use more than 2000). As such, for an ANOVA of the Viagra data based on 20% trimmed means, with 599 bootstrap samples, we execute:

```
t1waybt(viagraWide)
```

However, if we wanted, for example, a 5% trimmed mean with 2000 bootstrap samples we would execute:

```
t1waybt(viagraWide, tr = .05, nboot = 2000)
```

If you execute the *t1waybt()* function with the default settings you will see Output 10.7, which shows that, based on this robust test, there is not a significant difference in trimmed mean libido scores across the three dose groups, $F_t = 3$, $p = .089$. In short, all three robust methods suggest that dose does not have a significant impact on libido.

t1way() output	*med1way()* output	*t1waybt()* output

`$TEST`
`[1] 4.320451`
 	
`$nu1`
`[1] 2`
 	
`$nu2`
`[1] 7.943375`
 	
`$siglevel`
`[1] 0.05373847`
	`$TEST`
	`[1] 4.782879`
	
	`$crit.val`
	`[1] 5.472958`
	
	`$p.value`
	`[1] 0.07`
		`$test`
		`[1] 3`
		
		`$p.value`
		`[1] 0.0886076`

Output 10.7

10.6.7. Planned contrasts using R ②

To do planned comparisons in **R** we have to set the contrast attribute of our grouping variable using the *contrast()* function and then re-create our ANOVA model using *aov()*. By default, dummy coding is used, which was explained in section 10.2.3. We can see this if we summarize our existing *viagraModel* using the *summary.lm()* function rather than *summary()*. By using *summary.lm()* we are asking for a summary of the parameters of the linear model (rather than the overall ANOVA). Assuming you still have the *viagraModel* object (if not, re-create it) execute this command:

```
summary.lm(viagraModel)
```

You should get Output 10.8. Note that this is basically the same as Output 10.1, which we used to explain how dummy coding works. So, the 'low dose' effect is the effect of low dose compared to placebo and is non-significant ($t = 1.13$, $p = .282$), whereas the effect of high dose compared to the placebo group is significant ($t = 3.16$, $p = .008$).

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.2000     0.6272   3.508  0.00432 **
doseLow Dose    1.0000     0.8869   1.127  0.28158
doseHigh Dose   2.8000     0.8869   3.157  0.00827 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.402 on 12 degrees of freedom
Multiple R-squared: 0.4604,       Adjusted R-squared: 0.3704
F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Output 10.8

This is all very well, but what if we do not want dummy coding, but want to use our own planned comparisons, use another built-in comparison, or do a trend analysis? In general, we do this by resetting the contrast attribute associated with our predictor variable (in this case **dose**), using the following general command:

```
contrasts(predictor variable)<-contrast instructions
```

The *contrast instructions* can be either a set of weights for the contrasts that you want to do, or one of the built-in contrasts listed in Table 10.6. These built in functions can be:

```
contr.helmert(n)
contr.poly(n)
contr.treatment(n, base = x)
contr.SAS(n)
```

In all cases, *n* is the number of groups in the predictor variable (for **dose**, this value will be 3). The *contr.treatment()* function has an additional option, *base*, which allows you to specify the group that you want to use as a baseline. Therefore, if you want dummy coding (i.e., the first category is the baseline) you would use *contr.treatment(n, base = 1)*. The function *contr.SAS()* is the same as using *contr.treatment()* when you select the last category as the baseline.

To put this all together, if we wanted to set the contrast property of **dose** to be a Helmert contrast then we would execute:

```
contrasts(viagraData$dose)<-contr.helmert(3)
```

Note that the *3* is the number of groups present in the **dose** variable. We're not going to use this contrast, though, we're going to specify our own.

10.6.7.1. Your own contrasts ②

To conduct the planned comparisons described in section 10.4, we follow the general procedure just described. We need to tell **R** what weights to assign to each group. The first step is to decide which comparisons you want to do and then what weights must be assigned to each group for each of the contrasts. We have already gone through this process in section 10.4.2, so we know that the weights for contrast 1 were −2 (placebo group), +1 (low-dose group) and +1 (high-dose group). If we wanted to express these weights we could create a new object called *contrast1* and use the function *c()* to list the weights:

```
contrast1<-c(-2,1,1)
```

This variable indicates that the first group has a weight of –2, and the second and third groups a weight of 1. The order of the numbers is important because it corresponds to the order of groups in your predictor variable. In the Viagra data, remember that the order of groups was: placebo (because it was coded with the lowest value, 1), low dose (because it was coded using the next lowest number, 2), and high dose (because it was coded with the highest number, 3). As such, *contrast1* has the weights for placebo, low dose and high dose, in that order.

We can do the same for the second contrast. We know from section 10.4.2 that the weights for contrast 2 were: 0 (placebo group), −1 (low-dose group) and +1 (high-dose group). Remembering that the first weight we enter will be for the placebo group, we must enter the value 0 as the first weight, then –1 for the low-dose group and finally 1 for the high-dose group. It is imperative that you remember to input zero weights for any groups that are not in the contrast. We can specify this contrast by executing:

```
contrast2<-c(0,-1,1)
```

which creates a variable called *contrast2* that contains the weights for the second contrast.

Having created these variables we now need to bind them together using **cbind()**, which literally binds two columns of data together, and set them as the contrast attached to our predictor variable, **dose**. We can do this by executing:

```
contrasts(viagraData$dose)<-cbind(contrast1, contrast2)
```

This command sets the contrast property of **dose** to contain the weights for the two contrasts that we want to conduct.[9] If you have a look at the **dose** variable by executing:

```
viagraData$dose
```

You'll see this:

```
[1] Placebo   Placebo   Placebo   Placebo   Placebo   Low Dose  Low
Dose  Low Dose  Low Dose  Low Dose  High Dose High Dose High Dose
[14] High Dose High Dose
attr(,"contrasts")
          contrast1 contrast2
Placebo          -2         0
Low Dose          1        -1
High Dose         1         1
Levels: Placebo Low Dose High Dose
```

Note that the variable now has a contrast attribute that contains the weights that we just specified. This is very useful to look at to check that you have entered the weights correctly. Remember that when we do planned comparisons we arrange the weights such that we compare any group with a positive weight against any group with a negative weight. Therefore, the table of weights shows that contrast 1 compares the placebo group against the two experimental groups, and contrast 2 compares the low-dose group to the high-dose group. These are the contrasts we wanted. Happy days.

Once we have set the contrast attribute we create a new model using *aov()*, in exactly the same way as we did before, by executing:

```
viagraPlanned<-aov(libido ~ dose, data = viagraData)
```

If you use the *summary()* command you'll see that the model is the same as the *viagraModel* that we created earlier. However, to access the contrasts we need the model parameters, which are obtained by executing:

```
summary.lm(viagraPlanned)
```

[9] I think that creating the *contrast1* and *contrast2* variables makes what we're doing a bit easier to understand, but in reality I would normally create these contrasts by executing this single command:

```
contrasts(viagraData$dose)<-cbind(c(-2,1,1), c(0,-1,1))
```

The resulting Output 10.9 is the same as Output 10.2, which we looked at earlier when explaining how these contrasts work. Re-read that earlier material to see from where the values of the parameters come. The table gives the standard error of each contrast and a *t*-statistic. The significance value of the contrast is given in the final column, and this value is two-tailed. Using the first contrast as an example, if we had used this contrast to test the general hypothesis that the experimental groups would differ from the placebo group, then we should use this two-tailed value. However, in reality we tested the hypothesis that the experimental groups would increase libido above the levels seen in the placebo group: this hypothesis is one-tailed. Provided the means for the groups bear out the hypothesis we can divide the significance values by 2 to obtain the one-tailed probability (i.e., .0293/2 = .0147). Hence, for contrast 1, we can say that taking Viagra significantly increased libido compared to the control group (p = .0147). For contrast 2 we also had a one-tailed hypothesis (that a high dose of Viagra would increase libido significantly more than a low dose) and the means bear this hypothesis out. The significance of contrast 2 tells us that a high dose of Viagra increased libido significantly more than a low dose (p(one-tailed) = .0652/2 = .0326). Notice that had we not had a specific hypothesis regarding which group would have the highest mean, then we would have had to conclude that the dose of Viagra had no significant effect on libido. For this reason it can be important as scientists that we generate hypotheses before collecting any data, because this method of scientific discovery is more powerful.

In summary, the planned contrasts revealed that taking Viagra significantly increased libido compared to a control group, $t(12) = 2.47$, $p < .05$, and taking a high dose significantly increased libido compared to a low dose, $t(12) = 2.03$, $p < .05$ (one-tailed).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4667     0.3621   9.574 5.72e-07 ***
dose1         0.6333     0.2560   2.474   0.0293 *
dose2         0.9000     0.4435   2.029   0.0652 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.402 on 12 degrees of freedom
Multiple R-squared: 0.4604,      Adjusted R-squared: 0.3704
F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Output 10.9

10.6.7.2. Trend analysis ②

To conduct a trend analysis we can use *contr.poly()*. It is important that we have coded the predictor variable groups in a meaningful order. We expect libido to be smallest in the placebo group, to increase in the low-dose group and then to increase again in the high-dose group. To detect a meaningful trend, we need to have coded these groups in ascending order. We have done this by coding the placebo group with the lowest value 1, the low-dose group with the middle value 2 and the high-dose group with the highest coding value of 3. If we coded the groups differently, this would influence both whether a trend is detected and, if a trend is detected, whether it is statistically meaningful.

To obtain a trend analysis we follow the general procedure of setting the contrast attribute of the predictor variable, which in this case we can do by executing:

```
contrasts(viagraData$dose)<-contr.poly(3)
```

The '3' just tells *contr.poly()* how many groups there are in the predictor variable. Having set the contrast we again create a new model using *aov()*, by executing:

```
viagraTrend<-aov(libido ~ dose, data = viagraData)
```

To access the contrasts we need the model parameters, which are obtained by executing:

```
summary.lm(viagraTrend)
```

The resulting Output 10.10 breaks down the experimental effect to see whether it can be explained by either a linear (*dose.L*) or a quadratic (*dose.Q*) relationship in the data. First, let's look at the linear component. This comparison tests whether the means increase across groups in a linear way. The most important things to note are the value of the t and the corresponding significance value. For the linear trend $t = 3.16$ and this value is significant at $p = .008$. Therefore, we can say that as the dose of Viagra increased from nothing to a low dose to a high dose, libido increased proportionately.

Moving onto the quadratic trend, this comparison is testing whether the pattern of means is curvilinear (i.e., is represented by a curve that has one bend). The error bar graph of the data suggests that the means cannot be represented by a curve and the results for the quadratic trend bear this out: $t = 0.52$ and this value is significant at $p = .612$, which is not very significant at all.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4667     0.3621   9.574 5.72e-07 ***
dose.L        1.9799     0.6272   3.157  0.00827 **
dose.Q        0.3266     0.6272   0.521  0.61201
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.402 on 12 degrees of freedom
Multiple R-squared: 0.4604,      Adjusted R-squared: 0.3704
F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Output 10.10

10.6.8. *Post hoc* tests using R ②

How you conduct *post hoc* tests in **R** depends on which test you'd like to do. Bonferroni and related methods (such as Holm and Benjamini–Hochberg) are done using the *pairwise.t.test()* function, which is part of the **R** base system. However, Tukey and Dunnett's test (and some others that we're not going to look at) can be done using the *glht()* function in the *multcomp()* package. Finally, Wilcox (2005) has some robust methods implemented in his functions *lincon()* and *mcpp20()*. This section is divided according to these different methods.

10.6.8.1. Bonferroni and related methods ②

Bonferroni and related methods (e.g., Holm, Benjamini–Hochberg, Hommel, Hochberg) can be implemented using the *pairwise.t.test()* function that is built into **R**. This function takes the general form:

```
pairwise.t.test(outcome, predictor, paired = FALSE, p.adjust.method =
"method")
```

in which:

- *outcome* is the name of your outcome variable (in this case it will be **libido** (*viagraData$libido*).

- *predictor* is the name of your grouping variable (in this case it will be **dose** (*viagraData$dose*).
- *paired* is a logical statement that by default is FALSE but can be set to TRUE (the capital letters matter). This specifies whether you want paired *t*-tests or not. For these data we have independent groups so we do not want paired *t*-tests and the default of FALSE is fine, but we'll revisit this option in Chapter 13.
- *p.adjust.method* is a string that specifies which correction you would like to apply to your *p*-values. You can replace *"method"* in the command above with *"bonferroni"*, *"holm"*, *"hochberg"*, *"hommel"*, *"BH"* (which produces the Benjamini–Hochberg method), *"BY"* (which produces the more recent Benjamini–Yekutieli method), *"fdr"* (the general false discovery rate method), and *"none"* (you don't correct the *p*-value at all, you just do lots of *t*-tests – not advisable).

As such, we can obtain Bonferroni and Benjamini–Hochberg *post hoc* tests for the current data by executing these two commands:

```
pairwise.t.test(viagraData$libido, viagraData$dose, p.adjust.method =
"bonferroni")

pairwise.t.test(viagraData$libido, viagraData$dose, p.adjust.method = "BH")
```

Both commands specify **libido** as the outcome variable, and **dose** as the grouping variable, but they differ in the method that is set for correcting the *p*-values. The results can be seen in Output 10.11. Both methods produce a grid of *p*-values for all combinations of the groups. First of all, let's look at the Bonferroni corrected values: the placebo group is compared to the low-dose group and reveals a non-significant difference (.845 is greater than .05), but when compared to the high-dose group there is a significant difference (.025 is less than .05).

SELF-TEST

✓ Our planned comparison showed that any dose of Viagra produced a significant increase in libido, yet the *post hoc* tests indicate that a low dose does not. Why is there this contradiction?

In section 10.4.2, I explained that the first planned comparison would compare the experimental groups to the placebo group. Specifically, it would compare the average of the two group means of the experimental groups ((3.2 + 5.0)/2 = 4.1) to the mean of the placebo group (2.2). So, it was assessing whether the difference between these values (4.1 − 2.2 = 1.9) was significant. In the *post hoc* tests, when the low dose is compared to the placebo, the contrast is testing whether the difference between the means of these two groups is significant. The difference in this case is only 1, compared to a difference of 1.9 for the planned comparison. This explanation illustrates how it is possible to have apparently contradictory results from planned contrasts and *post hoc* comparisons. More important, it illustrates how careful we must be in interpreting planned contrasts.

The final comparison is the low-dose group compared to the high-dose group, which is not significant (because 0.196 is greater than .05). This result contradicts the planned

comparisons (remember that contrast 2 compared these groups and found a significant difference).

SELF-TEST

✓ Why does the *post hoc* test show a non-significant difference between high and low dose, when the planned comparison showed a significant difference?

This contradiction occurs for two possible reasons. First, *post hoc* tests by their nature are two-tailed (you use them when you have made no specific hypotheses and you cannot predict the direction of hypotheses that don't exist!) and contrast 2 was significant only when considered as a one-tailed hypothesis. However, even at the two-tailed level the planned comparison was closer to significance than the *post hoc* test and this fact illustrates that *post hoc* procedures are more conservative (i.e., have less power to detect true effects) than planned comparisons.

Looking now at the BH corrected tests, we find the same pattern of results as for Bonferroni: placebo is significantly different from a high dose (because .025 is less than .05), but not a low dose (.282 is greater than .05) and low and high doses did not significantly differ (.098 is greater than .05).

Bonferroni

```
Pairwise comparisons using t
tests with pooled SD

data:  viagraData$libido and
viagraData$dose

          Placebo Low Dose
Low Dose  0.845   -
High Dose 0.025   0.196

P value adjustment method:
bonferroni
```

BH

```
Pairwise comparisons using t
tests with pooled SD

data:  viagraData$libido and
viagraData$dose

          Placebo Low Dose
Low Dose  0.282   -
High Dose 0.025   0.098

P value adjustment method: BH
```

Output 10.11

10.6.8.2. Tukey and Dunnett ②

Tukey and Dunnett can be implemented using the *glht()* function that is part of the *multcomp* package (so remember to install and load it). This function takes the general form:

```
newModel<-glht(aov.Model, linfct = mcp(predictor = "method"), base = x)
```

in which:

- *newModel* is an object containing the information from the *post hoc* tests. To see this information we can use *summary(newModel)* for the basic *post hoc* tests and *confint(newModel)* to see the confidence intervals.

- *aov.Model* is the name of a model that has already been created with the *aov()* function (in this case it will be *viagraModel*).
- *predictor* is the name of your grouping variable (in this case it will be **dose** (*viagraData$dose*).
- *linfct = mcp(predictor = "method")* specifies which correction you would like to apply to your *p*-values. You can replace "*method*" in the command above with "*Dunnett*", "*Tukey*", "*Sequen*", "*AVE*", "*Changepoint*", "*Williams*", "*Marcus*", "*McDermott*", "*UmbrellaWilliams*", and "*GrandMean*".
- *base* is used only when "*Dunnett*" is specified. This option allows you to specify the baseline group using a group number. In this case if we wanted the placebo as the baseline we would use *base = 1*, but if we wanted the high-dose group we could specify *base = 3*.

For the Viagra data, we can obtain Tukey *post hoc* tests by executing:

```
postHocs<-glht(viagraModel, linfct = mcp(dose = "Tukey"))
summary(postHocs)
confint(postHocs)
```

The first command creates an object (which I've called *postHocs*) that is based on the *viagraModel* that we created in section 10.6.6.1. The *linfct* command is set to perform Tukey tests on the variable dose (the reason why we can type 'dose' rather than 'viagraData$dose' is because the function will look for 'dose' within *viagraModel*, which has been specified within the function). To access the information within *postHocs* we execute *summary()* to get the *post hoc* tests (Output 10.12) and *confint()* to get the corresponding confidence intervals (Output 10.13).

Output 10.12 shows the three comparisons (low dose vs. placebo, high dose vs. placebo, high dose vs. low dose), the estimate (which is the difference between the group means), the standard error associated with the difference between means, the *t*-test (which is simply the difference between means divided by the standard error, so for the first contrast it is 1/0.8869 = 1.127), and its associated *p*-value. As with the tests in the previous section, this output confirms significant differences between the high dose and placebo groups, $t = 3.16$, $p < .05$, but not between the low-dose group and the placebo, $t = 1.13$, $p = .52$, and high dose, $t = 2.03$, $p = .15$, groups. The confidence intervals (Output 10.13) also confirm this because they do not cross zero for the comparison of the high dose and placebo group, which means that the true difference between group means is likely not to be zero (i.e., no difference); conversely, for the other contrasts the confidence intervals cross zero, implying that the true difference between means could be zero.

```
Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = libido ~ dose, data = viagraData)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Low Dose - Placebo == 0   1.0000     0.8869   1.127   0.5162
High Dose - Placebo == 0  2.8000     0.8869   3.157   0.0208 *
High Dose - Low Dose == 0 1.8000     0.8869   2.029   0.1474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Output 10.12

```
Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = libido ~ dose, data = viagraData)

Quantile = 2.6671
95% family-wise confidence level

Linear Hypotheses:
                          Estimate lwr     upr
Low Dose - Placebo == 0    1.0000  -1.3656  3.3656
High Dose - Placebo == 0   2.8000   0.4344  5.1656
High Dose - Low Dose == 0  1.8000  -0.5656  4.1656
```

Output 10.13

We can obtain Dunnett *post hoc* tests for the Viagra data by executing:

```
postHocs<-glht(viagraModel, linfct = mcp(dose = "Dunnett"), base = 1)
summary(postHocs)
confint(postHocs)
```

The first command is the same as before, except that we have replaced "Tukey" with "Dunnett". We have also added the base command (because we're using Dunnett) to specify which group to use as the control group. We have used *base = 1*, which means 'use the first group', which in this case is the placebo group. To access the information we again execute *summary()* and *confint()*. The results are in Output 10.14. I won't labour the point because the conclusions are the same as for Tukey; all I will say is that you should note that Dunnett's test compares groups to a baseline so we end up with two tests rather than three. In this case we asked every group to be compared to the placebo group, so there is no comparison of the high and low-dose groups.

```
Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = libido ~ dose, data = viagraData)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
Low Dose - Placebo == 0    1.0000     0.8869   1.127   0.4459
High Dose - Placebo == 0   2.8000     0.8869   3.157   0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = libido ~ dose, data = viagraData)

Quantile = 2.5023
95% family-wise confidence level

Linear Hypotheses:
                          Estimate lwr     upr
Low Dose - Placebo == 0    1.0000  -1.2194  3.2194
High Dose - Placebo == 0   2.8000   0.5806  5.0194
```

Output 10.14

10.6.8.3. Run for cover – it's robust *post hoc* tests ③

As with the robust ANOVA, to run robust *post hoc* tests we need to (1) source Rand Wilcox's functions (see section 10.6.6.3 for how to do this); and (2) input data in the wide format – therefore, we'll use the object *viagraWide* that we created in section 10.6.6.3. We are going to use two functions: *lincon()*, which is based on trimmed means; and *mcppb20()*, which uses a percentile bootstrap to compute *p*-values as well as trimming the group means. The latter method, in particular, seems good at controlling the Type I error rate. The general forms of these functions are similar to *t1way()* and *t1waybt()*, which we encountered earlier in the chapter:[10]

```
lincon(dataframe, tr = .2, grp = c(x, y, …, z))
mcppb20(dataframe, tr = .2, nboot = 2000, grp = c(x, y, …, z))
```

The options for each function are the same as described in section 10.6.6.3. Note that these functions take the same parameters, except that *mcppb20()* has an additional *nboot* command to control the number of bootstrap samples (the default is 2000, which is fine). Trimming on the means defaults to 20% (*tr* = .2). If you are happy with the default values then we can execute these commands on the *viagraWide* dataframe as follows:

```
lincon(viagraWide)
mcppb20(viagraWide)
```

It's as easy as that. Output 10.15 comes from *lincon()*. Note that the confidence intervals are corrected for the number of tests, but the *p*-values are not. As such, we should ascertain significance from whether or not the confidence intervals cross zero. In this case they all do, which implies that none of the groups are significantly different. This is different from what we found when we did not trim the means (see the previous two sections).

SELF-TEST

✓ Repeat the analysis with 10% trimmed means. How do your conclusions differ?

```
[1] "Note: confidence intervals are adjusted to control FWE"
[1] "But p-values are not adjusted to control FWE"
$test
     Group Group      test crit       se df
[1,]     1     2 0.8660254 3.74 1.154701  4
[2,]     1     3 2.5980762 3.74 1.154701  4
[3,]     2     3 1.7320508 3.74 1.154701  4
$psihat
     Group Group psihat ci.lower ci.upper    p.value
[1,]     1     2     -1 -5.31858  3.31858 0.43533094
[2,]     1     3     -3 -7.31858  1.31858 0.06016985
[3,]     2     3     -2 -6.31858  2.31858 0.15830242
```

Output 10.15

[10] They actually have a few extra options, but I'm keeping things simple.

Output 10.16 comes from *mcpp20()*. Unlike *lincon()*, both the confidence intervals and *p*-values are corrected for the number of tests. The main table lists three contrasts. To make sense of these we have to look at the contrast codes listed under *$con*. These are like the contrast weights that we looked at earlier in the chapter, so groups with positive weights are compared to those with negative weights. From the contrast codes we can see that contrast 1 compares groups 1 and 2 (i.e., placebo vs. low dose), contrast 2 compares groups 1 and 3 (i.e., placebo vs. high dose), and contrast 3 compares groups 2 and 3 (i.e., low dose vs. high dose).

Looking at the confidence intervals, it's clear that only the interval for contrast 2 does not cross zero, implying a significance difference between the high dose and placebo group (which is confirmed by the associated *p*-value, which is smaller than .05). For the other two comparisons the confidence intervals cross zero (and the *p*s are greater than .05), implying non-significant differences in libido between the low-dose group and both placebo (contrast 1) and high-dose (contrast 3) groups. Essentially, this profile of results is consistent with what we found using non-robust *post hoc* tests.

```
[1] "Taking bootstrap samples. Please wait."
$psihat
     con.num psihat       se  ci.lower   ci.upper p-value
[1,]       1     -1 1.154701 -3.333333  1.3333333  0.3250
[2,]       2     -3 1.154701 -5.333333 -0.3333333  0.0055
[3,]       3     -2 1.154701 -4.333333  0.6666667  0.0840

$crit.p.value
[1] 0.017

$con
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]   -1    0    1
[3,]    0   -1   -1
```

Output 10.16

CRAMMING SAM'S TIPS One-way ANOVA

- The one-way independent ANOVA compares several means, when those means have come from different groups of people; for example, if you have several experimental conditions and have used different participants in each condition.
- When you have generated specific hypotheses before the experiment use *planned comparisons*, but if you don't have specific hypotheses use *post hoc* tests.
- There are lots of different *post hoc* tests: when you have equal sample sizes and homogeneity of variance is met, use Tukey's HSD. If there is any doubt about the underlying assumptions then use a robust method.
- Test for homogeneity of variance using Levene's test. Find the table with this label: if the *p*-value is less than .05 then the assumption is violated. If homogeneity of variance has been met (the significance of Levene's test is greater than .05), run a normal ANOVA. If, however, the assumption is violated (the significance of Levene's test is less than .05) compute Welch's *F* instead of the normal ANOVA, or use a robust method based on trimmed means and/or a bootstrap.
- In the main ANOVA, if the value of *p* is less than .05 then the means of the groups are significantly different.
- For contrasts and *post hoc* tests, look at the confidence intervals and *p*-values to discover if your comparisons are significant. If the confidence intervals do not contain zero or the *p*-value is less than .05 then the effect is significant.

Labcoat Leni's Real Research 10.1 Scraping the barrel? ①

Gallup, G. G. J., et al. (2003). *Evolution and Human Behavior*, *24*, 277–289.

Evolution has endowed us with many beautiful things (cats, dolphins, the Great Barrier Reef, etc.), all selected to fit their ecological niche. Given evolution's seemingly limitless capacity to produce beauty, it's something of a wonder how it managed to produce such a monstrosity as the human penis. One theory is that the penis evolved into the shape that it is because of sperm competition. Specifically, the human penis has an unusually large glans (the 'bell end', as it's affectionately known) compared to other primates, and this may have evolved so that the penis can displace seminal fluid from other males by 'scooping it out' during intercourse. To put this idea to the test, Gordon Gallup and his colleagues came up with an ingenious study (Gallup et al., 2003). Armed with various female masturbatory devices from Hollywood Exotic Novelties, an artificial vagina from California Exotic Novelties, and some water and cornstarch to make fake sperm, they loaded the artificial vagina with 2.6 ml of fake sperm and inserted one of three female sex toys into it before withdrawing it. Over several trials, three different female sex toys were used: a control phallus that had no coronal ridge (i.e., no bell end), a phallus with a minimal coronal ridge (small bell end) and a phallus with a coronal ridge.

They measured sperm displacement as a percentage using the following equation (included here because it is more interesting than all of the other equations in this book):

$$\frac{\text{weight of vagina with semen} - \text{weight of vagina following insertion and removal of phallus}}{\text{weight of vagina with semen} - \text{weight of empty vagina}} \times 100$$

As such, 100% means that all of the sperm was displaced by the phallus, and 0% means that none of the sperm was displaced. If the human penis evolved as a sperm displacement device, then Gallup et al. predicted: (1) that having a bell end would displace more sperm than not; and (2) the phallus with the larger coronal ridge would displace more sperm than the phallus with the minimal coronal ridge. The conditions are ordered (no ridge, minimal ridge, normal ridge) so we might also predict a linear trend. The data can be found in the file **Gallup et al.csv**. Draw an error bar graph of the means of the three conditions. Conduct a one-way ANOVA with planned comparisons to test the two hypotheses above. What did Gallup et al. find?

Answers are in the additional material on the companion website (or look at pages 280–281 in the original article).

10.7. Calculating the effect size ②

One thing you will notice is that **R** doesn't routinely provide an effect size for one-way independent ANOVA. However, we saw in equation (7.4) that:

$$R^2 = \frac{SS_M}{SS_T}$$

We can actually get this value from the main ANOVA by using *summary.lm()* on the object you create with *aov()*. For example, for the *viagraModel* this function gives us Output 10.8, at the bottom of which we see that $r^2 = .46$. For some bizarre reason, in the context of ANOVA, r^2 is usually called **eta squared**, η^2. It is then a simple matter to take the square root of this value to give us the effect size, r ($\sqrt{.46} = .68$). Using the benchmarks for effect sizes this represents a large effect (it is above the .5 threshold for a large effect). Therefore, the effect of Viagra on libido is a substantive finding.

However, this measure of effect size is slightly biased because it is based purely on sums of squares from the sample and no adjustment is made for the fact that we're trying to estimate the effect size in the population. Therefore, we often use a slightly more complex measure called **omega squared (ω^2)**. This effect size estimate is still based on the sums of squares that we've met in this chapter, but like the *F*-ratio it uses the variance explained by the model, and the error variance (in both cases the average variance, or mean squared error, is used):

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

All of these values can be found in Output 10.5 (although SS_T is not in the output, it is easily calculated as $SS_T = SS_M + SS_R$). In this example we'd get:

$$\begin{aligned}\omega^2 &= \frac{20.13 - (2 \times 1.97), \text{ or } 20.13 - (2)1.97}{43.73 + 1.97} \\ &= \frac{16.19}{45.70} \\ &= .35 \\ \omega &= .60\end{aligned}$$

As you can see, this has led to a slightly lower estimate than using *r*, and in general ω is a more accurate measure. Although in the sections on ANOVA I will use ω as my effect size measure, think of it as you would *r* (because it's basically an unbiased estimate of *r* anyway). People normally report ω^2, and it has been suggested that values of .01, .06 and .14 represent small, medium and large effects respectively (Kirk, 1996). Remember, though, that these are rough guidelines and that effect sizes need to be interpreted within the context of the research literature.

OLIVER TWISTED

Please Sir, can I have some more … omega?

'There's no place like omega', chants Oliver as he clicks the heels of his red shoes together. Much as you want to wake up in Kansas, Oliver, you're going to find yourself in bubo-infested Dickensian London. If you'd like to join him there, read the online material, which shows you how to write a function to calculate ω^2 in **R**. I think you'll agree it's not entirely different from a bubo infestation.

Most of the time it isn't that interesting to have effect sizes for the overall ANOVA because it's testing a general hypothesis. Instead, we really want effect sizes for the differences between pairs of groups. We can obtain these using the *mes()* function of the *calculate.es* package. This function takes the general form:

```
mes(mean_group1, mean_group2, sd_group1, sd_group2, n_group1, n_group2)
```

In other words, we simply input the mean, standard deviation (*sd*) and sample size (*n*) of the two groups that we want to compare. We have this information in Output 10.3. For example, if we want to compare the placebo and low-dose group we would execute:

```
mes(2.2, 3.2, 1.3038405, 1.3038405, 5, 5)
```

We have entered the mean of the placebo group (2.2), the mean of the low-dose group (3.2), the standard deviation of the placebo group (1.3038), the standard deviation of the low-dose group (also 1.3038), and both groups have a sample size of 5. Similarly, we can get effect sizes for the difference between the placebo and high-dose group by executing:

```
mes(2.2, 5, 1.3038405, 1.5811388, 5, 5)
```

Finally, the difference between the low- and high-dose groups can be quantified by executing:

```
mes(3.2, 5, 1.3038405, 1.5811388, 5, 5)
```

The outputs of these commands are shown in Output 10.17 (I have edited them to show only the effect sizes *d* and *r*). The difference between the placebo and low-dose group is a medium-sized effect (the means are about three-quarters of a standard deviation different), $d = -0.77$, $r = -.36$; the difference between the placebo and high-dose group is a very large effect (a difference between the group means of almost 2 standard deviations), $d = -1.93$, $r = -.69$; finally, the difference between the low- and high-dose groups is a largish effect (more than a standard deviation difference between the group means), $d = -1.24$, $r = -.53$.

```
Placebo vs. Low Dose:
$MeanDifference
         d      var.d          g      var.g
-0.7669650  0.4294118 -0.6927426  0.3503214

$Correlation
          r       var.r
-0.35805743  0.07113067

Placebo vs. High Dose:
$MeanDifference
         d      var.d          g      var.g
-1.9321836  0.5866667 -1.7451981  0.4786126

$Correlation
          r       var.r
-0.69480834  0.02029603

Low Dose vs. High Dose:
$MeanDifference
         d      var.d          g      var.g
-1.2421180  0.4771429 -1.1219130  0.3892612

$Correlation
          r       var.r
-0.52758935  0.04482986
```

Output 10.17

An alternative is to compute effect sizes for the orthogonal contrasts. We can use the same equation as in section 9.5.2.8:

$$r_{\text{contrast}} = \sqrt{\frac{t^2}{t^2 + 26}}$$

We could write a function (see R's Souls' Tip 6.2) to do this computation for us in **R**:

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

Executing this command creates a function called *rcontrast*. First, we tell **R** that we want to be able to input *t* and *df* into the function (these are specified in brackets). This means that to use the function we have to input these values in brackets in the correct order. The rest of the function uses these values to compute *r* and then print the result. The first command takes the value of *t* and *df* entered into the function and places them into the equation written above in **R**-speak (because of how I have labelled everything in the function you should be able to compare directly the command with the equation above) to get a value of *r*. The command prints some text (in speech marks) followed by the value of *r*. If you can't be bothered to write out the command, you should be able to use it directly if you have the package associated with this book, *DSUR*, loaded (see section 3.4.5).

Having executed this function, we can use it to calculate *r* for the contrasts. Output 10.9 gives us the value of *t* for each contrast (2.474 and 2.029). The degrees of freedom can be calculated as in normal regression (see section 7.2.4) as $N - p - 1$, in which N is the total sample size (in this case 15), and p is the number of predictors (in this case 2, the two contrast variables). Therefore, the degrees of freedom are $15 - 2 - 1 = 12$. Therefore, we can execute the following commands:

```
rcontrast(2.474, 12)
rcontrast(2.029, 12)
```

The resulting values of *r* are

```
[1] "r =  0.581182458413787"
[1] "r =  0.505407970122564"
```

Both effects are fairly large.

10.8. Reporting results from one-way independent ANOVA ②

When we report an ANOVA, we have to give details of the *F*-ratio and the degrees of freedom from which it was calculated. For the experimental effect in these data the *F*-ratio was derived by dividing the mean squares for the effect by the mean squares for the residual. Therefore, the degrees of freedom used to assess the *F*-ratio are the degrees of freedom for the effect of the model ($df_M = 2$) and the degrees of freedom for the residuals of the model ($df_R = 12$). Therefore, the correct way to report the main finding would be:

- ✓ There was a significant effect of Viagra on levels of libido, $F(2, 12) = 5.12$, $p < .05$, $\omega = .60$.

Notice that the value of the *F*-ratio is preceded by the values of the degrees of freedom for that effect. Also, we rarely state the exact significance value of the *F*-ratio: instead we report that the significance value, *p*, was less than the criterion value of .05 and include an effect size measure. The linear contrast can be reported in much the same way:

- There was a significant linear trend, $F(1, 12) = 9.97$, $p < .01$, $\omega = .62$, indicating that as the dose of Viagra increased, libido increased proportionately.

Notice that the degrees of freedom have changed to reflect how the *F*-ratio was calculated. I've also included an effect size measure (have a go at calculating this as we did for the main *F*-ratio and see if you get the same value). Also, we have now reported that the *F*-value was significant at a value less than the criterion value of .01. We can also report our planned contrasts or group comparisons:

- Planned contrasts revealed that taking any dose of Viagra significantly increased libido compared to having a placebo, $t(12) = 2.47$, $p < .05$ (one-tailed), and that taking a high dose significantly increased libido compared to taking a low dose, $t(12) = 2.03$, $p < .05$ (one-tailed).
- Despite fairly large effect sizes, Bonferroni tests revealed non-significant differences between the low-dose group and both the placebo, $p = .845$, $d = -0.77$, and high-dose, $p = .196$, $d = -1.24$, groups. The high-dose group, however, had a mean almost 2 standard deviations bigger than the placebo group, $p = .025$, $d = -1.93$.

What have I discovered about statistics? ①

This chapter has introduced you to analysis of variance (ANOVA), which is the topic of the next few chapters also. One-way independent ANOVA is used in situations when you want to compare several means, and you've collected your data using different participants in each condition. I started off explaining that if we just do lots of *t*-tests on the same data then our Type I error rate becomes inflated. Hence we use ANOVA instead. I looked at how ANOVA can be conceptualized as a general linear model (GLM) and so is in fact the same as multiple regression. Like multiple regression, there are three important measures that we use in ANOVA: the total sum of squares, SS_T (a measure of the variability in our data), the model sum of squares, SS_M (a measure of how much of that variability can be explained by our experimental manipulation), and SS_R (a measure of how much variability can't be explained by our experimental manipulation). We discovered that, crudely speaking, the *F*-ratio is just the ratio of variance that we can explain to the variance that we can't. We also discovered that a significant *F*-ratio tells us only that our groups differ, not how they differ. To find out where the differences lie we have two options: specify specific contrasts to test hypotheses (*planned contrasts*), or test every group against every other group (*post hoc tests*). The former are used when we have generated hypotheses before the experiment, whereas the latter are for exploring data when no hypotheses have been made. Finally, we discovered how to implement these procedures in **R**.

We also saw that my life was changed by a letter that popped through the letterbox one day saying only that I could go to the local grammar school if I wanted to. When my parents told me, rather than being in celebratory mood, they were very downbeat; they knew how much it meant to me to be with my friends and how I had got used to my apparent failure. Sure enough, my initial reaction was to say that I wanted to go to the local school. I was unwavering in this view. Unwavering, that is, until my brother convinced me that being at the same school as him would be really cool. It's hard to measure how much I looked up to him, and still do, but the fact that I willingly subjected myself to a lifetime of social dysfunction just to be with him is a measure of sorts. As it turned out, being at school with him was not always cool – he was bullied for being a boffin (in a school of boffins) and being the younger brother of a boffin made me a target. Luckily, unlike my brother, I was not a boffin and played football, which seemed to be good enough reasons for them to leave me alone. Most of the time.

R packages used in this chapter

car
compute.es
ggplot2
multcomp
pastecs
Rcmdr
WRS

R functions used in this chapter

aov()
by()
cbind()
contrasts()
contr.helmert()
contr.poly()
contr.SAS()
contr.treatment()
gl()
glht()
levene.test()
lincon()
lm()
mcppb20()
med1way()
mes()
oneway.test()
pairwise.t.test()
read.csv()
read.delim()
stat.desc()
summary()
summary.lm()
t1way()
t1waybt()
unstack()

Key terms that I've discovered

Analysis of variance (ANOVA)
Bonferroni correction
Cubic trend
Eta squared, η^2
Experimentwise error rate
Familywise error rate
Grand variance
Harmonic mean
Helmert contrast
Independent ANOVA
Omega squared (ω^2)
Orthogonal
Pairwise comparisons
Planned contrasts
Polynomial contrast
Post hoc tests
Quadratic trend
Quartic trend
Treatment contrast
Weights
Welch's *F*

Smart Alex's tasks

- **Task 1**: Imagine that I was interested in how different teaching methods affected students' knowledge. I noticed that some lecturers were aloof and arrogant in their teaching style and humiliated anyone who asked them a question, while others

were encouraging and supporting of questions and comments. I took three statistics courses where I taught the same material. For one group of students I wandered around with a large cane and beat anyone who asked daft questions or got questions wrong (*punish*). In the second group I used my normal teaching style, which is to encourage students to discuss things that they find difficult and to give anyone working hard a nice sweet (*reward*). The final group I remained indifferent to and neither punished nor rewarded students' efforts (*indifferent*). As the dependent measure I took the students' exam marks (*percentage*). Based on theories of operant conditioning, we expect punishment to be a very unsuccessful way of reinforcing learning, but we expect reward to be very successful. Therefore, one prediction is that reward will produce the best learning. A second hypothesis is that punishment should actually retard learning such that it is worse than an indifferent approach to learning. The data are in the file **Teach.dat**. Carry out a one-way ANOVA and use planned comparisons to test the hypotheses that: (1) reward results in better exam results than either punishment or indifference; and (2) indifference will lead to significantly better exam results than punishment. ②

- **Task 2**: Earlier in this chapter we encountered some data relating to children's injuries while wearing superhero costumes. Children reporting to the emergency centre at hospitals had the severity of their injury (**injury**) assessed (on a scale from 0, no injury, to 100, death). In addition, a note was taken of which superhero costume they were wearing (**hero**): Spiderman, Superman, the Hulk or a Teenage Mutant Ninja Turtle. Use one-way ANOVA and multiple comparisons to test the hypotheses that different costumes are associated with more severe injuries. ②

- **Task 3**: In Chapter 15 (section 15.6) there are some data looking at whether eating soya meals reduces your sperm count. Have a look at this section, access the data for that example, but analyse them with ANOVA. What's the difference between what you find and what is found in section 15.6.4? Why do you think this difference has arisen? ②

- **Task 4**: Students (and lecturers for that matter) love their mobile phones, which is rather worrying given some recent controversy about links between mobile phone use and brain tumours. The basic idea is that mobile phones emit microwaves, and so holding one next to your brain for large parts of the day is a bit like sticking your brain in a microwave oven and hitting the 'cook until well done' button. If we wanted to test this experimentally, we could get six groups of people and strap a mobile phone to their heads (so that they can't remove it). Then, by remote control, we turn the phones on for a certain amount of time each day. After 6 months, we measure the size of any tumour (in mm^3) close to the site of the phone antenna (just behind the ear). The six groups experienced 0, 1, 2, 3, 4 or 5 hours per day of phone microwaves for 6 months. The data are in **Tumour.dat** (from Field & Hole, 2003, so there is a very detailed answer in there). ②

- **Task 5**: Using the Glastonbury data from Chapter 7 (**GlastonburyFestivalRegression.dat**), carry out a one-way ANOVA on the data to see if the change in hygiene (**change**) is significantly different across people with different musical tastes (**music**). Do a contrast to compare each group against 'No Affiliation'. Compare the results to those described in section 7.12. ②

- **Task 6**: Labcoat Leni's Real Research 15.2 describes an experiment (Çetinkaya & Domjan, 2006) on quails with fetishes for terrycloth objects (really, it does). In this example, you are asked to analyse two of the variables that they measured with a Kruskal–Wallis test. However, there were two other outcome variables (time spent near the terrycloth object and copulatory efficiency). These data can be analysed

with one-way ANOVA. Read Labcoat Leni's Real Research 15.2 to get the full story, then carry out two one-way ANOVAs and Bonferroni *post hoc* tests on the aforementioned outcome variables. ②

Answers can be found on the companion website.

Further reading

Howell, D. C. (2006). *Statistical methods for psychology* (6th ed.). Belmont, CA: Duxbury. (Or you might prefer his *Fundamental statistics for the behavioral sciences,* also in its 6th edition, 2007. Both are excellent texts that provide very detailed coverage of the standard variance approach to ANOVA but also the GLM approach that I have discussed.)

Iversen, G. R., & Norpoth, H. (1987). *ANOVA* (2nd ed.). Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001. Newbury Park, CA: Sage. (Quite high level, but a good read for those with a mathematical brain.)

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-061. Newbury Park, CA: Sage. (High-level but thorough coverage of multiple comparisons – in my opinion this book is better than Toothaker for planned comparisons.)

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioural research: A correlational approach*. Cambridge: Cambridge University Press. (Fantastic book on planned comparisons by three of the great writers on statistics.)

Rosnow, R. L., & Rosenthal, R. (2005). *Beginning behavioral research: A conceptual primer* (5th ed.). Upper Saddle River, NJ: Pearson/Prentice Hall. (Look, they wrote another great book!)

Toothaker, L. E. (1993). *Multiple comparison procedures*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-089. Newbury Park, CA: Sage. (Also high level, but gives an excellent précis of *post hoc* procedures.)

Wright, D. B.,& London, K. (2009). *First steps in statistics* (2nd ed.). London: Sage. (If this chapter is too complex then Wright and London's book is a very readable basic introduction to ANOVA.)

Interesting real research

Davies, P., Surridge, J., Hole, L., & Munro-Davies, L. (2007). Superhero-related injuries in paediatrics: A case series. *Archives of Disease in Childhood, 92*(3), 242–243.

Gallup, G. G. J., Burch, R. L., Zappieri, M. L., Parvez, R., Stockwell, M., & Davis, J. A. (2003). The human penis as a semen displacement device. *Evolution and Human Behavior*, *24*, 277–289.