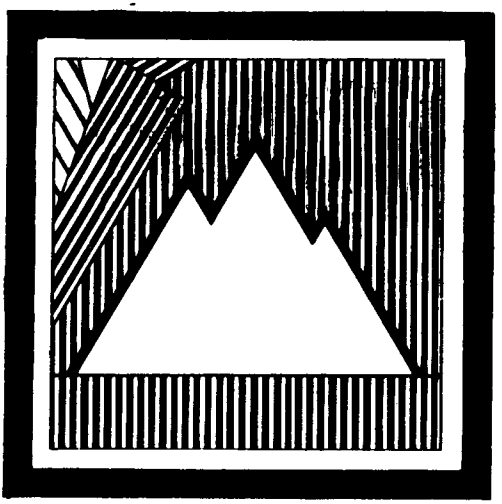
Statistical Methods in Management 2

TOM CASS

Cassell



STATISTICAL METHODS IN MANAGEMENT 2

The analysis of variance, experimental design, multiple regression analysis, time series analysis and forecasting

TOM CASS

CASSELL • LONDON

CASSELL LTD.
35 Red Lion Square, London WC1R 4SG
and at Sydney, Auckland, Toronto, Johannesburg,

an affiliate of Macmillan Publishing Co. Inc.,
New York

First published 1980

I.S.B.N. 0 304 30346 1

Printed and bound in Great Britain at
The Camelot Press Ltd, Southampton

Typeset by Colset Pte. Ltd., Singapore

PREFACE

This book assumes a knowledge of basic statistics up to the level covered in **Statistical Methods in Management 1**. Its objective is to proceed beyond that level and to provide a management-orientated text covering what are sometimes regarded as the more advanced techniques of analysis of variance and multiple regression analysis.

The approach adopted, as in Book 1, is practical rather than mathematical. A lot of attention is devoted to considering the practical implications of the assumptions underlying the techniques. Liberal use is made of examples and exercises, for all of which solutions are provided. Full benefit can only be derived by working all the exercises, since they are sometimes used to develop supplementary points.

It is a moot point whether the chapter on time series analysis belongs here or in Book 1, since it is often included in a first course on statistics. However, as the topic was not originally included in Book 1, it seems convenient and appropriate to include it here.

As Book 1 contains the basic statistical tables, it was not thought necessary to duplicate them here. Only such additional tables as are needed are included.

This book is intended primarily for managers and students on management or business studies courses, but experience with Book 1 has shown that it will be of value over a much wider range of interests.

I would like to record my gratitude for the help I have received from people at Cranfield, particularly Marjorie Dawe for her painstaking work in typing the script. I am indebted to the Biometric Society for permission to reprint Table 1 from *Critical Values for Duncan's New Multiple Range Test* by H. L. Harter (*Biometrics* **16**: 671-685, 1960).

A note on rounding

Many of the calculations in this book are quite lengthy. If intermediate values are not retained with great accuracy (in some cases 5 or 6 decimal places) substantial rounding errors may result in the later stages. However, it would be very cumbersome when writing out

the various steps in the calculations to show them all to this degree of accuracy.

Accordingly, therefore, in many cases intermediate values have been shown to only 2 or 3 decimal places, but greater accuracy than this has been retained for subsequent calculations in order to avoid rounding errors in final results. This occasionally leads to apparent slight inaccuracies if the rounded values are used, for example, 15.26 + 3.14 may be shown as 4.85 and not 4.86. The actual calculation carried out may well have been 15.25672 + 3.14484 = 4.851.

When numbers ending in 5 have been rounded, they have been taken to the nearest even number. For example, 3.125 becomes 3.12, but 4.735 would be written as 4.74.

CONTENTS

1 THE ANALYSIS OF VARIANCE — ONE-WAY CLASSIFICATION

A knitting wool manufacturer produces a range of different coloured wools in 4 different factories. Some of the more popular colours are produced in all 4 factories, and the manufacturer is concerned about maintaining uniformity of shade for each colour in the range. Inevitably, there will be slight differences from batch to batch of a particular colour, but the manufacturer suspects that there are systematic differences between the factories.

He has obtained a random sample of 5 balls of a particular colour from each factory, and had the depth of colour assessed against a standard scale, with the results shown in Table 1.1. Are there significant differences between the 4 sample means?

TABLE 1.1

Sample			
1	*2*	*3*	*4*
42	53	48	40
47	55	52	44
42	51	45	36
39	51	46	37
46	56	51	43

Invalidity of t-test

If there were only 2 samples, a comparison of the 2 averages would be made by means of a *t*-test (see page 95 of Book 1). However, the *t*-test is not valid when differences between more than 2 means are being investigated. This is because the *t*-test is based on the assumption that the 2 sample means are selected purely at random. If we have more than two means, the highest and lowest of the set will not be a random pairing. Indeed, the difference between the two extreme values will show a tendency to increase as the number of means in the set increases.

Look, for example, at the data in Table 1.2. They consist of two random samples from the same population.

TABLE 1.2

Sample 1	*Sample 2*
5.6	6.3
4.3	3.5
6.7	6.4
4.4	5.9
3.9	4.3

To assess whether or not the two means differ significantly, we carry out a standard *t*-test.

$$\bar{x}_1 = 4.98 \quad s_1 = 1.15 \quad n_1 = 5$$
$$\bar{x}_2 = 5.28 \quad s_2 = 1.30 \quad n_2 = 5$$

$$\text{Pooled standard deviation(s)} = \sqrt{\left[\frac{(4 \times 1.15^2) + (4 \times 1.30^2)}{8}\right]}$$
$$= 1.23$$
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$= \frac{4.98 - 5.28}{1.23\sqrt{(\frac{1}{5} + \frac{1}{5})}}$$
$$= \frac{-0.3}{0.78}$$
$$= -0.38$$

This is clearly not significant.

However, now consider the 6 samples in Table 1.3. They are again all random samples from the same population, but the difference between the two extreme values is large enough to show a spurious significant difference on a *t*-test.
The two extreme estimates of the mean are 5.58 and 4.00 coming from Samples 4 and 6. The *t*-test for these two samples gives:

$$t = \frac{5.58 - 4.00}{0.67\sqrt{(\frac{1}{5} + \frac{1}{5})}}$$
$$= \frac{1.58}{0.42}$$
$$= 3.73$$

This value of t is significant, even at the 0.01 level.

TABLE 1.3

	Sample					
	1	*2*	*3*	*4*	*5*	*6*
	5.6	4.3	4.8	5.6	6.3	2.8
	4.3	7.3	4.1	5.8	3.5	5.1
	6.7	5.8	4.8	5.0	6.4	4.2
	4.4	3.6	5.4	5.3	5.9	4.1
	3.9	4.1	2.8	6.2	4.3	3.8
MEANS	4.98	5.02	4.38	5.58	5.28	4.00
			s_4 =	0.46	s_6 =	0.83

Comparing More Than Two Means

The appropriate approach when more than 2 means are to be compared is based on a technique known as the **analysis of variance**. As its name implies, this is a procedure for splitting the total variation present in a set of data into separate components, each of which can be associated with a particular cause or factor. The technique is useful in many situations other than this multiple-means comparison one. In fact, it is probably the single most widely applicable statistical technique of all, and much of this book is devoted to various aspects of it.

To return to our wool example, if there are no significant differences in mean colour level between the 4 samples, the 20 balls of wool have come from a homogeneous population. The variance of this population would be a measure of overall apparent variability of colour. ('Apparent' because some of the variation in the sample results will almost certainly be due to errors of measurement; this point will be taken up again later.) To estimate this overall variance, we would carry out the standard calculation:

$$\frac{\Sigma(x - \bar{x})^2}{N - 1},$$

where N = total number of observations (20 in this case).

The top part of this formula, $\Sigma(x - \bar{x})^2$, is a measure of the total variation present in the data, being the sum of the squared deviations of all the items of data around the overall mean. It is this variation, known as the **total sum of squares**, which analysis of

variance procedures enable us to partition. For computational purposes, it is preferable to use an alternative way of writing the total sum of squares. It is easier to calculate $\Sigma x^2 - \frac{(\Sigma x)^2}{N}$ instead of $\Sigma(x - \bar{x})^2$. The two expressions will, of course, give identical results, and one may easily be derived algebraically from the other. However, before considering the problem from the analysis of variance viewpoint, it will be instructive to develop our required multiple-means test from another angle.

Within-sample Variance Estimates

The overall variance estimate $\frac{\Sigma(x - \bar{x})^2}{N - 1}$ has no practical meaning if there actually are differences between the mean colour levels of the wool produced at each of the 4 factories. In such a case, the output of each factory comprises a separate, distinguishable sub-population within the total population, and we must assess the variance of each sub-population separately. These within-sample estimates will not be affected by any between-samples differences which may be present.

To calculate the within-sample variances, we find the sum of squares *within* each sample, and divide by $(n - 1)$, where n is the number of items per sample. Σx_1, Σx_2 . . . denote the sum of values in sample 1, 2 . . . and $\bar{x}_1$, $\bar{x}_2$ . . . are the sample averages. Σx_1^2, Σx_2^2 . . . denote the sums of the squared sample values.

Sample 1

$$\Sigma x_1 = 42 + 47 + 42 + 39 + 46 = 216$$
$$\Sigma x_1^2 = 42^2 + 47^2 + 42^2 + 39^2 + 46^2 = 9374$$

The sum of squares within Sample 1 may now be calculated from the formula

$$\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n}$$
$$9374 - \frac{(216)^2}{5} = 42.8$$

As there are 5 values in the sample, we now divide this by $(5 - 1) = 4$ degrees of freedom to obtain:

$$\text{Variance estimate} = \frac{42.8}{4} = 10.7$$

We now carry out the same procedure for each of the other samples.

Sample 2

$$\begin{aligned} \Sigma x_2 &= 266 \\ \Sigma x_2^2 &= 14\,172 \\ n &= 5 \end{aligned}$$

$$\text{Sum of squares within Sample 2} = 14\,172 - \frac{(266)^2}{5}$$

$$= 20.8$$

$$\text{Variance estimate} = \frac{20.8}{4} = 5.2$$

Sample 3

$$\begin{aligned} \Sigma x_3 &= 242 \\ \Sigma x_3^2 &= 11\,750 \\ n &= 5 \end{aligned}$$

$$\text{Sum of squares within Sample 3} = 11\,750 - \frac{(242)^2}{5}$$

$$= 37.2$$

$$\text{Variance estimate} = \frac{37.2}{4} = 9.3$$

Sample 4

$$\begin{aligned} \Sigma x_4 &= 200 \\ \Sigma x_4^2 &= 8050 \\ n &= 5 \end{aligned}$$

$$\text{Sum of squares within Sample 4} = 8050 - \frac{(200)^2}{5}$$

$$= 50$$

$$\text{Variance estimate} = \frac{50}{4} = 12.5$$

Combined Variance Estimate

Each of these variance estimates is based on only 4 degrees of freedom, but if it is reasonable to assume that variability is the same within each factory, we can pool the 4 estimates together to obtain a better one based on 16 degrees of freedom. This assumption of

equality of within-sample variances is always necessary in the analysis of variance. The implications of this assumption are discussed in more detail later. It will be recalled that a similar assumption is necessary in the *t*-test.

To combine variances, we take a weighted average, using the number of degrees of freedom of each variance as the weights. This is achieved in this case by adding all the within-sample sums of squares and dividing by the total degrees of freedom.

Combined within-samples sum of squares	$= 42.8 + 20.8 + 37.2 + 50$
	$= 150.8$
Combined degrees of freedom	$= 4 + 4 + 4 + 4$
	$= 16$
Combined variance estimate	$= \frac{150.8}{16}$
	$= 9.43$

Between-samples Variance Estimate

The calculation we have just carried out is valid, whether or not there are significant differences between the sample averages. This is because the method of calculation has eliminated the between-samples differences. However, if there are no significant between-samples differences (i.e. all the data do indeed come from one homogeneous population), we could get an alternative estimate of the overall variance by using sample totals. The 4 sample totals are as follows:

Sample			
1	*2*	*3*	*4*
216	266	242	200

The variance of sample totals can now be calculated by first obtaining the between-sample-totals sum of squares as follows:

$$(216^2 + 266^2 + 242^2 + 200^2) - \frac{(924)^2}{4} = 2532$$

There are 4 sample totals, so the appropriate number of degrees of freedom will be 3. Dividing the between-sample-totals sum of

squares by 3 will thus produce the 'between-sample-totals' variance.

$$\frac{2532}{3} = 844$$

However, each sample total is the sum of 5 individual observations. This variance must, therefore, be divided by 5 in order to estimate the variance of the population of individual balls of wool. Finally, therefore, we have as our between-samples variance estimate:

$$\frac{844}{5} = 168.8$$

Comparison of the Two Estimates

We now have two estimates of the variance of the population from which the balls of wool were drawn (let us call this variance σ^2.)

The first of the estimates is based on differences *within* each sample and is unaffected by any differences which may exist between samples. The second estimate, however, is based on differences *between* samples. It will thus be a true estimate of σ^2 only if there are no significant differences between the samples, i.e. if the overall population from which the samples have been drawn really is homogeneous. If there are significant differences between the sample means, then these differences will be reflected in the estimate of σ^2. More specifically, the between-samples variance will overestimate σ^2. This suggests a way to test whether or not there are significant differences between the sample means. Under the null hypothesis of no significant difference, both these estimates are of the same σ^2. This hypothesis may be tested by comparing the two variance estimates, using an *F*-test (see page 101 of Book 1). If the second estimate is significantly greater than the first, the null hypothesis may be rejected. Because of the nature of the situation, a one-tail test will always be appropriate.

Another assumption is implicit if we use the *F*-test; that the populations from which the samples have been drawn are normally distributed. However, the *F*-test is fairly insensitive to departures from normality, so long as the errors are randomly distributed.

Within-samples estimate, based on 16 d.f. = 9.43
Between-samples estimate, based on 3 d.f. = 168.8

$$F = \frac{168.8}{9.43} = 17.90$$

$$F_{.01;3,16} = 5.29$$

The null hypothesis may thus be clearly rejected. We may conclude that there are significant differences between the 4 sample means.

Computational Procedure

The above calculations have been laid out in a detailed way in order to explain the reasoning behind them. In practice, it is possible to streamline them considerably. This shortened procedure also reveals more clearly the sense in which an analysis of variance has taken place. Note that the following symbols are used:

N = total number of items of data,
n = number in each sample,
k = number of samples,
Σx_1 = sum of the items in Sample 1,
Σx = sum of all the items of data.

Step 1 – Code the data. All the variance calculations are carried out in terms of deviations around means. We may thus often substantially reduce the size of the numbers involved, without affecting these deviations, by subtracting a constant from each item of data.

Subtracting 30 from each item of our wool data produces the following:

	Sample			
	1	*2*	*3*	*4*
	12	23	18	10
	17	25	22	14
	12	21	15	6
	9	21	16	7
	16	26	21	13
TOTALS	66	116	92	50

The choice of 30 as the constant is, of course, purely arbitrary. We could just as easily have taken 40, which would have reduced the size of the numbers even more. However, this would have made some of the values negative. There is nothing wrong with having negative numbers, but it makes the possibility of an error in the computations that much greater!

Step 2 – Calculate total sum of squares (TSS). The total sum of squares was defined earlier as the sum of squared deviations of all the items around the overall mean $\Sigma(x - \bar{x})^2$. It is thus a measure of the total amount of variation present in the data.

For computational purposes, $\Sigma (x - \bar{x})^2$ may be expressed in the more convenient form $\Sigma x^2 - \frac{(\Sigma x)^2}{N}$.

$$\Sigma x = 12 + 17 + \ldots + 13 = 324$$
$$\Sigma x^2 = 12^2 + 17^2 + \ldots + 13^2 = 5906$$

$$\frac{(\Sigma x)^2}{N} = \frac{(324)^2}{20} = 5248.8$$

$$\text{Total sum of squares} = \Sigma x^2 - \frac{(\Sigma x)^2}{N}$$
$$= 5906 - 5248.8$$
$$= 657.2$$

The quantity $\frac{(\Sigma x)^2}{N}$ is known as the **correction factor** and will be used again later.

Step 3 – Calculate between-samples sum of squares (BSSS). This is the **between-sample-totals sum of squares,** divided by n immediately, rather than at the end as in the earlier calculation. Notationally, the BSSS may be expressed as:

$$\frac{1}{n} [(\Sigma x_1)^2 + (\Sigma x_2)^2 + \ldots + (\Sigma x_k)^2] - \frac{(\Sigma x)^2}{N}$$
$$= \tfrac{1}{5}(66^2 + 116^2 + 92^2 + 50^2) - 5248.8$$
$$= 506.4$$

Step 4 – Calculate the within-samples sum of squares (WSSS). The **within-samples sum of squares** was calculated in detail earlier. If you check back, you will find that it was 150.8. However, this detailed calculation is not necessary because it can be shown that

the WSSS and the BSSS will always add up to the TSS. To get the WSSS, therefore, we simply have:

$$\begin{aligned} \text{WSSS} &= \text{TSS} - \text{BSSS} \\ &= 657.2 - 506.4 = 150.8 \end{aligned}$$

This illustrates how the total variation, as measured by the TSS, has been 'analysed' into two parts – the 'Within-Samples' variation and the 'Between-Samples' variation.

Step 5 – Calculate the degrees of freedom.

The TSS is based on $(N-1)$ d.f.	$= 19$
The BSSS is based on $(k-1)$ d.f.	$= 3$
The WSSS is based on $k(n-1)$	$= 16$

Notice that the d.f. associated with the BSSS and the WSSS add up to the d.f. associated with the TSS. Thus the total number of degrees of freedom has also been separated into two components.

Step 6 – Draw up an analysis of variance table. All the above results may be neatly summarized in tabular form, and the required significance test carried out:

Source	*d.f.*	*Sums of squares*	*Variance estimate*	*F*
Between samples	3	506.4	168.8	17.90
Within samples	16	150.8	9.43	
TOTALS	19	657.2		

As we always test whether the between-samples variance estimate is significantly greater than the within-samples, a one-tail test is appropriate. The value of 17.90 for the F-ratio is well above the level necessary to establish that there is a significant difference between the sample means ($F_{.01;3,16} = 5.29$).

Confidence Limits

Having established that the sample means do differ significantly, it does not necessarily follow that each one differs significantly from every other. It may be that just one of them is different from the remainder.

To help interpret the practical meaning of the established significance, we can calculate confidence limits for each mean and sort them into groups. The 4 sample averages are as follows:

Sample			
1	*2*	*3*	*4*
43.2	53.2	48.4	40.0

The within-samples variance estimate in the analysis of variance table provides an estimate of the variance of the population of balls of wools within each factory. The square root of this will thus be an estimate of the standard deviation (σ) and as each sample consists of 5 items, the standard error applicable to each of the mean estimates will be $\frac{\sigma}{\sqrt{n}}$, i.e. $\frac{\sigma}{\sqrt{5}}$ in this case.

$$\sigma^2 = 9.43, \text{ therefore } \sigma = 3.07$$

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{3.07}{\sqrt{5}} = 1.37$$

The estimate of σ used in this standard error calculation is based on 16 d.f., therefore 95% confidence limits for each sample mean will be $\pm\, t_{.025;16}\,\text{SE} = \pm\, 2.12 \times 1.37 = \pm\, 2.90$

We thus have the following 95% confidence intervals for each mean:

Sample 1 43.2 ± 2.9 = 40.3 to 46.1
Sample 2 53.2 ± 2.9 = 50.3 to 56.1
Sample 3 48.4 ± 2.9 = 45.5 to 51.3
Sample 4 40.0 ± 2.9 = 37.1 to 42.9

We can get a better feel for the relationship between these intervals by showing them diagramatically, as in Figure 1.1.
There is a large overlap between Samples 1 and 4 which suggests that they are not significantly different from each other. There is much less overlap between 1 and 3, which strongly suggests that they are significantly different from each other, although at what level of significance is not clear. A similar remark applies to Samples 3 and 2. However, Samples 3 and 2 are clearly different from Sample 4.