Chapter 10

Introduction to the analysis of variance

chapter 10

THE *t* TEST IS LIMITED in two ways. First, it allows a comparison of only two samples at a time, such as old men versus young men on a particular task. In many cases we want to compare a number of samples, not just two, such as young men, middle-aged men and old men on the same task, and the *t* test cannot do this. Second, the *t* test examines the effect of only one independent variable, such as *age* or *teaching method*, at a time whereas we may want to compare them in combination. The analysis of variance is similar to the *t* test but is without these restrictions. It is for this reason that the analysis of variance (or ANOVA as it is known) is a very popular statistical technique in a range of research fields.

Factors and conditions

In the following chapters I shall be referring to independent variables as factors as that is the term used in the analysis of variance, so *age*, *hair colour* and *type of school attended* are all examples of factors. The conditions are the categories of the independent variable we choose to study. These are also referred to by other terms such as groups, levels or treatments, but I shall use conditions throughout. If we were investigating the independent variable of age we might select the conditions: 20 year olds, 40 year olds and 60 year olds. These age groups are the three conditions of the factor under study. We could, of course, choose different conditions for the variable *age* if we wish.

The problem of many conditions and the *t* test

Consider the situation where you want to compare more than two conditions. Rather than comparing children in a small school with children in a large school, you might want to compare a range of schools of different sizes (that we could label A, B, C, etc.). Similarly, you might want to compare three different teaching methods (A, B and C) on a group of children. The problem is to find a way to analyse the findings statistically. One solution

would be to perform a number of *t* tests, comparing each different pair of conditions: A and B, B and C, C and A, when there are three conditions. But we do not do this for the following reasons.

We have to perform three tests instead of one. If we had four conditions we would have to undertake six different tests and if there were ten conditions the number of tests would be forty-five! We really need one single test that allows us to deal with more than two conditions simultaneously, in fact a test that we do once and not have to do forty-five times.

The second and more important reason why we do not do lots of *t* tests is that the more *t* tests we perform on the data the more likely we are to make a Type I error (accept a result as significant when it occurred by chance). With one test, with $\alpha = 0.05$, we have a probability of 0.05 of making a Type I error. This means we have a probability of $1 - \alpha$ or 0.95 of *not* making a Type 1 error. If we perform two tests, each at the 0.05 level of significance, the probability of not making a Type I error becomes $0.95 \times 0.95 = 0.90$. The probability of making a Type I error in the tests is $1 - 0.90 = 0.10$. Already the probability of making at least one Type I error has doubled. With ten tests the probability of at least one Type I error rises to 0.40, or a 40 per cent chance.

If we want the *overall* significance level of a number of tests to be 0.05, then we have to set the significance level of each of the individual tests at a much more conservative level. If, for example, we undertake five tests then the significance level for each individual test has to be set at $p = 0.01$ for the overall risk of a Type I error to be 0.05 (as $1 - 0.99 \times 0.99 \times 0.99 \times 0.99 \times 0.99 = 0.05$).

The alternative is to devise a single test which has the same effect as the multiple comparisons but with an overall significance level set at $p = 0.05$. It is this alternative test that we consider now.

Why do scores vary in an experiment?

If we look at a set of data we find that not all the scores are identical. Why is there this variability in the data? The answer to this question holds the key to the analysis of variance as a means of hypothesis testing. Let us take an example to demonstrate this. We want to know the effect of the frequency of a word in the language on anagram solution times. We select a number of conditions, such as Common Words, Less Frequent Words, and Rare Words. We might use a computer-based store of words in the language (accessible

over the Internet, which gives the frequency of a word in a vast body of text) to select words appropriate to our conditions. In choosing words we make sure they differ in frequency but not in word length or other possible confounding factors. We then record the time it takes participants to solve a set of anagrams in each of the three conditions.

The null hypothesis predicts that the scores in all three conditions come from the same distribution. If there are differences between the mean solution times for the three conditions can we reject the null hypothesis and claim a genuine difference between the distributions of solution times according to word frequency? Unfortunately not, because even when the null hypothesis is true we will still find that we do not get equal means in the various conditions. What we need to find out is what causes the variation in the scores and how we can detect when the variation has arisen because of the manipulation of the factor, *word frequency*, and not for other reasons, such as the chance variation we would expect even when the null hypothesis is true.

Random variation in an experiment

It is unlikely that the participants in the same condition will produce exactly the same time for solving the anagrams. These are scores from a distribution and some participants will be fast and others slow rather than every one producing the population mean. The result is a sample of scores from a population and even if we select our sample randomly from the population there will be unsystematic or random errors that can lead to differences in the scores, and differences between the sample mean and the population mean. Even when the null hypothesis is true we would still expect the scores in the conditions to vary by random error and the means of the conditions to vary for the same reason.

When the scores come from different subjects one major category of random error is that of individual differences: participants will differ on their anagram solving ability, crossword puzzle experience and so forth. We can see from this why we need to select randomly from the population. If we select in a biased way, such as choosing only good crossword puzzlers, then their times would be systematically distorted from the population mean making them a poor estimate of it and we would not be able to generalise from our result to the wider population.

As well as individual differences, there will be a collection of other random errors, due to the difficulty of setting up equivalent conditions for

the participants. Someone might drop a pencil on the floor, another might remember a word from the crossword in that morning's newspaper and a third might be distracted by a noise. These could influence the anagram solution times. Thus we would expect scores to differ in an experiment due to a range of random errors regardless of whether the null hypothesis is true or not.

Systematic variation in the scores

If the null hypothesis is true and there are no differences in the populations of solution times for the different conditions of word frequency then any differences we find between condition means should be due to random error only. However, when the null hypothesis is false, the scores between conditions might be drawn from different populations (unlike the scores within a condition) and when this is the case we should find systematic differences between the conditions. We have deliberately chosen the anagrams so that they differ on word frequency between the conditions. If Common Word anagrams really are easier to solve than Less Frequent Word anagrams then we would expect this difference in population means to be reflected in our scores. If word frequency does affect solution times then we should expect systematic differences in the scores between conditions (known as a treatment effect). This is what we are looking for, evidence that there are genuine differences in the population means of the anagram solution times between the conditions.

Random errors and systematic differences

Scores in an experiment will vary due to random errors and systematic differences. If we have selected our subjects appropriately we would expect random errors to occur anywhere in the data rather than focused in any one condition. However, if there is a genuine effect of the independent variable and it does affect the scores then we would expect systematic differences between the scores in the different conditions. The random errors will provide a certain level of variability in the data both within and between the conditions, a sort of 'background noise' in the results. If the null hypothesis is false and there really are differences between the conditions we would expect this to appear as a systematic difference in the scores from the different conditions, over and above the 'background noise'.

Look at the three examples of results to this experiment in the table below.

	(a)			*(b)*			*(c)*		
	CW	*LFW*	*RW*	*CW*	*LFW*	*RW*	*CW*	*LFW*	*RW*
	17	16	19	18	18	40	20	30	40
	16	18	25	21	18	44	19	30	41
	22	21	19	16	20	38	21	31	39
	16	18	25	21	18	42	20	29	41
	23	24	18	18	23	37	21	29	40
	20	23	20	20	23	39	19	31	39
Mean	19	20	21	19	20	40	20	30	40

(The initials CW, LFW and RW in the table stand for Common Words, Less Frequent Words and Rare Words respectively. The scores are in minutes.)

What can we say about the causes of the variability of the scores in (a), (b) and (c)? The key thing is to decide whether there are systematic differences in the scores between the conditions. In example (a) there are differences between the condition means but only of 1. This is actually quite small compared to the 'background noise' of the random variability: there are both high and low scores in all three conditions. A set of results like this could quite easily occur when the null hypothesis is true and there are no genuine differences between the populations from which the samples are drawn. Example (b) looks more indicative of an underlying difference, but only between the Rare Words and the other conditions. All the high scores are in the Rare Word condition and a mean of 40 differs by at least 10 from the other condition means and looks larger than the variability in the data that could arise from random variability alone. In this example, there appears to be a difference in the underlying population distribution for Rare Words compared to the other two but not between Common Words and Less Frequent Words. Finally, in example (c) we have large differences between all three means that seem to dominate any random variability, indicating that the three conditions have drawn samples from different distributions.

What we need to do now is to produce a statistic that formally analyses the variability of the scores in an experiment, in an equivalent manner to my informal 'eyeballing' of the above examples and allows us to decide when the variability of the scores between conditions indicates genuine differences between populations (such as in examples (b) and (c)) and when it indicates only the random variation that we would expect by chance, when the null hypothesis is true (example (a)).

Calculating the variability of scores

We need to express the variability of the scores statistically. Up to now we have used the standard deviation to do this for a sample of scores: $\sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$. Now we are interested in comparing different sources of variability, to find whether there are systematic differences between conditions as well as random variability in the data, rather than seeking a standard difference from the mean. For this reason, and because we don't want to have to keep working out square roots, it is much easier for us to use variance, the square of the standard deviation:

$$\text{Sample variance, } s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$

At the heart of the variance calculation is the sums of squares: $\sum(X - \bar{X}^2)$. This measures the variability of the scores from the mean of the sample. When the scores vary wildly from the mean the sums of squares is large and when they cluster round the mean the sums of squares is small. This is what we want for our analysis of variability.

The sums of squares is also affected by the number of scores in the sample. The more scores we have the larger the sums of squares, even though the variability of the scores is no greater, as each extra score (unless exactly the same as the mean) will add to it. Consider the two samples, Sample 1 with scores 1, 1, 2, 3, 3 and Sample 2 with scores 1, 2, 3. Their variability looks about the same, with scores deviating from the mean by no more than 1. We can see from the table below that because there are more scores in Sample 1 the sums of squares is much larger.

Sample 1

X	$X - \bar{X}$	$(X - \bar{X})^2$
1	−1	1
1	−1	1
2	0	0
3	1	1
3	1	1

Sums of squares = $\sum(X - \bar{X}^2)$ = 4

Sample 2

X	$X - \bar{X}$	$(X - \bar{X})^2$
1	−1	1
2	0	0
3	1	1

Sums of squares = $\sum(X - \bar{X}^2)$ = 2

In order to take account of this we need to divide the sums of squares by the degrees of freedom, $df = n - 1$, to produce an 'average' variability of a score in the sample. (Recall from Chapter 5 that we use the degrees of freedom when dealing with samples as this produces a better estimate of the population parameter we are interested in.) There are five scores in Sample 1 so $n = 5$ and $df = n - 1 = 5 - 1 = 4$. This produces a variance of 1. In Sample 2 $n = 3$ and $df = 2$. This also produces a variance of 1. This matches our intuitive view that there is the same variability in these two samples.

We are interested in the variability produced by different factors in our data: random error and systematic differences and we can use the variance formula to find it.

The process of analysing variability

The useful thing about sums of squares is that we can calculate it for different portions of the data. We can work out the total sums of squares, taking into account every single score irrespective of condition. Using the data below, the overall mean is 10, taking into account all 18 scores, and the total sums of squares is 328.

	Condition 1	*Condition 2*	*Condition 3*
	5	11	14
	6	10	15
	7	9	17
	5	11	13
	3	9	17
	4	10	14
Mean	5	10	15

We can also work out the sums of squares for the scores within a single condition. The scores in Condition 1 have a mean of 5 and a sums of squares of the six scores is 10, for Condition 2 the sums of squares is 4 and for Condition 3 it is 14. If we add these up it will provide us with a measure of the variability of the scores within the conditions. The within conditions sums of squares is therefore 28 (the sum: 10 + 4 + 14). The scores also vary between the conditions. If we take just the three condition means 5, 10 and 15 they have a mean of 10 and a sums of squares of 50. These are not scores but means and each mean is composed of 6 scores so we multiply the figure of 50 by 6 to get the variability of the scores (rather than the means) between the conditions. The between conditions sums of squares is 300. If we use the label SS_{total} for the total sums of squares and $SS_{with.conds}$ and $SS_{bet.conds}$ for the within and between conditions sums of squares respectively, we can see that:

$$SS_{total} = SS_{with.conds} + SS_{bet.conds}$$
$$328 = 28 + 300$$

We can also separate the degrees of freedom in the same way. There are 18 scores in the experiment so the total degrees of freedom, $df_{total} = 18 - 1 = 17$. There are 6 scores in each condition giving $6 - 1 = 5$ degrees of freedom within each condition. Adding up the degrees of freedom within the three conditions we produce the within conditions degrees of freedom, $df_{with.conds}$, of 15. There are 3 conditions so there are $3 - 1 = 2$ between conditions degrees of freedom, $df_{bet.conds}$. We also see that:

$$df_{total} = df_{with.conds} + df_{bet.conds}$$
$$17 = 15 + 2$$

As we can partition both the sums of squares and the degrees of freedom into components we can also work out the variance within and between the conditions.

The variance ratio

What we want to do is to work out how much variability in the experiment is due to our manipulation, that is, the systematic differences between the conditions. The between conditions variance will tell us the 'average' variability between the conditions. This will arise from systematic differences between the conditions (if there are any) plus random errors (that will occur anywhere). This is not enough on its own to detect a difference in populations because this variance could be large for more than one reason; the systematic differences might be large or the random errors might be large, or both. What we need to do now is estimate the size of the variability due to the random errors.

Within a condition the scores will only vary due to random errors but not systematic differences (as the subjects within a condition will be performing in the same circumstances – we are not manipulating the independent variable within a condition). Assuming that random errors affect all scores equally (otherwise they would not be random) we can take the variance within the conditions as an estimate of the variance due to random errors, the error variance.[7]

Now if we compare the variance between conditions with the variance within conditions we will have a statistic for uncovering systematic differences between our conditions if there are any. We call this statistic, F, the variance ratio:

$$\text{Variance ratio } (F) = \frac{\text{Between conditions variance}}{\text{Error variance}}$$

This can also be expressed as follows:

$$\text{Variance ratio } (F) = \frac{\text{Systematic differences + Error variance}}{\text{Error variance}}$$

Note that the only difference between the top and the bottom of our equation is the systematic differences between the conditions, the error variance affecting the top and bottom equally. If there really are systematic differences between the conditions this should show up by a large value of F.

Alternatively, if the null hypothesis is true, and there are no differences between the distributions that the samples are drawn from, then we would expect to find no systematic differences between the conditions. Thus, when the null hypothesis is true, we would expect:

$$F = \frac{0 + \text{Error variance}}{\text{Error variance}} = \frac{\text{Error variance}}{\text{Error variance}} = 1$$

When the null hypothesis is true we expect F to equal 1 as the top and bottom of the equation are the same. When the null hypothesis is false we expect to find systematic differences between conditions and F to be greater than 1, with large systematic differences producing a large value of F.

The *F* distribution

Clearly, we need to know how large our calculated value of F must be for it to be significant at the level of significance chosen. What we need is the sampling distribution of F when the null hypothesis is true. If we select samples from the same distribution for our experimental conditions and calculated F, what values of F would we get?

First, the F values would cluster around 1 as there are no systematic differences between the conditions and the two variances making up the equation are likely to be equal. Second, F will never be less than zero as it is a ratio of numbers that have been squared and squares are never negative. This also means that we are only interested in one tail of the F distribution, the upper end: how much bigger than 1 the F value must be in order to reject the null hypothesis.

Like the t distribution F is also an estimate. We are using the variances of samples to estimate population values. Again, like t, the accuracy of this estimation will depend on the degrees of freedom of the estimate. Unlike t, however, the F statistic depends on two variances, the between conditions variance and the error variance, and so will be influenced by the degrees of freedom of both. This means that there is a different F distribution for each combination of the two degrees of freedom. Fortunately, the F distributions are known and the critical values for significance have been calculated for each combination (Table A.3 of the Appendix). As a result we can compare our calculated value of F with the appropriate table value to decide whether there are significant differences between the conditions or not.

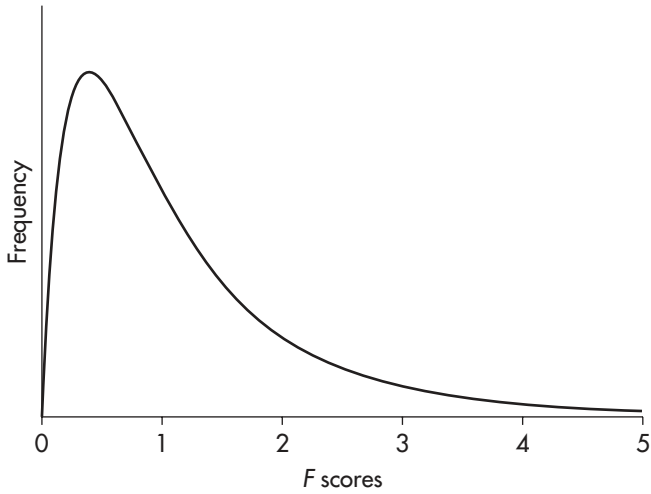

FIGURE 10.1 An example of an *F* distribution (degrees of freedom = 4,8)

In order to use the *F* distribution for comparison we have to make a number of assumptions: the samples for our conditions come from normally distributed populations, the samples come from distributions with equal variance, and the samples are randomly selected. These are very much the same assumptions that underlie the *t* test. When we perform the analysis of variance we must make these assumptions too, otherwise it may be inappropriate to compare our calculated value of *F* with those in the tables.

It is not surprising that I have been saying 'like *t*' throughout this section, for there is a simple relationship between *F* and *t* in the case where we can compare them (with two conditions): $F = t^2$. There is a demonstration of this in the next chapter. Figure 10.1 gives an example of an *F* distribution. It may look a little strange but imagine that all the scores in a *t* distribution (such as in Figure 6.2) were squared. All the negative values would become positive and it would turn into an *F* distribution like Figure 10.1. Another point to note about the fact that *F* is made up of squared numbers is that we no longer have the distinction between one-tailed and two-tailed tests. The squared values mean that any differences between the condition means will add to the size of *F*. Our prediction for *F* is simply that there are significant systematic differences between the conditions somewhere. A large value of *F* could mean that all the conditions differ significantly from each other or it could mean that only one differs from the others. It often needs further investigation to pin down the meaning of a significant *F* value.

Conclusion

By studying the variability in the data we have produced a statistic, the variance ratio F, that analyses the variance due to various factors in the data. The variance between the conditions contains the systematic differences between the conditions that we are seeking out. It also comprises the random errors that we expect with any data that we collect. Fortunately, we can estimate this error variance by looking at the data that is not affected by the systematic differences between conditions, the within conditions variance. When we examine the ratio of these two variances we have a statistic that provides an estimate of the systematic differences between conditions. If the calculated value of F is greater than the critical value of the F distribution at the chosen level of significance (say $p = 0.05$ or $p = 0.01$) then we can reject the null hypothesis and conclude that there are significant differences between at least some of the conditions.

By performing an analysis of variance we no longer have the problem of increasing the risk of Type I errors, as all conditions are compared in the one test, examined at a chosen level of significance. In the following chapters we shall see how the analysis of variance can be used to analyse data from a variety of different experimental designs.

Chapter 11

One factor independent measures ANOVA

chapter 11

THE ONE FACTOR independent measures ANOVA is similar to the independent *t* test but allows us to compare more than two conditions. It analyses data from an independent measures design, that is, employing different subjects in each condition. If we wanted to compare only two groups, such as 5 year old children to 7 year old children on a reading test then we could use either the *t* test or the ANOVA. We would get the same outcome regardless of which test we used. However, if we wanted to compare more groups, say, 5, 6 and 7 year olds then we would undertake the analysis of variance. (This form of ANOVA is also called the completely randomised design ANOVA.)

Analysing variability in the independent measures ANOVA

In the previous chapter we saw that the variability of the scores between the conditions arose from systematic differences between conditions plus random errors. In the independent measures design there are different subjects providing the scores for the different conditions, so part of the between conditions variance will be due to individual differences between the subjects. This is a random error as we are not systematically varying subjects across the conditions. The other random errors can be termed experimental error as we will always get some random errors in any experiment despite our attempts to provide equivalent conditions for the subjects. The between conditions variance can be seen as arising from three sources: systematic differences between the conditions, individual differences and experimental error.

If we look at the variability of the scores within the conditions we see that there are no systematic differences (if we have carried out the experiment properly) but there are still different subjects within a condition so we do expect variability due to individual differences. Again, as always, we expect other random errors that once again we can term experimental error as we expect it to occur at random anywhere in the experiment. The within conditions variance thus comprises two components: individual differences and experimental error. Therefore the within conditions variance provides us with the 'error variance' we need as it is influenced by the same variability as the between conditions variance apart from the systematic differences

between conditions. Comparing the between conditions variance with the within conditions variance will provide us with a variance ratio that we can compute and compare with the distribution of F in the search for an effect of our independent variable on the dependent variable.

We want to produce an F that is the following ratio:

$$F = \frac{\text{Systematic differences + Error variance}}{\text{Error variance}}$$

We can achieve this with the following:

$$F = \frac{\text{Between conditions variance}}{\text{Within conditions variance}}$$

This is because these variances only differ in the systematic differences between the conditions:

$$F = \frac{\text{Systematic differences + Individual differences + Experimental Error}}{\text{Individual differences + Experimental Error}}$$

To calculate F we must work out the between and within conditions variance.

The ANOVA summary table

The calculation of F requires us to build up the various components of the analysis of variance: the sums of squares, the degrees of freedom, the variances etc. In order to do this correctly and to display the results of the calculation clearly we produce an ANOVA summary table.

The summary table lists the sources of the variation in the scores as rows in the table. In the one factor independent measures ANOVA we are concerned with the variance between conditions and within conditions. We also need the total variability in the data in order to calculate the various sums of squares required. The columns provide the intermediate stages in the production of the variances needed for the variance ratio along with the final calculation of F and whether it is significant or not. We need the sums of squares and degrees of freedom to calculate variance. In the terminology of the analysis of variance we refer to variance as mean square (MS). It is simply an alternative label for the same thing. It can be considered more

descriptive in this context because dividing the sums of squares by the degrees of freedom produces an 'average' of the 'squares'.

The significance or otherwise of the calculated value of F can be indicated in the table in two ways. One, the specific probability of the F score of this size arising from the null hypothesis can be given: for example, $p = 0.0145$. In this case the reader can observe whether the probability is larger or smaller than a chosen significance level, such as $p = 0.05$. Alternatively, the probability can be given in relation to the significance level, such as $p < 0.05$ to indicate that the F value is significant at the $p = 0.05$ level of significance and $p > 0.05$ to indicate that it is not significant at the 0.05 significance level. I will use the latter convention.

For the one factor independent measures ANOVA the summary table is laid out in the following manner:

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	Variance ratio (F)	Probability
Between conditions	$df_{bet.conds}$	$SS_{bet.conds}$	$MS_{bet.conds}$	F	p
Within conditions	df_{error}	SS_{error}	MS_{error}		
Total	df_{total}	SS_{total}			

Notice that we only fill in the cells in the table we need for the variance ratio calculation. For example, we do not need the total variance as this is not required in the calculation of F. Below are listed the formulae for the calculation.

Degrees of freedom:

$df_{total} = N - 1$ where N is the total number of scores.

$df_{bet.conds} = k - 1$ where k is the number of conditions.

$df_{error} = df_{total} - df_{bet.conds}$

Sums of squares:

$$SS_{total} = \sum X^2 - \frac{(\sum X)^2}{N}$$

where $\sum X^2$ is the sum of the squared scores and $(\sum X)^2$ is the square of the sum of the scores.[8]

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{(\sum X)^2}{N}$$

where T refers to a total of the scores in a condition. $\sum T^2$ is the sum of the squared totals of the conditions and n is the number of scores in each condition.

$$SS_{error} = SS_{total} - SS_{bet.conds}$$

Mean square:

$$MS_{bet.conds} = \frac{SS_{bet.conds}}{df_{bet.conds}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

Variance ratio:

$$F = \frac{MS_{bet.conds}}{Ms_{error}}$$

We must always include the two degrees of freedom with our F value. We write it thus:

$$F(df_{bet.conds}, df_{error}) = \text{calculated value}$$

We compare the calculated value with the critical value in the F distribution tables at our chosen level of significance. When we look up the table value (Table A.3 in the Appendix) we use $df_{bet.conds}$ as our first degrees of freedom (the columns in the table) and df_{error} as our second degrees of freedom (the rows in the table). Our calculated value of F is only significant if it is equal to or larger than the table value.

A worked example

A researcher was interested in the effects of hints on anagram solution. The time it took a participant to solve five eight-letter anagrams was measured. The same five anagrams were used in three conditions: First Letter (where the first letter of the word was given), Last Letter (where the last letter was given) and No Letter (where no help was given). Thirty participants were chosen and ten were randomly allocated to each condition. The number of minutes it took to solve the five anagrams was recorded. These results are shown below. Is there an effect of *type of hint* (the independent variable) on solution times (the dependent variable)?

	First Letter Condition 1 X_1	*Last Letter Condition 2* X_2	*No Letter Condition 3* X_3
	15	21	28
	20	25	30
	14	29	32
	13	18	28
	18	26	26
	16	22	30
	13	26	25
	12	24	36
	18	28	20
	11	21	25
Mean	$\overline{X}_1 = 15.00$	$\overline{X}_2 = 24.00$	$\overline{X}_3 = 28.00$
Total	$T_1 = 150$	$T_2 = 240$	$T_3 = 280$
Squared total	$T_1^2 = 22500$	$T_2^2 = 57600$	$T_3^2 = 78400$

Sum of the scores (overall total): $\sum X = 670$
Square of the sum of the scores: $(\sum X)^2 = 448900$
Sum of the squared scores: $\sum X^2 = 16210$

Number of conditions: $k = 3$
Number of scores per condition: $n = 10$
Total number of scores: $N = 30$

Degrees of freedom:

$$df_{total} = N - 1 = 30 - 1 = 29$$
$$df_{bet.conds} = k - 1 = 3 - 1 = 2$$
$$df_{error} = df_{total} - df_{bet.conds} = 29 - 2 = 27$$

Sums of squares:

$$SS_{total} = \sum X^2 - \frac{(\sum X)^2}{N} = 16210 - \frac{448900}{30}$$
$$= 16210 - 149363.33 = 1246.67$$

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{(\sum X)^2}{N} = \frac{22500 + 57600 + 78400}{10} - \frac{448900}{30}$$
$$= 15850 - 14963.33 = 886.67$$

$$SS_{error} = SS_{total} - SS_{bet.conds} = 1246.67 - 886.67 = 360.00$$

Mean square:

$$MS_{bet.conds} = \frac{SS_{bet.conds}}{df_{bet.conds}} = \frac{886.67}{2} = 443.33$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{360.00}{27} = 13.33$$

Variance ratio (F):

$$F = \frac{MS_{bet.conds}}{MS_{error}} = \frac{443.33}{13.33} = 33.26$$

From the tables of the F distribution (Table A.3 in the Appendix) we find that $F(2,27) = 3.35$, at $p = 0.05$. As our value of 33.26 is greater than the table value we can reject the null hypothesis and claim that anagram solution times are affected by the type of hint given. Note that the result is highly significant, so we can adopt an even more conservative significance level. From the tables $F(2,27) = 5.49$, at $p = 0.01$, so our finding is still significant at $p < 0.01$.

The fact that we have found a significant effect does not tell us which conditions are significantly different although we can infer this by looking at the means. We will be able to be more specific in the following chapter. Also the *F* test has found significant differences between the conditions but it does not give the cause. We hope the experiment is so well controlled that it can only be due to *type of hint* but if the researcher introduced any inadvertent confounding factor this could also have produced the systematic differences picked up by the analysis of variance.

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	Variance ratio (*F*)	Probability
Between conditions	2	886.67	443.33	33.26	$p < 0.01$
Within conditions	27	360.00	13.33		
Total	29	1246.67			

The above table clearly summarises the analysis. It also allows us to check our calculations: do the degrees of freedom and the sums of squares add up to the correct totals? You must never get a negative sums of squares as a sum of squares has to be positive. If you do, check the calculations, there is definitely an error.

Rejecting the null hypothesis

When we reject the null hypothesis in an ANOVA, as we have done in the example above, we are only concluding that there are systematic differences between the conditions but not where they lie. In the case of three conditions there are four alternative hypotheses to the null hypothesis:

1 All three conditions are significantly different, their samples come from different population distributions.
2 Condition 1 is significantly different to conditions 2 and 3 but conditions 2 and 3 are not significantly different. The sample in condition 1 comes from a different distribution to the samples of conditions 2 and 3.

3 Condition 2 is significantly different to conditions 1 and 3 but conditions 1 and 3 are not significantly different. The sample in condition 2 comes from a different distribution to the samples of conditions 1 and 3.
4 Condition 3 is significantly different to conditions 1 and 2 but conditions 1 and 2 are not significantly different. The sample in condition 3 comes from a different distribution to the samples of conditions 1 and 2.

With more conditions the number of alternative hypothesis increases. A significant F value simply indicates that the null hypothesis is very unlikely and hence we can reject it. We need to perform further tests to decide which one of the alternative hypotheses to accept.

Unequal sample sizes

Researchers often organise the samples in the independent measures ANOVA so that there are equal numbers of subjects in each condition. It is not necessary but makes the calculation slightly easier. Yet the test, like the independent t test, allows for different sample sizes. The formulae given above are for equal sample sizes. However, the only change we need to make for unequal sample sizes is to the first term in the $SS_{bet.conds}$ formula. We replace $SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{(\sum X)^2}{N}$ with $SS_{bet.conds} = \sum\left(\frac{T^2}{n}\right) - \frac{(\sum X)^2}{N}$. We have a different n for each sample and we divide the squared total of each condition by its sample size *before* we add them up. A worked example is shown below.

Unequal sample sizes usually occur when you have planned for equal numbers in each condition but for some reason a subject is unable to provide a score. In the anagram example we might find a person who simply cannot solve an anagram no matter how much time allowed. One solution is to replace the participant with another. However, the change to the formula is so small that unequal sample sizes are not really a problem (as long as the equal population variance assumption is still met).

A worked example

As an example of the calculation of unequal sample sizes I shall take the data we used to calculate the independent t test in Chapter 8. This compared the effects of a sleeping pill on 6 men and 8 women. The scores for the men

(Condition 1) were 4, 6, 5, 4, 5 and 6 extra hours slept and for the women (Condition 2) were 3, 8, 7, 6, 7, 6, 7 and 6 extra hours.

Sum of the scores (overall total): $\sum X = 80$
Square of the sum of the scores: $(\sum X)^2 = 6400$
Sum of the squared scores: $\sum X^2 = 482$

Number of conditions: $k = 2$
Number of scores per condition: $n_1 = 6$, $n_2 = 8$
Total of the scores in condition 1, $T_1 = 30$ and the squared total, $T_1^2 = 900$
Total of the scores in condition 2, $T_2 = 50$ and the squared total, $T_2^2 = 2500$
Total number of scores: $N = 14$

Degrees of freedom:

$$df_{total} = N - 1 = 14 - 1 = 13$$

$$df_{bet.conds} = k - 1 = 2 - 1 = 1$$

$$df_{error} = df_{total} - df_{bet.conds} = 13 - 1 = 12$$

Sums of squares:

$$SS_{total} = \sum X^2 - \frac{(\sum X)^2}{N} = 482 - \frac{6400}{14}$$
$$= 482 - 457.14 = 24.86$$

$$SS_{bet.conds} = \sum\left(\frac{T^2}{n}\right) - \frac{(\sum X)^2}{N} = \left(\frac{900}{6} + \frac{2500}{8}\right) - \frac{6400}{14}$$
$$= 462.5 - 457.14 = 5.36$$

where $\sum\left(\frac{T^2}{n}\right) = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2}\right)$ as there are two conditions.

$$SS_{error} = SS_{total} - SS_{bet.conds} = 24.86 - 5.36 = 19.50$$

Mean square:

$$MS_{bet.conds} = \frac{SS_{bet.conds}}{df_{bet.conds}} = \frac{5.36}{1} = 5.36$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{19.50}{12} = 1.625$$

Variance ratio (F):

$$F = \frac{MS_{bet.conds}}{MS_{error}} = \frac{5.36}{1.625} = 3.30$$

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	Variance ratio (F)	Probability
Between conditions	1	5.36	5.36	3.30	$p > 0.05$
Within conditions	12	19.50	1.625		
Total	13	24.86			

From the F distribution tables (Table A.3 in the Appendix) we find $F(1,12) = 4.75$ at $p = 0.05$. As the calculated value of 3.30 is less than the table value we cannot reject the null hypothesis at this level of significance.

The relationship of *F* to *t*

The example in the section above allows us to compare an ANOVA with an independent t test on the same two samples. If you look back to the t calculations you can see the similarity in the calculations; for example note the SS_{error} of 1.625 in the bottom of the t calculation. If we explored further we could see how the two formulae are related. The calculated F of 3.30 is

indeed the square of the calculated t of 1.82.[9] Similarly, the table values of F and t are also related in the same way and so we have the same outcome whichever of the tests we perform on the data.

Details on calculating the one factor independent measures ANOVA using the SPSS computer statistical package can be found in Chapter 10 of Hinton *et al*. (2004).

Chapter 12

Multiple comparisons

chapter 12

When we compare more than two groups in an ANOVA a significant F value does not indicate where the effect lies, simply that there is an effect between the conditions somewhere. A researcher compared four groups of children (6, 8, 10 and 12 year olds) on a test of social skills. She found a significant F value and concluded that the scores from the four conditions were not drawn from the same distribution. But this conclusion does not really provide the researcher with the information about which ages show the significant differences. Let us assume that the means were respectively 10, 12, 18 and 23 on the test (out of 50). Given that there is a significance variance ratio it seems likely that the scores of the 6 year olds differ significantly from those of the 12 year olds as this comparison provides the largest difference in means. Is the difference between the 6 and 8 year olds or 8 year olds and 10 year olds significant? And what about the smallest difference, between the 6 and 8 year olds? The data needs to be inspected further to find the source of the significant F value.

The way we answer these questions is to perform post hoc tests. The name comes from the Latin, meaning 'after this'. The first stage in the analysis is to find a significant F value in the ANOVA. Only then do we perform a post hoc test. These tests are called multiple comparison tests as they allow us to undertake various comparisons between the conditions. In the example above we want to compare each of the four groups with each of the others to show where the significant differences lie.

The problem with multiple comparisons is that the more comparisons we make using the same data the greater the risk of making at least one Type I error. We saw in Chapter 10 that this was the same problem we had with undertaking multiple t tests: when we start undertaking multiple tests on the data we increase the risk of finding differences by chance. The solution is to find a post hoc test that takes account of this increased risk and controls for it.

There are a range of multiple comparison tests. Some of these ignore the problem completely. The Least Significant Difference test takes no account of the number of comparisons being made and the increased risk of a Type I error is simply accepted. Other tests such as the Newman–Keuls and the Duncan tests take account of the number of comparisons being made and compute different values accordingly. At the more conservative

end of the scale the Tukey and Scheffé tests allow all comparisons to be made as the test corrects for the increased risk of Type I errors by reducing the significance level of the individual comparisons. The simplest and most conservative method is to apply a Bonferroni correction to the significance level. For example, if a one factor independent measures ANOVA had shown a significant F value then follow-up t tests on each of the 6 pairs of conditions could be undertaken with a Bonferroni correction to the significance level for these tests. The Bonferroni correction requires us to divide the significance level by the number of tests, so in this case we would compare each test against the 0.05/6 level of significance ($p = 0.0083$) rather than the 0.05 significance level. This does influence the power of the test (see Chapter 9) and can be viewed as overly conservative due to the reduction in power.

I am going to describe the Tukey and the Scheffé tests, both conservative tests, for the following reasons. Usually, after we have found a significant varaince ratio in the ANOVA, we want to compare all the conditions to find the interesting (significant) differences, such as in the social skills test example above. The Tukey and Scheffé tests allow us to do this without worrying unduly about the risks of Type I errors. Second, they are easy to carry out, particularly the Tukey test. The fact that they set high critical values for significance need not lead us to miss out on potentially significant findings because we have set too rigorous a criterion for significance. We might not accept some differences as significant when using these tests when we would with some other tests but this does not have to be a problem if we remember to use our judgment as researchers. If there is a difference which does not quite reach significance using these tests yet we have reason to believe that it is an important difference then, as in other cases of this kind, we should trust our judgement and follow it up: replicate the experiment, run more subjects, use a more sensitive design, essentially adopt measures to improve the power of our test. If it is a genuine difference it will eventually show, even with a Tukey test. Statistics are only tools to help us. They do not replace experimenter skill and intelligence. I happen to like a conservative test as it gives me confidence in the results of the analysis. But I do not let it disturb my interest in the comparisons that 'bubble under' (do not quite reach significance). I check these out in subsequent experiments.

The reason for presenting both the Tukey and the Scheffé tests is that the Tukey test is more sensitive for pairwise comparisons, comparing two conditions at a time, than the Scheffé test, in that it is more likely to accept a difference as significant. The Scheffé test, however, is more sensitive than the Tukey test for complex comparisons, combining conditions and

comparing the composite condition with others, such as comparing the 8 year olds with the combination of the 10 and 12 year olds on the social skills test.

The Tukey test (for all pairwise comparisons)

The Tukey HSD (honestly significant difference) test allows us to compare each pair of conditions to see if their difference is significant. What the Tukey test does is to look at the random variation that exists between any pair of means. This is the standard error of the difference between pairs of means. If we then compare a specific difference between two means with this standard error we have a statistic for telling us how big the difference between the mean is compared to the random variation between means. We call this statistic q:

$$q = \frac{\text{the difference between any two means}}{\text{the standard error of the difference between any two means}}$$

We already have a measure of the error variance that we can take from the ANOVA, MS_{error}. The standard deviation is the square root of the variance: $\sqrt{MS_{error}}$ and so the standard error of the differences in means is $\frac{\sqrt{MS_{error}}}{\sqrt{n}}$, where n is the number of subjects in each condition. Hence:

$$q = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{MS_{error}}{n}}}$$

where $\overline{X}_i$ and $\overline{X}_j$ are any two means (the i and j standing for 1, 2, 3, etc. or whichever means we choose to compare).

Notice the similarity of q to t. This is not by chance; the logic of the two statistics is the same. With a t test we use a different standard error for every pair of means:

$$t = \frac{\text{the difference between two means}}{\text{the standard error of the difference between the two means}}$$

With q, however, we are using a 'general purpose' standard error that can be used for any pair of means. Like t we can find the distribution of q

under the null hypothesis. Using this distribution we can decide whether a specific difference in means is significant by observing whether the calculated q exceeds the table value of q for the level of significance chosen. The Tukey test overcomes the problem of the increased risk of Type I errors that occurs with multiple t tests by setting an overall level of significance. This means that the risk of a Type 1 error has a probability of, say, 0.05 when we compare every pair of means. Thus the Tukey test allows *all* pairwise comparisons so we can work out q for each pair of means knowing that the risk of a Type I error will not exceed 0.05. In the social skills test example we can make six pairwise comparisons as we have four conditions. If we had five age groups: 6, 8, 10, 12 and 14 year olds, as long as we achieved a significant F in the ANOVA, the Tukey test would allow us to make every one of the 24 pairwise comparisons between condition means.

Rather than working out a q every time we compare a pair of means we can rearrange the formula as follows:

$$\bar{X}_i - \bar{X}_j = q\sqrt{\frac{MS_{error}}{n}}$$

If we no longer calculate q but use the critical value (for significance) of q from the table in the formula we can write:

$$\text{An honestly significant difference between means, HSD} = q\sqrt{\frac{MS_{error}}{n}}$$

All we need to do is look up q at the chosen significance level, work out Tukey's HSD and use HSD to compare any or all of the differences in means. If a difference in means is greater than HSD then that difference is significant (honestly!).

The statistic q is called the Studentized range statistic (after a famous statistician who wrote under the pseudonym of Student. You also see t referred to as Student's t for the same reason). We find the appropriate value of q in the table (Table A.4 in the Appendix) by deciding on the level of significance we require (usually either 0.05 or 0.01), and then looking up the critical value of q in the table using df_{error}, the degrees of freedom of the error variance in the ANOVA and k, the number of conditions in the experiment.

(Normally, with equal numbers of subjects in each condition the Tukey HSD test is easy to undertake but with different sample sizes we cannot put

a single n in the equation for HSD as there are different ns: n_1, n_2, etc. To deal with this we can be cautious and simply take the smallest sample size as n. A more sophisticated way of producing a single ('average') n is by the following formula:

$$n = \frac{k}{\left(\frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_k}\right)}$$

where n_1 to n_k are the sample sizes. However, we should be wary of using the test with any but relatively small differences in sample sizes as the basic assumptions of the test may be violated.)

A worked example

The anagram example of the previous chapter provides a good example as we found a significant effect of *type of hint* on anagram solution times. The significant F value allows us to undertake post hoc tests, and see which differences in means are significant. The means are shown in the table below.

	First Letter $\overline{X}_1$	*Last Letter* $\overline{X}_2$	*No Letter* $\overline{X}_3$
Mean	15	24	28

Taking each pair of means we can work out the difference between them:

Difference of means	$\overline{X}_2$	$\overline{X}_3$
$\overline{X}_1$	−9	−13
$\overline{X}_2$		−4

The differences in the table are calculated by subtracting the column mean from the row mean. The fact that our differences are negative is a result of the way we have subtracted the means. This indicates $\overline{X}_1$ is faster than $\overline{X}_2$ by

9 minutes, etc. For the moment we are only concerned about the difference in the size of the means not whether the difference is positive or negative at this point. For the Tukey test we treat all the differences as positive.

From the ANOVA summary table we have $df_{error} = 27$ and $MS_{error} = 13.33$. The number of conditions, k, is 3, and the number of participants in each condition, $n = 10$. Selecting a significance level of $p = 0.05$, we can work out HSD. From the tables (Table A.4 in the Appendix) at $p = 0.05$, for $k = 3$ and $df_{error} = 27$, we find a value of q of 3.51. (As $df = 27$ is not in the table we take the figure midway between that for $df = 24$ and $df = 30$ for our value for $df = 27$.)

$$\text{HSD} = q\sqrt{\frac{MS_{error}}{n}} = 3.51 \times \sqrt{\frac{13.33}{10}} = 3.51 \times 1.15 = 4.04$$

The differences between the First Letter and No Letter conditions (13) and the First Letter and Last Letter conditions (9) are highly significant at $p = 0.05$ as they both exceed HSD. The difference between the Last Letter and No Letter conditions (4) is not significant at $p = 0.05$ but further investigations might find an effect here as the difference does approach significance but does not reach it. Now we know where the significant differences lie we check to see which way the differences occur (which condition produces the faster times) for our conclusion.

We can conclude that the First Letter condition results in significantly faster solution times than both the Last Letter and No Letter conditions. The Last Letter times are not significantly faster than the No Letter condition (although there appears be a non-significant tendency for the Last Letter times to be faster).

We can very easily work out confidence intervals for our comparisons, as we know the difference in means, we have the appropriate critical value and we also have a standard error (see Chapter 6 for an introduction to confidence intervals). So we can write the confidence interval as follows:

$$95\%\text{CI} = \overline{X}_i - \overline{X}_j \pm q\sqrt{\frac{MS_{error}}{n}}$$

where $\overline{X}_i$ and $\overline{X}_j$ are the two means we are comparing, q is the critical value (and we found that above) and $\sqrt{\frac{MS_{error}}{n}}$ is the standard error of the comparison (which we also found out above). Furthermore, $q\sqrt{\frac{MS_{error}}{n}} = \text{HSD}$, so:

$$95\%\text{CI} = \overline{X}_i - \overline{X}_j \pm \text{HSD}$$
$$95\%\text{CI} = \overline{X}_i - \overline{X}_j \pm 4.04$$

And now we can produce the confidence intervals for our three comparisons:

For $\overline{X}_1 - \overline{X}_2$, 95%CI = −9 ± 4.04, producing 95%CI = (−13.04, −4.96)
For $\overline{X}_1 - \overline{X}_3$, 95%CI = −13 ± 4.04, producing 95%CI = (−17.04, −8.96)
For $\overline{X}_2 - \overline{X}_3$, 95%CI = −4 ± 4.04, producing 95%CI = (−8.04, +0.04)

It is interesting to note that for the first two comparisons the differences are consistent across the confidence interval and even in the 'worst case' are still quite large (4.96 and 8.96 seconds difference). However, the third confidence interval includes zero so, even though the 'best case' gives us a difference of 8.04 seconds, the difference might still be zero. Even though the zero is near the end of the interval we cannot confidently exclude the possibility. The confidence intervals are expressing the findings in a different way to the significance test but the same implication arises: we can be confident that only the first two differences imply genuine population differences.

The Scheffé test (for complex comparisons)

Out of the 'between conditions sums of squares' the Scheffé test calculates the part of it relevant to the comparison being made. From the sums of squares of the comparison we can then go on to produce a mean square and then an *F* value for the comparison. We can test this against the *F* distribution to see if the comparison is significant. To correct for the increase in the risk of a Type I error that could arise with multiple comparisons we adjust the size of the table value of *F* according to the Scheffé correction. The calculated value of *F* for the comparison has to be larger than the corrected table value before we can claim a significant difference between the conditions being compared.

The Scheffé test is most useful for complex post hoc comparisons. In the example of the social skills experiment cited at the beginning of this chapter we shall assume that the researcher was interested in the difference between the children under 10 years old and the 10 year old group. Here we have a complex comparison, as two groups are being combined (the 6 and 8 year olds) to compare with the 10 year olds with one group being left out of the comparison (the 12 year olds) altogether.

The Scheffé test calculates a sums of squares for the comparison of interest by the following formula:

$$SS_{comp} = \frac{(\sum cT)^2}{n\sum c^2}$$

where the Ts are the totals of the scores in the conditions (T_1 is the total of the scores in condition 1, etc.), n is the number of subjects in each condition, and the cs are the coefficients of the conditions (c_1 is the coefficient of condition 1, etc.).

The choice of coefficients allows us to select the conditions we are interested in, in the correct combination, and exclude the conditions we do not wish to be included in the comparison. Essentially they 'weight' the contribution of the condition to the comparison. The conditions on one side of the comparison are given positive coefficients and the ones of the other side given negative coefficients. In order to properly balance the comparison the coefficients must sum to zero, $\sum c = 0$. In an experiment with three conditions where the comparison to be made is between condition 1 on one side with a combination of conditions 2 and 3 on the other then the coefficients could be $c_1 = +1$, $c_2 = -0.5$, $c_3 = -0.5$. Notice that the sum of the coefficients equal zero: $c_1 + c_2 + c_3 = 1 - 0.5 - 0.5 = 0$. Conditions 2 and 3 are equally weighted on their side of the comparison, as each is given the same coefficient of −0.5. The two sides of the comparison are equally weighted with +1 on one side and −1 on the other. (The actual numbers we choose for the coefficients can be anything as long as the above restrictions are met, so we could have chosen +2, −1, −1 for the coefficients or +10, −5, −5. We usually choose the ones that make the calculations easiest.)

The choice of coefficients results in a sums of squares for the comparison only. This comparison is always between two new conditions that are combinations of the experimental conditions. In the above paragraph the two new conditions are: condition 1 from the original experiment as the first new condition and a combination of conditions 2 and 3 as the second new condition. As there are always two conditions in the comparison the degrees of freedom for the comparison is always 1.

Hence the mean square for the comparison is:

$$MS_{comp} = \frac{SS_{comp}}{df_{comp}} = \frac{SS_{comp}}{1} = SS_{comp}$$

The calculated variance ratio for the comparison uses the error mean square from the original ANOVA, so the F value for the comparison is:

$$F = \frac{MS_{comp}}{MS_{error}}$$

At this point we must select the correct table value to compare our calculated F with. This depends on whether the comparison is planned prior to the calculation of the ANOVA or whether it was unplanned; that is, a post hoc comparison made after the significant ANOVA F value had been found.

A planned comparison

With a planned comparison we are saying that, prior to knowing whether the ANOVA F value was significant or not, we were interested in this comparison in particular. In this case we are not concerned with the increased risk of Type I errors with multiple comparisons as this is the only comparison of interest. Hence we can look up the table value using the degrees of freedom contributing to the comparison F value: df_{comp} and df_{error} at the chosen level of significance.

Unplanned comparisons

Unplanned comparisons are more usual in the use of the Scheffé test as post hoc tests are used to seek out the interesting results after a significant ANOVA. We may have certain comparisons in mind prior to the experiment but the data can lead us to follow up the most interesting, and unexpected, lines of research. As we wish to make any comparison *post hoc* we need to correct for the increased risk of a Type I error. The Scheffé test does this by creating a new, larger table value F'. Only if the calculated value exceeds F' can we say the comparison is significant. We calculate F' by the following formula: $F' = (k - 1)F$, where k is the number of conditions in the original experiment and F is the table value used in the original ANOVA, found using degrees of freedom $k - 1$ and $k(n - 1)$. The calculation of F' allows us to undertake any post hoc comparison without worrying about increasing the risk of a Type I error.

A worked example

At the beginning of this chapter I briefly mentioned a social skills study looking at four different age groups of children. The researcher was looking for an effect of age on the social skills test. The analysis produced the following summary table for the one factor independent measures ANOVA, with a highly significant F value:

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	Variance ratio (F)	Probability
Between conditions	3	838.00	279.33	12.415	$p < 0.01$
Within conditions	28	630.00	22.50		
Total	31	1468.00			

In this experiment there were eight children in each condition. The totals of the scores of the four conditions are shown below:

Condition 1 *6 year olds*	*Condition 2* *8 year olds*	*Condition 3* *10 year olds*	*Condition 4* *12 year olds*
T_1 80	T_2 96	T_3 144	T_4 184

The researcher decided post hoc that she wanted to know whether there was a significant difference between the 10 year olds and the younger children, combining the 6 and 8 year olds. To produce this comparison she chose the coefficients $c_1 = +1$, $c_2 = +1$, $c_3 = -2$ and $c_4 = 0$. These coefficients exclude condition 4 and combine conditions 1 and 2, which are then balanced on the other side of the comparison to condition 3.

The sums of squares of the comparison is calculated from the formula:

$$SS_{comp} = \frac{(c_1T_1 + c_2T_2 + c_3T_3 + c_4T_4)^2}{n(c_1^2 + c_2^2 + c_3^2 + c_4^2)}$$

$$SS_{comp} = \frac{((+1 \times 80) + (+1 \times 96) + (-2 \times 144) + (0 \times 184))^2}{8((+1)^2 + (+1)^2 + (-2)^2 + (0)^2)}$$

$$= \frac{-112^2}{8 \times 6} = \frac{12544}{48} = 261.33$$

As the degrees of freedom of the comparison is 1,

$$MS_{comp} = \frac{SS_{comp}}{df_{comp}} = \frac{261.33}{1} = 261.33$$

Using the error variance from the ANOVA,

$$F = \frac{MS_{comp}}{MS_{error}} = \frac{261.33}{22.50} = 11.61$$

We now calculate F':

$$F' = (k - 1)F(k - 1,k(n - 1)) = (4 - 1)F(4 - 1,4(8 - 1))$$

$$= 3F(3,28)$$

From the tables $F(3,28) = 2.95$, $p = 0.05$, so

$$F' = 3 \times 2.95 = 8.85$$

As the calculated value of F is greater than F' we can conclude that there is a significant difference in the performance of the 10 year olds compared to the combination of the 6 and 8 year olds on the social skills test, with the 10 year olds scoring significantly higher than the younger children.