

Analysis of Variance

44.1	One-Way Analysis of Variance	2
44.2	Two-Way Analysis of Variance	15
44.3	Experimental Design	40

Learning Outcomes

In this Workbook you will learn the basics of this very important branch of Statistics and how to do the calculations which enable you to draw conclusions about variance found in data sets. You will also be introduced to the design of experiments which has great importance in science and engineering.

One-Way Analysis of Variance

44.1

Introduction

Problems in engineering often involve the exploration of the relationships between values taken by a variable under different conditions. HELM 41 introduced hypothesis testing which enables us to compare two population means using hypotheses of the general form

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

or, in the case of more than two populations,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_1 : H_0 \text{ is not true}$$

If we are comparing more than two population means, using the type of hypothesis testing referred to above gets very clumsy and very time consuming. As you will see, the statistical technique called Analysis of Variance (ANOVA) enables us to compare several populations **simultaneously**. We might, for example need to compare the shear strengths of five different adhesives or the surface toughness of six samples of steel which have received different surface hardening treatments.



Prerequisites

Before starting this Section you should ...

- be familiar with the general techniques of hypothesis testing
- be familiar with the F -distribution



Learning Outcomes

On completion you should be able to ...

- describe what is meant by the term one-way ANOVA.
- perform one-way ANOVA calculations.
- interpret the results of one-way ANOVA calculations

1. One-way ANOVA

In this Workbook we deal with one-way analysis of variance (one-way ANOVA) and two-way analysis of variance (two-way ANOVA). One-way ANOVA enables us to compare several means simultaneously by using the F -test and enables us to draw conclusions about the variance present in the set of samples we wish to compare.

Multiple (greater than two) samples may be investigated using the techniques of two-population hypothesis testing. As an example, it is possible to do a comparison looking for variation in the surface hardness present in (say) three samples of steel which have received different surface hardening treatments by using hypothesis tests of the form

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

We would have to compare all possible pairs of samples before reaching a conclusion. If we are dealing with three samples we would need to perform a total of

$${}^3C_2 = \frac{3!}{1!2!} = 3$$

hypothesis tests. From a practical point of view this is not an efficient way of dealing with the problem, especially since the number of tests required rises rapidly with the number of samples involved. For example, an investigation involving ten samples would require

$${}^{10}C_2 = \frac{10!}{8!2!} = 45$$

separate hypothesis tests.

There is also another crucially important reason why techniques involving such batteries of tests are unacceptable. In the case of 10 samples mentioned above, if the probability of correctly accepting a given null hypothesis is 0.95, then the probability of correctly accepting the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_{10}$$

is $(0.95)^{45} \approx 0.10$ and we have only a 10% chance of correctly accepting the null hypothesis for all 45 tests. Clearly, such a low success rate is unacceptable. These problems may be avoided by simultaneously testing the significance of the difference between a set of more than two population means by using techniques known as the analysis of variance.

Essentially, we look at the variance *between* samples and the variance *within* samples and draw conclusions from the results. Note that the variation between samples is due to assignable (or controlled) causes often referred in general as **treatments** while the variation within samples is due to **chance**. In the example above concerning the surface hardness present in three samples of steel which have received different surface hardening treatments, the following diagrams illustrate the differences which may occur when between sample and within sample variation is considered.

Case 1

In this case the variation within samples is roughly on a par with that occurring between samples.

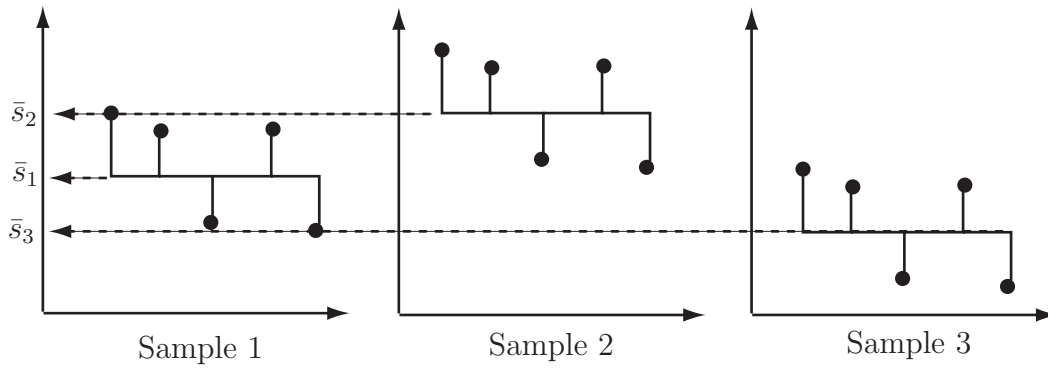


Figure 1

Case 2

In this case the variation within samples is considerably less than that occurring between samples.

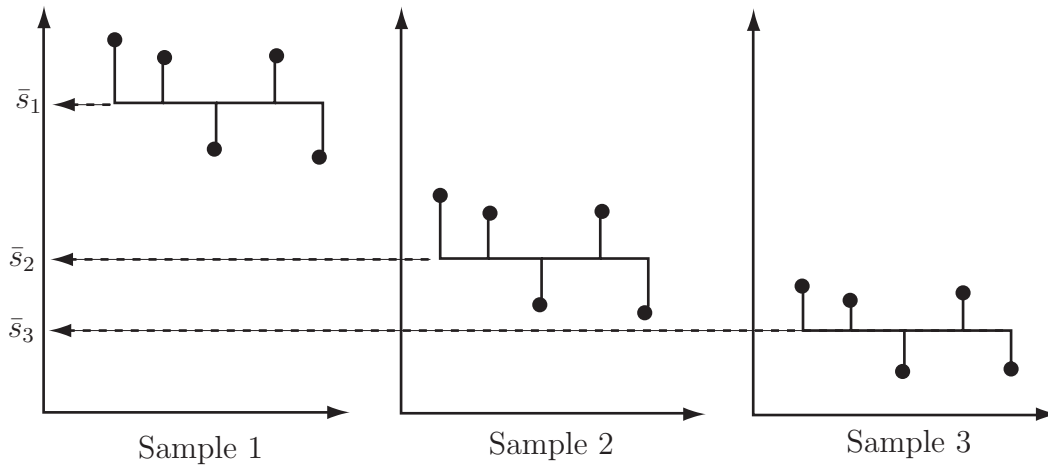


Figure 2

We argue that the greater the variation present *between* samples in comparison with the variation present *within* samples the more likely it is that there are ‘real’ differences between the population means, say μ_1, μ_2 and μ_3 . If such ‘real’ differences are shown to exist at a sufficiently high level of significance, we may conclude that there is sufficient evidence to enable us to reject the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$.

Example of variance in data

This example looks at variance in data. Four machines are set up to produce alloy spacers for use in the assembly of microlight aircraft. The spaces are supposed to be identical but the four machines give rise to the following varied lengths in mm.

Machine A	Machine B	Machine C	Machine D
46	56	55	49
54	55	51	53
48	56	50	57
46	60	51	60
56	53	53	51

Since the machines are set up to produce identical alloy spacers it is reasonable to ask if the evidence we have suggests that the machine outputs are the same or different in some way. We are really asking whether the sample means, say $\bar{X}_A, \bar{X}_B, \bar{X}_C$ and \bar{X}_D , are different because of differences in the respective population means, say μ_A, μ_B, μ_C and μ_D , or whether the differences in $\bar{X}_A, \bar{X}_B, \bar{X}_C$ and \bar{X}_D may be attributed to chance variation. Stated in terms of a hypothesis test, we would write

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

H_1 : At least one mean is different from the others

In order to decide between the hypotheses, we calculate the mean of each sample and overall mean (the mean of the means) and use these quantities to calculate the variation present *between* the samples. We then calculate the variation present *within* samples. The following tables illustrate the calculations.

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

H_1 : At least one mean is different from the others

Machine A	Machine B	Machine C	Machine D
46	56	55	49
54	55	51	53
48	56	50	57
46	60	51	60
56	53	53	51
$\bar{X}_A = 50$	$\bar{X}_B = 56$	$\bar{X}_C = 52$	$\bar{X}_D = 54$

The mean of the means is clearly

$$\bar{\bar{X}} = \frac{50 + 56 + 52 + 54}{4} = 53$$

so the variation present between samples may be calculated as

$$\begin{aligned} S_{Tr}^2 &= \frac{1}{n-1} \sum_{i=A}^D (\bar{X}_i - \bar{\bar{X}})^2 \\ &= \frac{1}{4-1} ((50-53)^2 + (56-53)^2 + (52-53)^2 + (54-53)^2) \\ &= \frac{20}{3} = 6.67 \text{ to 2 d.p.} \end{aligned}$$

Note that the notation S_{Tr}^2 reflects the general use of the word 'treatment' to describe assignable causes of variation between samples. This notation is not universal but it is fairly common.

Variation within samples

We now calculate the variation due to chance errors present *within* the samples and use the results to obtain a pooled estimate of the variance, say S_E^2 , present *within* the samples. After this calculation we will be able to compare the two variances and draw conclusions. The variance present within the samples may be calculated as follows.

Sample A

$$\sum (X - \bar{X}_A)^2 = (46 - 50)^2 + (54 - 50)^2 + (48 - 50)^2 + (46 - 50)^2 + (56 - 50)^2 = 88$$

Sample B

$$\sum (X - \bar{X}_B)^2 = (56 - 56)^2 + (55 - 56)^2 + (56 - 56)^2 + (60 - 56)^2 + (53 - 56)^2 = 26$$

Sample C

$$\sum (X - \bar{X}_C)^2 = (55 - 52)^2 + (51 - 52)^2 + (50 - 52)^2 + (51 - 52)^2 + (53 - 52)^2 = 16$$

Sample D

$$\sum (X - \bar{X}_D)^2 = (49 - 54)^2 + (53 - 54)^2 + (57 - 54)^2 + (60 - 54)^2 + (51 - 54)^2 = 80$$

An obvious extension of the formula for a pooled variance gives

$$S_E^2 = \frac{\sum (X - \bar{X}_A)^2 + \sum (X - \bar{X}_B)^2 + \sum (X - \bar{X}_C)^2 + \sum (X - \bar{X}_D)^2}{(n_A - 1) + (n_B - 1) + (n_C - 1) + (n_D - 1)}$$

where n_A, n_B, n_C and n_D represent the number of members (5 in each case here) in each sample. Note that the quantities comprising the denominator $n_A - 1, \dots, n_D - 1$ are the number of degrees of freedom present in each of the four samples. Hence our pooled estimate of the variance present within the samples is given by

$$S_E^2 = \frac{88 + 26 + 16 + 80}{4 + 4 + 4 + 4} = 13.13$$

We are now in a position to ask whether the variation *between* samples S_{Tr}^2 is large in comparison with the variation *within* samples S_E^2 . The answer to this question enables us to decide whether the difference in the calculated variations is sufficiently large to conclude that there is a difference in the population means. That is, do we have sufficient evidence to reject H_0 ?

Using the F -test

At first sight it seems reasonable to use the ratio

$$F = \frac{S_{Tr}^2}{S_E^2}$$

but in fact the ratio

$$F = \frac{nS_{Tr}^2}{S_E^2},$$

where n is the sample size, is used since it can be shown that if H_0 is true this ratio will have a value of approximately unity while if H_0 is not true the ratio will have a value greater than unity. This is because the variance of a sample mean is σ^2/n .

The test procedure (three steps) for the data used here is as follows.

- (a) Find the value of F ;
- (b) Find the number of degrees of freedom for both the numerator and denominator of the ratio;

- (c) Accept or reject depending on the value of F compared with the appropriate tabulated value.

Step 1

The value of F is given by

$$F = \frac{nS_{Tr}^2}{S_E^2} = \frac{5 \times 6.67}{13.13} = 2.54$$

Step 2

The number of degrees of freedom for S_{Tr}^2 (the numerator) is

$$\text{Number of samples} - 1 = 3$$

The number of degrees of freedom for S_E^2 (the denominator) is

$$\text{Number of samples} \times (\text{sample size} - 1) = 4 \times (5 - 1) = 16$$

Step 3

The critical value (5% level of significance) from the F -tables (Table 1 at the end of this Workbook) is $F_{(3,16)} = 3.24$ and since $2.54 < 3.224$ we see that we cannot reject H_0 on the basis of the evidence available and conclude that in this case the variation present is due to chance. Note that the test used is one-tailed.

ANOVA tables

It is usual to summarize the calculations we have seen so far in the form of an ANOVA table. Essentially, the table gives us a method of recording the calculations leading to both the numerator and the denominator of the expression

$$F = \frac{nS_{Tr}^2}{S_E^2}$$

In addition, and importantly, ANOVA tables provide us with a useful means of checking the accuracy of our calculations. A general ANOVA table is presented below with explanatory notes.

Define a = number of treatments, n = number of observations per sample.

Source of Variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS	Value of F Ratio
<i>Between samples</i> (due to treatments) Differences between means \bar{X}_i and \bar{X}	$SS_{Tr} = n \sum_{i=1}^a (\bar{X}_i - \bar{X})^2$	$(a - 1)$	$MS_{Tr} = \frac{SS_{Tr}}{(a - 1)}$ $= nS_{\bar{X}}^2$	$F = \frac{MS_{Tr}}{MS_E}$ $= \frac{nS_{Tr}^2}{S_E^2}$
<i>Within samples</i> (due to chance errors) Differences between individual observations X_{ij} and means \bar{X}_i	$SS_E = \sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$	$a(n - 1)$	$MS_E = \frac{SS_E}{a(n - 1)}$ $= S_E^2$	
TOTALS	$SS_T = \sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X})^2$	$(an - 1)$		

In order to demonstrate this table for the example above we need to calculate

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X})^2$$

a measure of the total variation present in the data. Such calculations are easily done using a computer (Microsoft Excel was used here), the result being

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X})^2 = 310$$

The ANOVA table becomes

Source of Variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS	Value of F Ratio
<i>Between samples</i> <i>(due to treatments)</i> Differences between means \bar{X}_i and \bar{X}	100	3	$MS_{Tr} = \frac{SS_{Tr}}{(a-1)}$ $= \frac{100}{3}$ $= 33.33$	$F = \frac{MS_{Tr}}{MS_E}$ $= 2.54$
<i>Within samples</i> <i>(due to chance errors)</i> Differences between individual observations X_{ij} and means \bar{X}_i	210	16	$MS_E = \frac{SS_E}{a(n-1)}$ $= \frac{210}{16}$ $= 13.13$	
TOTALS	310	19		

It is possible to show theoretically that

$$SS_T = SS_{Tr} + SS_E$$

that is

$$\sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X})^2 = n \sum_{i=1}^a (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^a \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$$

As you can see from the table, SS_{Tr} and SS_E do indeed sum to give SS_T even though we can calculate them separately. The same is true of the degrees of freedom.

Note that calculating these quantities separately does offer a check on the arithmetic but that using the relationship can speed up the calculations by obviating the need to calculate (say) SS_T . As you might expect, it is recommended that you check your calculations! However, you should note that it is usual to calculate SS_T and SS_{Tr} and then find SS_E by subtraction. This saves a lot of unnecessary calculation but does not offer a check on the arithmetic. This shorter method will be used throughout much of this Workbook.

Unequal sample sizes

So far we have assumed that the number of observations in each sample is the same. This is not a necessary condition for the one-way ANOVA.



Key Point 1

Suppose that the number of samples is a and the numbers of observations are n_1, n_2, \dots, n_a . Then the between-samples sum of squares can be calculated using

$$SS_{Tr} = \sum_{i=1}^a \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

where T_i is the total for sample i , $G = \sum_{i=1}^a T_i$ is the overall total and $N = \sum_{i=1}^a n_i$.

It has $a - 1$ degrees of freedom.

The total sum of squares can be calculated as before, or using

$$SS_T = \sum_{i=1}^a \sum_{j=1}^{n_i} X_{ij}^2 - \frac{G^2}{N}$$

It has $N - 1$ degrees of freedom.

The within-samples sum of squares can be found by subtraction:

$$SS_E = SS_T - SS_{Tr}$$

It has $(N - 1) - (a - 1) = N - a$ degrees of freedom.



Three fuel injection systems are tested for efficiency and the following coded data are obtained.

System 1	System 2	System 3
48	60	57
56	56	55
46	53	52
45	60	50
50	51	51

Do the data support the hypothesis that the systems offer equivalent levels of efficiency?

Your solution

Answer

Appropriate hypotheses are

$$H_0 = \mu_1 = \mu_2 = \mu_3$$

H_1 : At least one mean is different to the others

Variation between samples

System 1	System 2	System 3
48	60	57
56	56	55
46	53	52
45	60	50
50	51	51
$\bar{X}_1 = 49$	$\bar{X}_2 = 56$	$\bar{X}_3 = 53$

The mean of the means is $\bar{\bar{X}} = \frac{49 + 56 + 53}{3} = 52.67$ and the variation present *between* samples is

$$S_{Tr}^2 = \frac{1}{n-1} \sum_{i=1}^3 (\bar{X}_i - \bar{\bar{X}})^2 = \frac{1}{3-1} ((49 - 52.67)^2 + (56 - 52.67)^2 + (53 - 52.67)^2) = 12.33$$

Variation within samples**System 1**

$$\sum (X - \bar{X}_1)^2 = (48 - 49)^2 + (56 - 49)^2 + (46 - 49)^2 + (45 - 49)^2 + (51 - 49)^2 = 76$$

System 2

$$\sum (X - \bar{X}_2)^2 = (60 - 56)^2 + (56 - 56)^2 + (53 - 56)^2 + (60 - 56)^2 + (51 - 56)^2 = 66$$

System 3

$$\sum (X - \bar{X}_3)^2 = (57 - 53)^2 + (55 - 53)^2 + (52 - 53)^2 + (50 - 53)^2 + (51 - 53)^2 = 34$$

Hence

$$S_E^2 = \frac{\sum (X - \bar{X}_1)^2 + \sum (X - \bar{X}_2)^2 + \sum (X - \bar{X}_3)^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)} = \frac{76 + 66 + 34}{4 + 4 + 4} = 14.67$$

The value of F is given by $F = \frac{nS_{Tr}^2}{S_E^2} = \frac{5 \times 12.33}{14.67} = 4.20$

The number of degrees of freedom for S_{Tr}^2 is No. of samples $- 1 = 2$

The number of degrees of freedom for S_E^2 is No. of samples \times (sample size $- 1$) $= 12$

The critical value (5% level of significance) from the F -tables (Table 1 at the end of this Workbook) is $F_{(2,12)} = 3.89$ and since $4.20 > 3.89$ we conclude that we have sufficient evidence to reject H_0 so that the injection systems are not of equivalent efficiency.

Exercises

1. The yield of a chemical process, expressed in percentage of the theoretical maximum, is measured with each of two catalysts, A, B, and with no catalyst (Control: C). Five observations are made under each condition. Making the usual assumptions for an analysis of variance, test the hypothesis that there is no difference in mean yield between the three conditions. Use the 5% level of significance.

Catalyst A	Catalyst B	Control C
79.2	81.5	74.8
80.1	80.7	76.5
77.4	80.5	74.7
77.6	81.7	74.8
77.8	80.6	74.9

2. Four large trucks, A, B, C, D, are used to move stone in a quarry. On a number of days, the amount of fuel, in litres, used per tonne of stone moved is calculated for each truck. On some days a particular truck might not be used. The data are as follows. Making the usual assumptions for an analysis of variance, test the hypothesis that the mean amount of fuel used per tonne of stone moved is the same for each truck. Use the 5% level of significance.

Truck	Observations										
A	0.21	0.21	0.21	0.21	0.20	0.19	0.18	0.21	0.22	0.21	
B	0.22	0.22	0.25	0.21	0.21	0.22	0.20	0.23			
C	0.21	0.18	0.18	0.19	0.20	0.18	0.19	0.19	0.20	0.20	0.20
D	0.20	0.20	0.21	0.21	0.21	0.19	0.20	0.20	0.21		

Answers

1. We calculate the treatment totals for A: 392.1, B: 405.0 and C: 375.7. The overall total is 1172.8 and $\sum \sum y^2 = 91792.68$.

The total sum of squares is

$$91792.68 - \frac{1172.8^2}{15} = 95.357$$

on $15 - 1 = 14$ degrees of freedom.

The between treatments sum of squares is

$$\frac{1}{5}(392.1^2 + 405.0^2 + 375.7^2) - \frac{1172.8^2}{15} = 86.257$$

on $3 - 1 = 2$ degrees of freedom.

By subtraction, the residual sum of squares is

$$95.357 - 86.257 = 9.100$$

on $14 - 2 = 12$ degrees of freedom.

The analysis of variance table is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Treatment	86.257	2	43.129	56.873
Residual	9.100	12	0.758	
Total	95.357	14		

The upper 5% point of the $F_{2,12}$ distribution is 3.89. The observed variance ratio is greater than this so we conclude that the result is significant at the 5% level and we reject the null hypothesis at this level. The evidence suggests that there are differences in the mean yields between the three treatments.

Answer

2. We can summarise the data as follows.

Truck	$\sum y$	$\sum y^2$	n
A	2.05	0.4215	10
B	1.76	0.3888	8
C	2.12	0.4096	11
D	1.83	0.3725	9
Total	7.76	1.5924	38

The total sum of squares is

$$1.5924 - \frac{7.76^2}{38} = 7.7263 \times 10^{-3}$$

on $38 - 1 = 37$ degrees of freedom.

The between trucks sum of squares is

$$\frac{2.05^2}{10} + \frac{1.76^2}{8} + \frac{2.12^2}{11} + \frac{1.83^2}{9} - \frac{7.76^2}{38} = 3.4581 \times 10^{-3}$$

on $4 - 1 = 3$ degrees of freedom.

By subtraction, the residual sum of squares is

$$7.7263 \times 10^{-3} - 3.4581 \times 10^{-3} = 4.2682 \times 10^{-3}$$

on $37 - 3 = 34$ degrees of freedom.

The analysis of variance table is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Trucks	3.4581×10^{-3}	3	1.1527×10^{-3}	9.1824
Residual	4.2682×10^{-3}	34	0.1255×10^{-3}	
Total	7.7263×10^{-3}	37		

The upper 5% point of the $F_{3,34}$ distribution is approximately 2.9. The observed variance ratio is greater than this so we conclude that the result is significant at the 5% level and we reject the null hypothesis at this level. The evidence suggests that there are differences in the mean fuel consumption per tonne moved between the four trucks.