The Analysis of Variation and Co-variation

'After two years Pharaoh had a dream. He thought he stood by the river out of which came up seven kine, very beautiful and fat.'

Undoubtedly one of the most elegant, powerful, and useful techniques in modern statistical method is that of the Analysis of Variation and Co-variation by which the total variation in a set of data may be reduced to components associated with possible sources of variability whose relative importance we wish to assess. The precise form which any given analysis will take is intimately connected with the structure of the investigation from which the data are obtained. A simple structure will lead to a simple analysis; a complex structure to a complex analysis. In this chapter we shall consider some of the more common types of analysis so that the reader may get hold of the basic principles and appreciate the beauty of the technique.*

It will be recalled that we calculate the variance of a set of data as the mean square deviation of the several items from their grand average. Thus, if the individual items be denoted by x, their grand average by $\bar{x}$, and the number of items by N, then the variance will be

$$V = \sigma^2 = \frac{1}{N}\Sigma(x - \bar{x})^2$$

This will be the sample variance. But we also know that a small sample tends to underestimate the variance of the parent population and that a better estimate of the population variance is obtained by dividing the 'Sum of Squares', $\Sigma(x - \bar{x})^2$ by the number of 'degrees of freedom', $(N - 1)$. We have then that the 'Population Variance Estimate' is

$$\hat{V} = \sigma^2 = \frac{\Sigma(x - \bar{x})^2}{N - 1}$$

We shall show, in a moment, by way of example, how the total variation may be resolved into components in suitable cases.

* Introduced by R. A. Fisher.

First, however, the reader should have it clearly in his mind that in the Analysis of Variance we compute for each source of variability in turn:

(a) the sum of squares, (b) the number of degrees of freedom. Consider, then, the following table of data which shows the values of 20 items which have been collected in four samples of 5 items each. Even if the data were collected at random from a perfectly homogeneous population, we should not expect each sample to have the same average value, since even sample averages must reflect the variance in the parent population. What we should expect in these circumstances is that the variation between sample averages should be commensurate with the population variance as indicated by the variation within the individual samples. If it should prove that the 'between sample variation' were significantly greater than the 'within sample variation', then we should suspect that the samples were not, in fact, drawn from the same population, but from populations whose average values differed, so that on top of the 'within population variation' there existed also a 'between population variation'.

	Sample 1	Sample 2	Sample 3	Sample 4
	2	3	6	5
	3	4	8	5
	1	3	7	5
	3	5	4	3
	1	0	10	2
	—	—	—	—
Sample totals	10	15	35	20
Sample means	2	3	7	4
	—	—	—	—

Total number of items $= N = 20$

Grand Total of all items $= T = 80$

Grand Average of all items $= \frac{T}{N} = \frac{80}{20} = 4$

The table on next page shows the squares of the deviations of the 20 items from their grand average value of 4.

The number of degrees of freedom on which this total sum of squares was computed is found as one less than the number of items on which the calculation was made. We had 20 items and so

Total Degrees of Freedom $= 19$

	Sample 1	Sample 2	Sample 3	Sample 4
	4	1	4	1
	1	0	16	1
	9	1	9	1
	1	1	0	1
	9	16	36	4
Totals	24	19	65	8

Grand Total of Squared Deviations from the Grand Average
= Total Sum of Squares $= 24 + 19 + 65 + 8 = 116$

Let us now try to partition the total sum of squares and the total degrees of freedom into components corresponding to 'between sample averages' and 'within samples' respectively. In order to get the between sample effect, we must eliminate the within sample effect. We can do this by replacing each item by its own sample average. Doing this, we obtain the following table:

Sample 1	Sample 2	Sample 3	Sample 4
2	3	7	4
2	3	7	4
2	3	7	4
2	3	7	4
2	3	7	4

For which the Grand Total is still $T = 80$, of course

In order to get the between sample sum of squares, we now proceed exactly as we did when we were calculating the total sums of squares. We set up the following table which shows the squares of the deviations of the entries in our new table from their grand average, thus:

	Sample 1	Sample 2	Sample 3	Sample 4
	4	1	9	0
	4	1	9	0
	4	1	9	0
	4	1	9	0
	4	1	9	0
Totals	20	5	45	0

Between sample sum of squares $= 20 + 5 + 45 = 70$.

To get the relevant degrees of freedom, we take one less than the number of sample averages on which the computation was based. Hence:

Between sample degrees of freedom $= 4 - 1 = 3$

It now remains for us to get the sum of squares and the degrees of freedom which correspond to within sample variation. In order to do this, we must remove the between sample average effect. We are now concerned only with the variability within the individual samples. To get this, we subtract from each item in our original table of data its own sample average. The result is shown in the following table:

	Sample 1	Sample 2	Sample 3	Sample 4
	0	0	-1	1
	1	1	1	1
	-1	0	0	1
	1	2	-3	-1
	-1	-3	3	-2
Totals	0	0	0	0

The grand average of the items in this new table is, of course, zero, and the sum of squares for the within sample source of variation is obtained by finding the sum of the squares of the deviations of the items in this table from their grand average, zero. All we have to do, then, is to square the items as they stand. The result is:

	Sample 1	Sample 2	Sample 3	Sample 4
	0	0	1	1
	1	1	1	1
	1	0	0	1
	1	4	9	1
	1	9	9	4
Totals	4	14	20	8

Within Sample Sum of Squares = 4 + 14 + 20 + 8 = 46

In order to get the within sample degrees of freedom, we argue as follows: each sample consists of five items. For each sample the number of degrees of freedom within that sample will be one less than the number of items within that sample, viz. 4. However, there are four such samples, so the total degrees of freedom within samples will be 4 × 4 = 16.

Let us now collect our results together in a Table of the Analysis of Variance.

TABLE OF ANALYSIS OF VARIANCE

Source of variation	Sums of squares	Degrees of freedom	Variance estimate
Between samples	70	3	$\frac{70}{3}=23{\cdot}3$
Within samples	46	16	$\frac{46}{16}=2{\cdot}9$
Total	116	19	

It will be seen that our procedure has neatly divided the total sum of squares and the total degrees of freedom into two independent components, which correspond to between sample and within sample variation.

Now let us think a little, and see if we can turn this device to practical account. When we divide a sum of squares by the corresponding number of degrees of freedom on which the sum of squares is based, we are estimating a variance. In our example, the Table of the Analysis of Variance shows this done for the two components of our variation. If we set up the Null Hypothesis that the between sample variation is only a reflexion of the variation of the items in the common parent population from which the items were drawn, the two Variance estimates are estimates of the same variance. What we are saying, in effect, is this: it does not matter whether we estimate the population variance on the basis of the variation between sample averages or on the basis of the variation of the items about their own sample average. Both are completely determined by the variance of the items in the common parent population. Since the two estimates are independent of each other, we shall not expect them to be identical in value. But we shall expect them not to differ more than is to be expected taking into account the number of degrees of freedom on which they are based. Now we already have a simple test for the mutual compatibility of two variance estimates, namely Snedecor's Variance Ratio Test which we dealt with in Chapter 13. If our Null Hypothesis is correct, and there is no specific between sample effect other than that introduced by the variance of the common parent population, then we should expect Snedecor's Test to yield a non-significant result.

On the face of it, judging from our Table of Analysis of Variance, there is a specific between sample effect, i.e. in addition to the between sample variation to be expected on our Null Hypothesis, there is an extra variation between the samples which is unaccounted for by the Null Hypothesis. Applying Snedecor's Test we get

$$F=\frac{23\cdot 3}{2\cdot 9}=8\cdot 1$$

For the greater variance estimate there are 3 degrees of freedom and for the lesser variance estimate there are 13 degrees of freedom.

Consulting the Table for Snedecor's F given in Chapter 13, we find that the 1% level of F is about 5·3. The 0·1% level is about 9. Our observed value of 8·1 is therefore well above the 1% level and very nearly at the 0·1% level. We conclude that the observed variance ratio is too great for the Null Hypothesis to be maintained and that there is a specific between sample variation. The implication is that, whatever we may have hoped or thought to the contrary, if we are wise we shall act on the assumption that the samples were, in fact, drawn from sources whose average values differed from each other. If for our purpose it were desirable to have the average value as large as possible, then we should do our business with the source which gave us Sample 3 for which the average value came out at 7. We should then have to remember that, although this particular sample gave an average value of 7, it might have been an optimistic-looking sample from a population with a rather lower average, and we should therefore be interested in setting up confidence limits for the mean value in the population from which the sample was drawn. This is a matter which the reader can follow up for himself along the lines laid down in Chapter 14.

Unless the reader is of a different psychological make-up from the author, he will feel that while this is a very useful device, there ought to be some quick method of arriving at the same result. This is not just slothfulness on our part. Decisions of the kind for which this technique would be useful have to be made speedily in