

1

An introduction to analysis of variance

1.1 Model formulae and geometrical pictures

The approach we are going to take centres around two concepts: model formulae and geometry. The model formulae are not mathematical formulae but ‘word formulae’, and are an expression of the question we are asking. If we have data on the weight of 50 male and 50 female squirrels, we could hypothesise that these two groups are different. The question we are asking is, ‘Is weight explained by the sex of the squirrel?’ If the two groups are different, then knowing the sex of the squirrel will help you to predict its weight. The corresponding model formula would be:

$$\text{WEIGHT} = \text{SEX}$$

Variable names such as `WEIGHT` and `SEX` will be represented in a different font. On the left hand side of the equation is the data variable: i.e. the variable we wish to explain. We would like to know which factors are important in determining the magnitude of this variable. On the right hand side is the explanatory variable: i.e. the variable which we suspect to be important in determining the weight of a squirrel. So this simple model formula embodies our question. Later we will see how to develop these formulae to ask more complex questions.

It is the aim of this book to provide an understanding of the concepts behind the statistical analyses done, but not to expound the mathematical details. It is not necessary to be able to do the matrix algebra that lies behind the tests now that statistical packages are provided to do that for you. It is necessary however to have an understanding of the underlying principles. These principles will be illustrated using geometrical pictures rather than maths. These pictures will be used to illustrate different concepts in the early chapters.

1.2 General Linear Models

The combined approach of using model formulae and geometrical analogies (and this by-pass of mathematical details) has been made possible by a technique

known as General Linear Modelling (GLM). Whilst this technique has been used for many years by statisticians, it is only relatively recently that it has been incorporated into user friendly packages used by non-specialists—Minitab being a notable example. GLMs are developed here in a framework applicable to a wide range of packages: most particularly Minitab, SAS, and SPSS, but others such as GENSTAT, BMDP, GLIM and SPLUS have a similar interface.

Having introduced the idea of General Linear Models, we will first of all turn our attention to the analysis of variance (ANOVA) for the rest of this chapter, followed by regression in the next chapter. It is a central message of Chapter 3 however that these two kinds of analysis are forms of GLM. Indeed, one of the great advantages of using GLMs is that a number of tests that have been traditionally considered separately, all come under one umbrella. They can all be expressed in terms of model formulae and the geometrical analogy, they are all subject to the same set of assumptions, and these assumptions can be tested using a common set of procedures (see Chapters 8 and 9).

1.3 The basic principles of ANOVA

The first and simplest problem to consider is the comparison of three means. This is done by the analysis of variance (ANOVA). The aim of this section is to look at an example in some detail. This will be done by actually working through the numerical mechanics, and relating it to the output. Once the origin of the output has been derived from first principles, it will not be necessary to do this again. This section will also provide you with your first introduction to model formulae, and the geometrical representation of the analysis you are conducting.

If we have three fertilisers, and we wish to compare their efficacy, this could be done by a field experiment in which each fertiliser is applied to 10 plots, and then the 30 plots are later harvested, with the crop yield being calculated for each plot. We now have three groups of ten figures, and we wish to know if there are any differences between these groups. The data were recorded in the *fertilisers* dataset as shown in Table 1.1.

When these data are plotted on a graph, it appears that the fertilisers do differ in the amount of yield produced (Fig. 1.1), but there is also a lot of

Table 1.1 Raw data from the *fertilisers* dataset

Fertiliser	Yields (in tonnes) from the 10 plots allocated to that fertiliser
1	6.27, 5.36, 6.39, 4.85, 5.99, 7.14, 5.08, 4.07, 4.35, 4.95
2	3.07, 3.29, 4.04, 4.19, 3.41, 3.75, 4.87, 3.94, 6.28, 3.15
3	4.04, 3.79, 4.56, 4.55, 4.53, 3.53, 3.71, 7.00, 4.61, 4.55

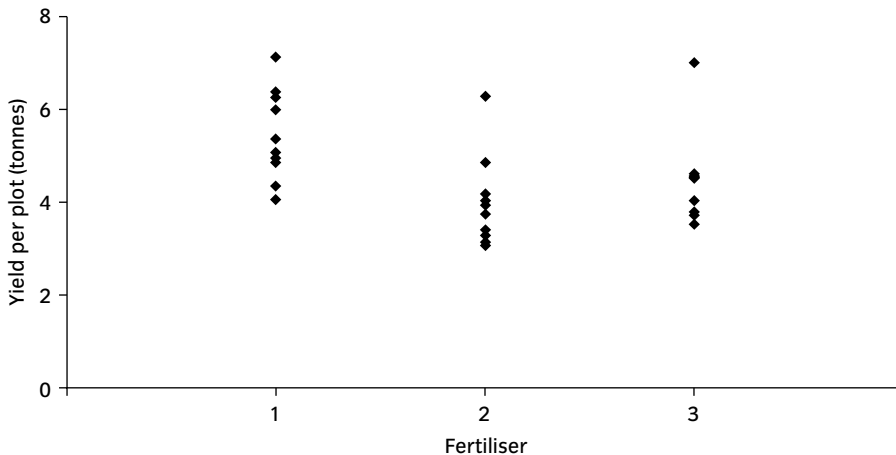


Fig. 1.1 The yield per plot for 30 plots treated with 3 fertilisers.

variation between plots given the same fertiliser. Whilst it appears that fertiliser 1 produces the highest yield on average, a number of plots treated with fertiliser 1 did actually yield less than some of the plots treated with fertilisers 2 or 3.

We now need to compare these three groups to discover if this apparent difference is statistically significant. When comparing two samples, the first step was to compute the difference between the two sample means (see revision section). However, because we have more than two samples, we do not compute the differences between the group means directly. Instead, we focus on the variability in the data. At first this seems slightly counter-intuitive: we are going to ask questions about the *means* of three groups by analysing the *variation* in the data. How does this work?

What happens when we calculate a variance?

The variability in a set of data quantifies the scatter of the data points around the mean. To calculate a variance, first the mean is calculated, then the deviation of each point from the mean. Deviations will be both positive and negative; and the sum will be zero. (This follows directly from how the mean was calculated in the first place). This will be true regardless of the size of the dataset, or amount of variability within a dataset, and so the raw deviations are not useful as a measure of variability. If the deviations are squared before summation then this sum is a useful measure of variability, which will increase the greater the scatter of the data points around the mean. This quantity is referred to as a **sum of squares** (SS), and is central to our analysis. The fertiliser dataset is illustrated in Fig. 1.2, along with the mean, and the deviation of each point from the mean. At this point, the fertiliser applied to each plot is not indicated.

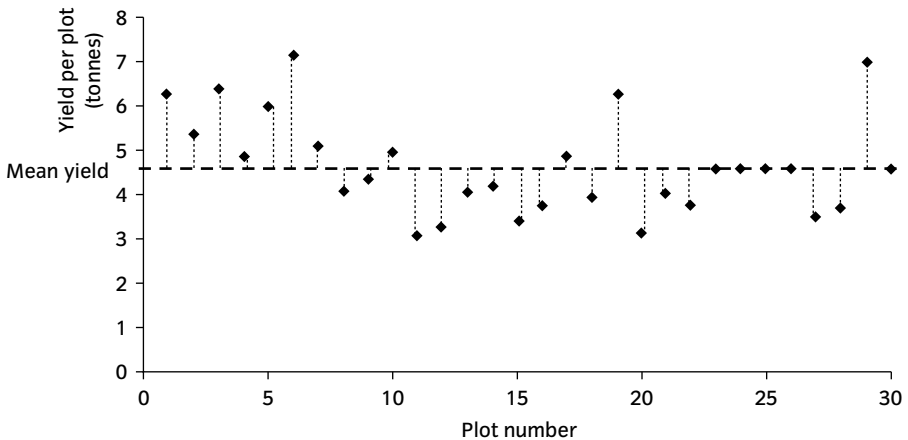


Fig. 1.2 Yield per plot by plot number.

The SS however cannot be used as a comparative measure between groups, because clearly it will be influenced by the number of data points in the group; the more data points, the greater the SS. Instead, this quantity is converted to a variance by dividing by $n - 1$, where n equals the number of data points in the group. A variance is therefore a measure of variability, taking account of the size of the dataset.

Why use $n - 1$ rather than n ?

If we wish to calculate the average squared deviation from the mean (i.e. the variance) why not divide by n ? The reason is that we do not actually have n independent pieces of information about the variance. The first step was to calculate a mean (from the n independent pieces of data collected). The second step is to calculate a variance with reference to that mean. If $n - 1$ deviations are calculated, it is known what the final deviation must be, for they must all add up to zero by definition. So we have only $n - 1$ independent pieces of information on the variability about the mean. Consequently, you can see that it makes more sense to divide the SS by $n - 1$ than n to obtain an average squared deviation around the mean. (A fuller explanation of this is given in Appendix 2). The number of independent pieces of information contributing to a statistic are referred to as the **degrees of freedom**.

Partitioning the variability

In an ANOVA, it is useful to keep the measure of variability in its two components; that is, a sum of squares, and the degrees of freedom associated with the sum of squares. Returning to the original question: what is causing the variation in yield between the 30 plots of the experiment? Numerous factors are likely to be involved: e.g. differences in soil nutrients between the plots,

differences in moisture content, many other biotic and abiotic factors, and also the fertiliser applied to the plot. It is only the last of these that we are interested in, so we will divide the variability between plots into two parts: that due to applying different fertilisers, and that due to all the other factors. To illustrate the principle behind partitioning the variability, first consider two extreme datasets. If there was almost no variation between the plots due to any of the other factors, and nearly all variation was due to the application of the three fertilisers, then the data would follow the pattern of Fig. 1.3a. The first step would be to calculate a grand mean, and there is considerable variation around this mean. The second step is to calculate the three group means that we wish to compare: that is, the means for the plots given fertilisers A, B and C. It can be seen that once these means are fitted, then little variation is left around the group means (Fig. 1.3b). In other words, fitting the group means has removed

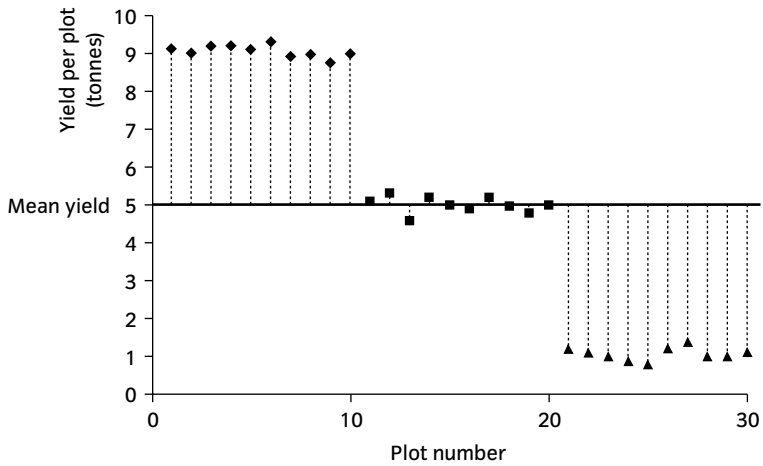


Fig. 1.3(a) Variability around the grand mean for fictitious dataset 1.

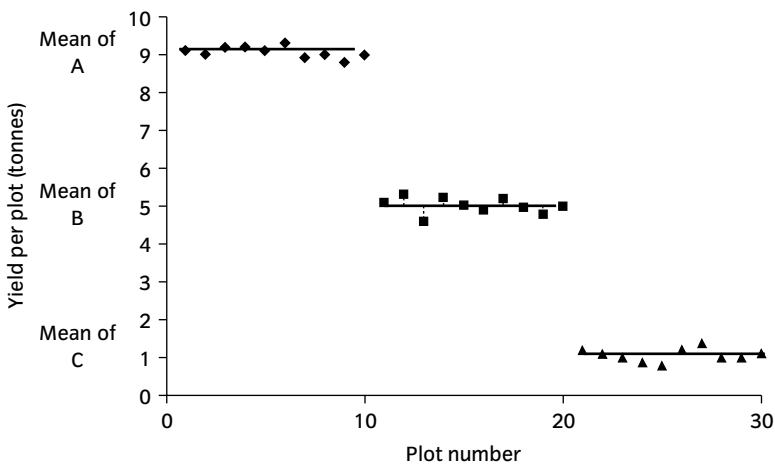


Fig. 1.3(b) Variability around three treatment means for fictitious dataset 1.

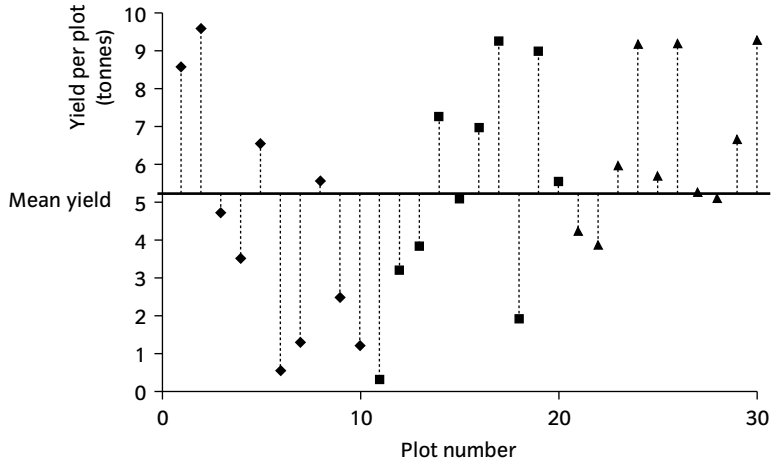


Fig. 1.4(a) Variability around the grand mean for fictitious dataset 2.

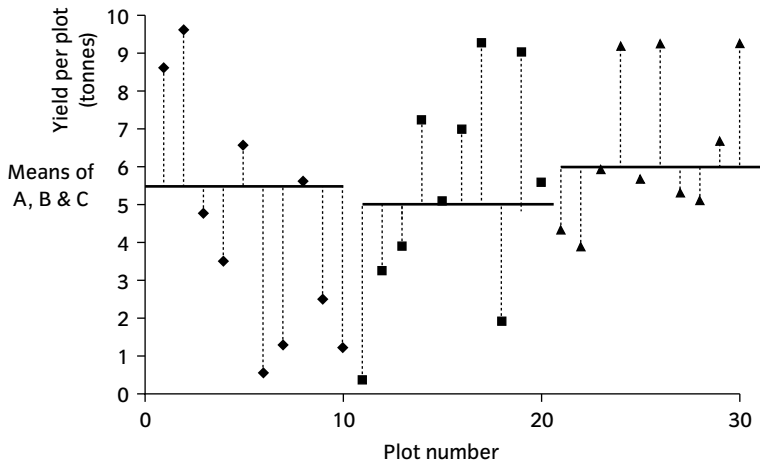


Fig. 1.4(b) Variability around three treatment means for fictitious dataset 2.

or **explained** nearly all the **variability** in the data. This has happened because the three means are distinct.

Now consider the other extreme, in which the three fertilisers are, in fact, identical. Once again, the first step is to fit a grand mean and calculate the sum of squares. Second, three group means are fitted, only to find that there is almost as much variability as before. Little variability has been explained. This has happened because the three means are relatively close to each other (compared to the scatter of the data).

The amount of variability that has been explained can be quantified directly by measuring the scatter of the treatment means around the grand mean. In the first of the two examples, the deviations of the group means around the grand mean are considerable (Fig. 1.5a), whereas in the second example these deviations are relatively small (Fig. 1.5b).

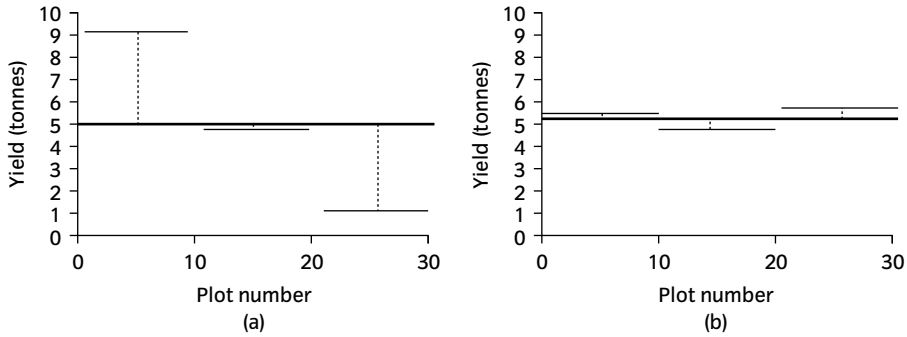


Fig. 1.5 Deviations of group means around the grand mean.

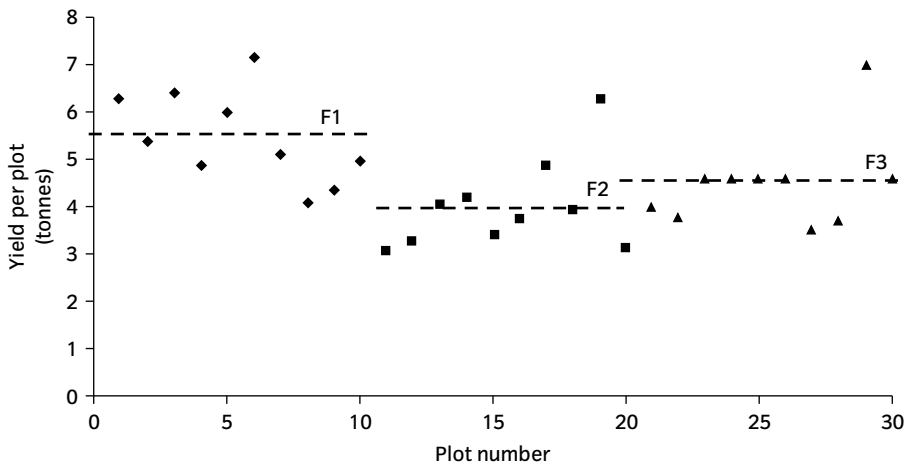


Fig. 1.6 Variability in yield around three means for the *fertiliser* dataset.

The dataset given in Table 1.1 represents an intermediate situation in which it is not immediately obvious if the fertilisers have had an influence on yield. When the three group means are fitted, there is an obvious reduction in variability around the three means (compared to the one mean) (Fig. 1.6). But at what point do we decide that the amount of variation explained by fitting the three means is **significant**? The word **significant**, in this context, actually has a technical meaning. It means ‘When is the variability between the group means greater than that we would expect by chance alone?’

At this point it is useful to define the three measures of variability that have been referred to. These are:

SSY = Total sum of squares.

Sum of squares of the deviations of the data around the grand mean.

This is a measure of the total variability in the dataset.

SSE = Error sum of squares.

Sum of squares of the deviations of the data around the three separate group means.

This is a measure of the variation between plots that have been given the same fertiliser.

SSF = Fertiliser sum of squares.

Sum of squares of the deviations of the group means from the grand mean.

This is a measure of the variation between plots given different fertilisers.

Variability is measured in terms of sums of squares rather than variances because these three quantities have the simple relationship:

$$SSY = SSF + SSE.$$

So the total variability has been divided into two components; that due to differences between plots given different treatments, and that due to differences between plots given the same treatment. Variability must be due to one or other of these two causes. Separating the total SS into its component SS is referred to as **partitioning the sums of squares**.

A comparison of SSF and SSE is going to indicate whether fitting the three fertiliser means accounts for a significant amount of variability in the data. For example, looking back at Fig. 1.3b, SSE was very small in this instance, and SSF large. In contrast in Fig. 1.4b, SSE was large, and SSF fairly small. Comparing the two raw figures however would not be useful, as the size of an SS is always related to the number of data points used to calculate it. In this instance, the greater the number of means fitted to the data, the greater SSF would be, because more variance would have been explained. Taken to the limit, if our aim was merely to maximise SSF, we should fit a mean for every data point, because in that way we could explain all the variability! For a valid comparison between these two sources of variability, we need to compare the variability per degree of freedom, i.e. the variances.

Partitioning the degrees of freedom

Every SS was calculated using a number of independent pieces of information. The first step in any analysis of variance is to calculate SSY. It has already been discussed that when looking at the deviations of data around a central grand mean, there are $n - 1$ independent deviations: i.e. in this case $n - 1 = 29$ degrees of freedom (df). The second step is to calculate the three treatment means. When the deviations of two of these treatment means from the grand mean have been calculated, the third is predetermined, as again by definition, the three deviations must sum to zero. Therefore, SSF, which measures the extent to which the group means deviate from the grand mean, has two df associated with it. Finally, SSE measures variation around the three group means. Within each of these groups, the ten deviations must sum to zero. Given nine deviations within the group, the last is predetermined. Thus SSE has $3 \times 9 = n - 3 = 27$ df associated with it. Just as the SS are additive, so are the df.

Mean squares

Combining the information on SS and df, we can arrive at a measure of variability per df. This is equivalent to a variance, and in the context of ANOVA is called a **mean square (MS)**. In summary:

Fertiliser Mean Square (FMS)	= SSF/2	The variation (per df) between plots given different fertilisers.
Error Mean Square (EMS)	= SSE/27	The variation (per df) between plots given the same fertiliser.
Total Mean Square (TMS)	= SSY/29	The total variance of the dataset.

Unlike the SS, the MS are not additive.

So now the variability per df due to differences between the fertilisers has been partitioned from the variability we would expect due to all other factors. Now we are in the position to ask: by fitting the treatment means, have we explained a significant amount of variance?

F-ratios

If none of the fertilisers influenced yield, then the variation between plots treated with the same fertiliser would be much the same as the variation between plots given different fertilisers. This can be expressed in terms of mean squares: the mean square for fertiliser would be the same as the mean square for error: i.e.

$$\frac{\text{FMS}}{\text{EMS}} = 1.$$

The ratio of these two mean squares is the **F-ratio**, and is the end result of the ANOVA. Even if the fertilisers are identical, it is unlikely to equal exactly 1, it could by chance take a whole range of values. The **F distribution** represents the range and likelihood of all possible F-ratios under the null hypothesis (i.e. when the fertilisers are identical), as illustrated in Fig. 1.7.

If the three fertilisers were very different, then the FMS would be greater than the EMS, and the F-ratio would be greater than 1. However, Fig. 1.7 illustrates that the F-ratio can be quite large even when there are no treatment differences. At what point do we decide that the size of the F-ratio is due to treatment differences rather than chance?

Just as with other test statistics, the traditional threshold probability of making a mistake is 0.05. In other words, we accept that the F-ratio is significantly greater than 1 if it will be that large or larger under the null hypothesis only 5% of the time. If we had inside knowledge that the null hypothesis was in fact true, then 5% of the time we would still get an F-ratio that large. When we conduct an experiment however we have no such inside knowledge, and

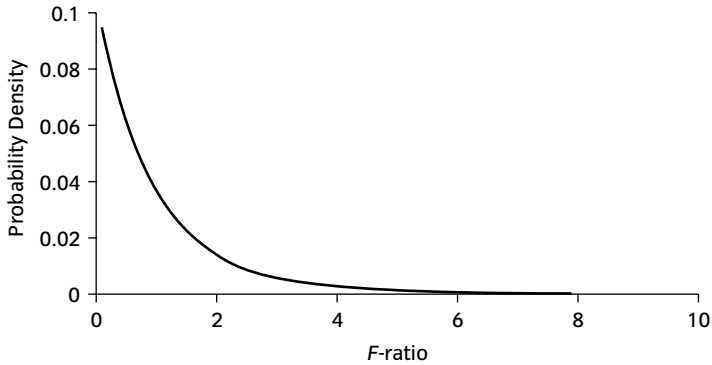


Fig. 1.7 The F distribution for 2 and 27 degrees of freedom (illustrates the probability of a F -ratio of different sizes when there are no treatment differences).

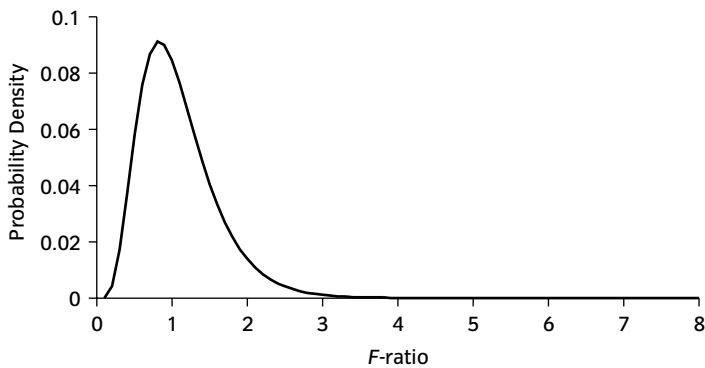


Fig. 1.8 The F distribution for 10 and 57 degrees of freedom.

we are trying to gather evidence against it. Our p -value is a measure of the strength of evidence against the null hypothesis. Only when it is less than 0.05 do we consider the evidence great enough to accept. A fuller discussion of the meaning of the p -value is given in Appendix 1.

It should be mentioned that the exact F distribution will depend upon the df with which the F -ratio was constructed. In this case, the df are 2 and 27, associated with the numerator and the denominator of the F -ratio respectively. The general shape will vary from a decreasing curve (Fig. 1.7) to a humped distribution, skew to the right (Fig. 1.8). When doing an ANOVA table in most packages the F -ratio, degrees of freedom and the p -value are provided in the output, or occasionally you are left to look up the F -ratio in statistical tables.

1.4 An example of ANOVA

Having explained the principles behind an analysis of variance, this section will provide an example of a one-way ANOVA. This requires two pieces of input from you.

Step 1: The data

The first point is to represent the two variables in a form that a statistical program will understand. To do this, the data should be converted from Table 1.1 to the ‘samples and subscripts’ form shown in Table 1.2. It can be seen here that FERTIL is represented by the subscripts 1, 2 and 3 which correspond to the

Table 1.2 Data presented as samples and subscripts

FERTIL	YIELD (tonnes)
1	6.27
1	5.36
1	6.39
1	4.85
1	5.99
1	7.14
1	5.08
1	4.07
1	4.35
1	4.95
2	3.07
2	3.29
2	4.04
2	4.19
2	3.41
2	3.75
2	4.87
2	3.94
2	6.28
2	3.15
3	4.04
3	3.79
3	4.56
3	4.55
3	4.55
3	4.53
3	3.53
3	3.71
3	7.00
3	4.61

three different fertilisers. This variable is categorical, and in this sense the values 1, 2 and 3 are arbitrary. In contrast, `YIELD` is continuous, the values representing true measurements. Data are usually continuous, whilst explanatory variables may be continuous (see Chapter 2) or categorical (this chapter) or both (later chapters).

Step 2: The question

This is the first use of model formulae—a form of language that will prove to be extremely useful. The question we wish to ask is: ‘Does fertiliser affect yield?’.

This can be converted to the **word equation**

$$\text{YIELD} = \text{FERTIL.}$$

This equation contains two variables: `YIELD`, the data we wish to explain and `FERTIL`, the variable we hypothesise might do the explaining.

`YIELD` is therefore the **response** (or dependent) **variable**, and `FERTIL` the **explanatory** (or independent) **variable**. It is important that the data variable is on the left hand side of the formula, and the explanatory variable on the right hand side. It is the right hand side of the equation that will become more complicated as we seek progressively more sophisticated explanations of our data. Having entered the data into a worksheet in the correct format, and decided on the appropriate model formula and analysis, the specific command required to execute the analysis will depend upon your package (see package specific supplements). The output is presented here in a generalised format.

BOX 1.1 Analysis of variance with one explanatory variable

Word equation: `YIELD = FERTIL`

`FERTIL` is categorical

One-way analysis of variance for `YIELD`

Source	DF	SS	MS	F	P
<code>FERTIL</code>	2	10.8227	5.4114	5.70	0.009
Error	27	25.6221	0.9490		
Total	29	36.4449			

Output

The primary piece of output is the ANOVA table, in which the partitioning of SS and df has taken place. This will either be displayed directly, or can be constructed by you with the output given. The total SS have been partitioned between treatment (`FERTIL`) and error, with a parallel partitioning of degrees of freedom. Each of the columns ends with the total of the preceding terms.

The calculation of the SS is displayed in Table 1.3. Columns *M*, *F* and *Y* give the grand mean, the fertiliser mean and the plot yield for each plot in turn.

Table 1.3 Calculating the SS and the DF

Datapoint	FERTIL	<i>M</i>	<i>F</i>	<i>Y</i>	<i>MY</i>	<i>MF</i>	<i>FY</i>
1	1	4.64	5.45	6.27	1.63	0.80	0.82
2	1	4.64	5.45	5.36	0.72	0.80	-0.09
3	1	4.64	5.45	6.39	1.75	0.80	0.94
4	1	4.64	5.45	4.85	0.21	0.80	-0.60
5	1	4.64	5.45	5.99	1.35	0.80	0.54
6	1	4.64	5.45	7.14	2.50	0.80	1.69
7	1	4.64	5.45	5.08	0.44	0.80	-0.37
8	1	4.64	5.45	4.07	-0.57	0.80	-1.38
9	1	4.64	5.45	4.35	-0.29	0.80	-1.10
10	1	4.64	5.45	4.95	0.31	0.80	-0.50
11	2	4.64	4.00	3.07	-1.57	-0.64	-0.93
12	2	4.64	4.00	3.29	-1.35	-0.64	-0.71
13	2	4.64	4.00	4.04	-0.60	-0.64	0.04
14	2	4.64	4.00	4.19	-0.45	-0.64	0.19
15	2	4.64	4.00	3.41	-1.23	-0.64	-0.59
16	2	4.64	4.00	3.75	-0.89	-0.64	-0.25
17	2	4.64	4.00	4.87	0.23	-0.64	0.87
18	2	4.64	4.00	3.94	-0.70	-0.64	-0.06
19	2	4.64	4.00	6.28	1.64	-0.64	2.28
20	2	4.64	4.00	3.15	-1.49	-0.64	-0.85
21	3	4.64	4.49	4.04	-0.60	-0.16	-0.45
22	3	4.64	4.49	3.79	-0.85	-0.16	-0.70
23	3	4.64	4.49	4.56	-0.08	-0.16	0.07
24	3	4.64	4.49	4.55	-0.09	-0.16	0.06
25	3	4.64	4.49	4.55	-0.09	-0.16	0.06
26	3	4.64	4.49	4.53	-0.11	-0.16	0.04
27	3	4.64	4.49	3.53	-1.11	-0.16	-0.96
28	3	4.64	4.49	3.71	-0.93	-0.16	-0.78
29	3	4.64	4.49	7.00	2.36	-0.16	2.51
30	3	4.64	4.49	4.61	-0.03	-0.16	0.12
DF		1	3	30	29	2	27
SS					36.44	10.82	25.62

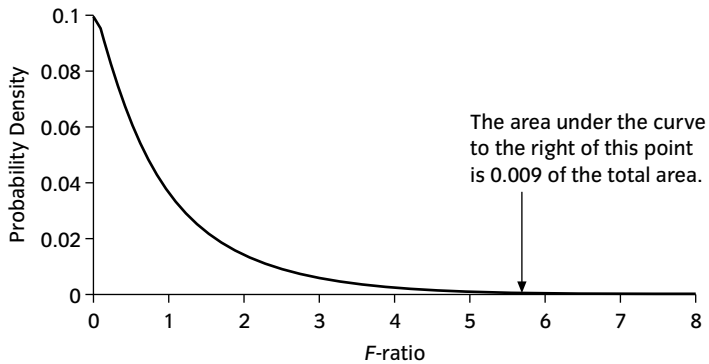


Fig. 1.9 The F distribution of 2 and 27 df. The area to the right of 5.7 represents the probability that the F -ratio is at least 5.7, and is 0.009 of the total area under the curve.

Column MY then represents the deviations from the grand mean for each plot. If these values are squared and summed, then the result is the total SS of 36.44. FY then represents the deviations from the group mean for each plot; these values squared and summed give the error SS.

Finally, MF represents the deviations of the fertiliser means from the grand mean; squaring and summing giving the treatment SS. Dividing by the corresponding df gives the mean square. Comparison of the two mean squares gives the F -ratio of 5.70. The probability of getting an F -ratio as large as 5.70 or larger, if the null hypothesis is true, is the p -value of 0.009. That is sufficiently small to conclude that these fertilisers probably do differ in efficacy.

Presenting the results

Having concluded that there is a significant difference between the fertilisers, it would be interesting to know where this difference lies. One useful way of displaying the results would be to tabulate the means for each group, and their 95% confidence intervals. What do we mean by a confidence interval, and how are they constructed?

To answer this, we need to draw together the basic principles reviewed in Revision Section 1, and apply them in the context of ANOVA. A confidence interval is an expression of how confident we are in our estimates (in this case, the three group means). For each confidence interval, we would expect the true mean for that group to lie within that range 95% of the time.

To construct a confidence interval, both the parameter estimate, and the variability in that estimate are required. In this case, the parameters estimated are means—we wish to know the true mean yield to be expected when we apply fertiliser 1, 2 or 3—which we will denote μ_A , μ_B , and μ_C respectively. These represent true population means, and as such we cannot know their exact values—but our three treatment means represent estimates of these three parameters. The reason why these estimates are not exact is because of the

Table 1.4 Constructing confidence intervals

Fertiliser	\bar{y}	t_{crit} with 27 df for 95% confidence	$\frac{s}{\sqrt{n}}$	Confidence interval
1	5.445	2.0518	0.3081	(4.81, 6.08)
2	3.999	2.0518	0.3081	(3.37, 4.63)
3	4.487	2.0518	0.3081	(3.85, 5.12)

unexplained variation in the experiment, as quantified by the **error variance** which we previously met as the error mean square, and will refer to as s^2 . From Revision Section 1, the 95% confidence interval for a population mean is:

$$\bar{y} \pm t_{\text{crit}} \frac{s}{\sqrt{n}}.$$

The key point is where our value for s comes from. If we had only the one fertiliser, then all information on population variance would come from that one group, and s would be the standard deviation for that group. In this instance however there are three groups, and the unexplained variation has been partitioned as the error mean square. This is using all information from all three groups to provide an estimate of unexplained variation—and the degrees of freedom associated with this estimate are 27—much greater than the 9 which would be associated with the standard deviation of any one treatment. So the value of s used is $\sqrt{\text{EMS}} = \sqrt{0.949} = 0.974$. This is also called the pooled standard deviation. Hence the 95% confidence intervals are as shown in Table 1.4.

These intervals, combined with the group means, are an informative way of presenting the results of this analysis, because they give an indication of how accurate the estimates are likely to be.

It is worth noting that we have assumed it is valid to take one estimate of s and apply it to all fertiliser groups. However, consider the following scenario. Fertiliser 1 adds nitrate, while Fertiliser 2 adds phosphate (and Fertiliser 3 something else altogether). The plots vary considerably in nitrate levels, and Fertiliser 1 is sufficiently strong to bring all plots up to a level where nitrate is no longer limiting. So Fertiliser 1 reduces plot-to-plot variation due to nitrate levels. The phosphate added by Fertiliser 2 combines multiplicatively with nitrate levels, so increasing the variability arising from nitrate levels. The mean yields from plots allocated to Fertiliser 2 would be very much more variable, while those allocated to Fertiliser 1 would have reduced variability, and our assumption of equal variability between plots within treatments would be incorrect. The 95% confidence interval for Fertiliser 2 will have been underestimated.

Fortunately in this case the group standard deviations do not look very different (Table 1.5), so it is unlikely that we have a problem. In Chapter 9 we shall discuss this and other assumptions we make in doing these analyses.

Table 1.5 Descriptive statistics for YIELD by FERTIL

Descriptive Statistics for YIELD by FERTIL			
FERTIL	N	Mean	Standard Deviation
1	10	5.445	0.976
2	10	3.999	0.972
3	10	4.487	0.975

1.5 The geometrical approach for an ANOVA

The analysis that has just been conducted can be represented as a simple geometrical picture. One advantage of doing this is that such pictures can be used to illustrate certain concepts. In this first illustration, geometry can be used to represent the partitioning and additivity of the SS.

The geometrical approach is actually a two-dimensional representation of multidimensional space. One dimension is represented by the position of a point on a line—one coordinate can be used to define that position. Two dimensions may be pictured as a graph, with a point being specified by two coordinates. This can be extended to three dimensions, in which the position of a point in a cube is specified by three coordinates. Beyond three dimensions it is no longer possible to visualise a geometrical picture to represent all dimensions simultaneously. It is possible however to take a slice through multidimensional space and represent it in two dimensions. For example, if a cube has axes x , y , and z , the position of three points can be specified by their x , y and z coordinates. A plane could then be drawn through those three points, so allowing them to be represented on a piece of paper (Fig. 1.10). There are still three coordinates associated with each point (and so defining that point), but for visual purposes, the three dimensions have been reduced to two. In fact, it is possible to do this for any three points, however many dimensions they are plotted in. This trick is employed by the geometrical approach.

In this case, there are as many dimensions as there are data points in the dataset (30). Each point is therefore represented by 30 coordinates. The three points themselves are the columns 3, 4 and 5 (M , F and Y) of Table 1.3.

Point Y

This point represents the data, so the 30 coordinates describing this point are the 30 measurements of yield.

Point M

This point represents the grand mean. Because we are dealing with 30-dimensional space (as dictated by the size of the dataset), this point also has 30 coordinates

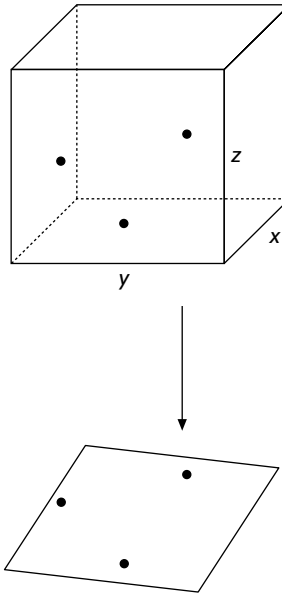


Fig. 1.10 Representing three dimensions in two dimensions.

specifying its position in multidimensional space. However, the values of these 30 coordinates are all the same (the grand mean).

Point F

This point represents the treatment means. While still 30 elements long, the first ten elements are the mean for treatment 1 (and are therefore the same value), the second ten the mean for treatment two etc. Therefore the first part of the geometrical approach is that the three variables, M , F and Y , are represented as points. These three points may be joined to form a triangle in two dimensional space as follows:

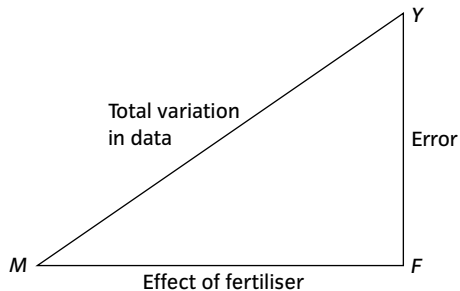


Fig. 1.11 The geometrical approach—variables represented as points, sources as vectors.

The triangle has been drawn with F at a right angle. There is a reason for this which will be explained more fully in Chapter 2. The lines joining the points are vectors, and these represent sources of variability. For example, the vector MY represents the variability of the data (Y) around the grand mean (M). In

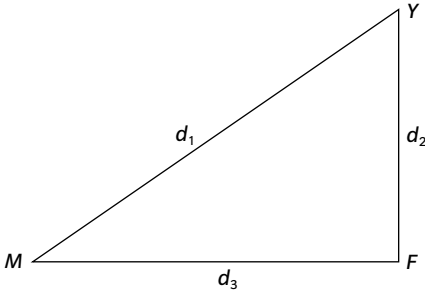


Fig. 1.12 The geometrical approach—Pythagoras theorem.

the same way that a vector can be decomposed into two components, so can the variability be partitioned into (i) *FY*—the variability of the data around their group means, and (ii) *MF*—the variability of the group means around the grand mean. The implication here is that sources of variability are **additive**. While this assumption is crucial in our approach, it is not necessarily true. Testing and correcting for this assumption are covered later (Chapters 9 and 10).

The third part of the geometrical approach relies on the fact that the triangle is right-angled. The **squared length of each vector** is then equivalent to the SS for that source. This is illustrated in Fig. 1.12.

Pythagoras states that:

$$d_1^2 = d_2^2 + d_3^2.$$

This is equivalent to:

$$SSY = SSF + SSE.$$

This illustrates geometrically the partitioning of sums of squares. It is precisely because the sums of squares can be summed in this way that they are central to the analysis of variance. Other ways of measuring variation (e.g. using variances) would not allow this, because the variances of the components would not add up to the variance of the whole.

The shape of the triangle can then provide information on the relative sizes of these different components. The triangle in Fig. 1.13a suggests the effect of fertiliser is weak, whereas Fig. 1.13b suggests that the effect is strong.

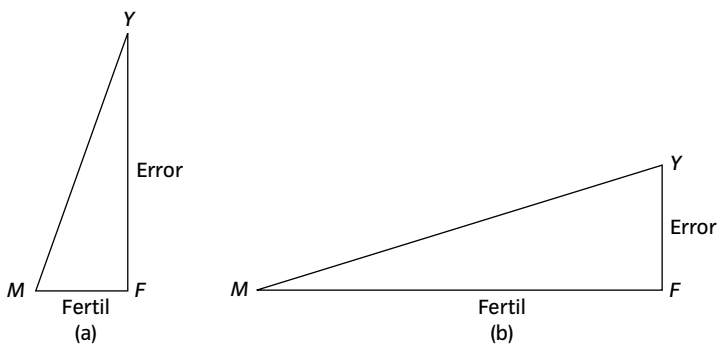


Fig. 1.13 (a) Impact of fertiliser on yield is weak; (b) Impact of fertiliser on yield is strong.

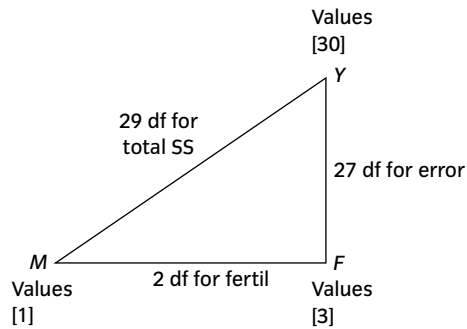


Fig. 1.14 The geometrical approach—partitioning the degrees of freedom.

It is also possible to represent the parallel partitioning of degrees of freedom in a similar manner. At each apex (Fig. 1.14), are the number of values in each variable (30 different data points, 3 treatment means, and 1 grand mean). The difference between these values gives the number of degrees of freedom associated with moving from one point to another. For example, calculating a grand mean is equivalent to moving from Y to M , and in doing so, twenty nine degrees of freedom are lost. Moving from M to F is equivalent to exchanging one grand mean for three treatment means, the difference being two degrees of freedom. These degrees of freedom are associated with the corresponding vectors and therefore with the sources represented by these vectors.

Figure 1.14 also illustrates the additivity of degrees of freedom. In moving from $M[1]$ to $Y[30]$, the total df are 29. This is true if we go directly (via vector MY), or indirectly (via MF with 2 df then FY with 27 df).

1.6 Summary

In this chapter an analysis of variance has been followed from first principles. A number of concepts have been discussed, including:

- Model formulae—a ‘word equation’ which encapsulates the question being asked.
- The fundamental principle behind an ANOVA—partitioning variability in data to ask questions about differences between groups.
- Degrees of freedom—the number of independent pieces of information which contribute to a particular measure.
- Three parallel partitions: sources of variability, sums of squares, degrees of freedom.
- The meaning of a p -value.
- Presenting the results of ANOVA in the form of confidence intervals.
- The geometrical approach for ANOVA, see Table 1.6.

Table 1.6 Geometric approach for ANOVA

Statistics		Geometry
Variable	↔	Point
Source	↔	Vector
Sums of squares	↔	Squared length of vector
DF	↔	The number of dimensions of freedom gained by moving from the variable at one end of the vector to the variable at the other

1.7 Exercises

Melons

An experiment was performed to compare four melon varieties. It was designed so that each variety was grown in six plots—but two plots growing variety 3 were accidentally destroyed. The data are plotted in Fig. 1.15, and can be found in the *melons* dataset under the variables `YIELDM` and `VARIETY`.

Table 1.7 shows some summary statistics and an ANOVA table produced from these data.

- (1) What is the null hypothesis in this case?
- (2) What conclusions would you draw from the analysis in Table 1.7?
- (3) How would you summarise the information provided by the data about the amount of error variation in the experiment?
- (4) Calculate the standard error of the mean for all four varieties.
- (5) How would you summarise and present the information from this analysis?

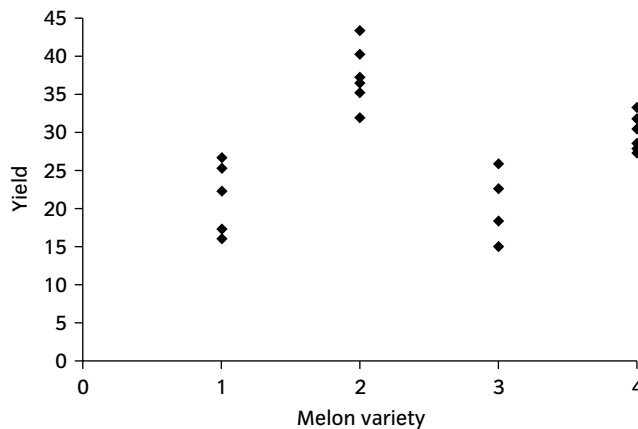
**Fig. 1.15** Melon yields.

Table 1.7 ANOVA for *melons*

VARIETY	N	Mean
1	6	20.490
2	6	37.403
3	4	20.462
4	6	29.897

One-way analysis of variance for YIELDM					
Source	DF	SS	MS	F	P
VARIETY	3	1115.3	371.8	23.80	0.000
Error	18	281.2	15.6		
Total	21	1396.5			

Dioecious trees

A plant species is dioecious if each individual produces all male flowers or all female flowers. The dataset *dioecious trees* contains data from 50 trees of one particular dioecious species, from a ten hectare area of mixed woodland. For each individual, the `SEX` was recorded (coded as 1 for male and 2 for female), the diameter at breast height in millimetres (`DBH`), and the number of flowers on the tree at the time of measurement (`FLOWERS`). This dataset will be revisited several times over the following chapters.

- (1) Test the null hypothesis that male and female trees produce the same number of flowers.
- (2) Show graphically how the number of flowers differs between the sexes.

Technical guidance on the analysis of these datasets is provided in the package specific supplements. Answers are presented at the end of this book.