29 Introduction to Analysis of Variance

29.1. Introduction

In this chapter, linear equations of observational data with constant coefficients and the method of calculating their variations are explained. Analysis of variance is also introduced. This chapter is based on Genichi Taguchi et al., *Design of Experiments.* Tokyo: Japanese Standards Association, 1973.

29.2. Decomposition of Variation

Letting n observational values be $y_1, y_2, \ldots, y_n$, the equation for $y_1, y_2, \ldots, y_n$ with constant coefficients is called a *linear equation.* Let n coefficients be $c_1, c_2, \ldots, c_n$; also let the linear equation be L:

$$L = c_1y_1 + c_2y_2 + \cdots + c_ny_n \tag{29.1}$$

Some of the coefficients $c_1, c_2, \ldots, c_n$, could be zero, but assume that not all the coefficients are zero.

The sum of these coefficients squared is denoted by D, which is the number of units comprising the linear equation, L. D is

$$D = c_1^2 + c_2^2 + \cdots + c^{2_n} \tag{29.2}$$

For example, the mean (deviation) of n measurements, denoted by $\bar{y}$, is

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

$$= \frac{1}{n}y_1 + \frac{1}{n}y_2 + \cdots + \frac{1}{n}y_n \tag{29.3}$$

Equation (29.3) is a linear equation in which all the coefficients are identical.

$$c_1 = c_2 = \cdots = c_n = \frac{1}{n} \tag{29.4}$$

Its number of units is

$$D = \left(\frac{1}{n}\right)^2 n = \frac{1}{n} \tag{29.5}$$

The square of a linear equation divided by its number of units forms a variation with one degree of freedom. Letting it be S_L, then

$$S_L = \frac{L^2}{c_1^2 + c_2^2 + \cdots + c_n^2} = \frac{L^2}{D} \tag{29.6}$$

The sum of $y_1, y_2, \ldots, y_n$ is expressed by a linear equation with $c_1 = c_2 = \cdots = c_n = 1$, with n units. Variation S_L is written from equation (29.6) as

$$S_L = \frac{(y_1 + y_2 + \cdots + y_n)^2}{n} \tag{29.7}$$

The variation corresponds to variation S_m, which is the correction factor. Following is an example to clarify linear equations and their variation.

❒ Example 1

There are six Japanese and four Americans with the following heights in centimeters:

A_1 (*Japanese*): 158, 162, 155, 172, 160, 168

A_2 (*American*): 186, 172, 176, 180

The quality of total variation among the 10 persons is given by the sum of deviations (from the mean) squared. Since there is neither an objective value nor a theoretical value, a working mean of 170 is subtracted and the total variation among the 10 persons is calculated.

- *Deviations from the working mean for* A_1: $-12, -8, -15, 2, -10, -2$
  Total: -45
- *Deviations from the working mean for* A_2: 16, 2 , 6, 10
  Total: 34

Therefore,

$$\text{CF} = \frac{(-45 + 34)^2}{10} = \frac{(-11)^2}{10} = 12 \tag{29.8}$$

$$S_T = (-12)^2 + (-8)^2 + \cdots + 10^2 - \text{CF} = 937 - 12 = 925 \tag{29.9}$$

Observing the data above, one must understand that the major part of the total variation of heights among the 10 persons is caused by the difference between the Japanese and the Americans. How much of the total variation of heights among the 10 persons, which was calculated to be 925, was caused by the differences between the two nationalities?

Assume that everyone expresses the difference between the mean heights of the Japanese and the Americans as follows:

$$\begin{aligned} L &= (\text{mean of Japanese}) - (\text{mean of Americans}) \\ &= \left(170 + \frac{-45}{6}\right) - \left(170 + \frac{34}{4}\right) = \frac{-45}{6} - \frac{34}{4} \end{aligned} \tag{29.10}$$

Usually, the total of the Japanese (A_1) is denoted by the same symbol, A_1, and the total of the Americans (A_2), by A_2. Thus,

$$L = \frac{A_1}{6} - \frac{A_2}{4} \tag{29.11}$$

Now let the heights of the 10 persons by y_1, y_2, y_3, y_4, y_5, y_6 (Japanese), y_7, y_8, y_9, y_{10} (Americans). The total variation, S_T, is the total variation of the 10 measurements of height y_1, y_2, ... , y_{10}. Equations (29.10) and (29.11) are the linear equations of 10 observational values y_1, y_2, ... , y_{10} with constant coefficients.

Accordingly, the variation S_L for equation (29.11) is

$$\begin{aligned} S_A = S_L &= \frac{(A_1/6 - A_2/4)^2}{(1/6)^2(6) + (-1/4)^2(4)} = \frac{(-45/6 - 34/4)^2}{1/6 + 1/4} \\ &= \frac{[(4)(-45) - (6)(34)]^2}{(4^2)(6) + (-6)^2(4)} = \frac{(-384)^2}{240} = 614 \end{aligned} \tag{29.12}$$

It is seen that with the total variation among 10 heights, which in this case is 925, the difference between Japanese and Americans can be as much as 614. By subtracting 614 from 925, we then identify the magnitude of individual differences within the same nationality. This difference is called the *error variation*, denoted by S_e:

$$S_e = S_T - S_A = 925 - 614 = 311 \tag{29.13}$$

The degrees of freedom are 9 for S_T, 1 for S_A, and $(9 - 1 = 8)$ for S_e. Therefore, the error variance V_e is

$$V_e = \frac{S_e}{8} = \frac{311}{8} = 38.9 \tag{29.14}$$

Error variance shows the height differences per person within the same nationality.

The square-root value of error variance (V_e) is called the *standard deviation* of individual difference of height, usually denoted by s:

$$s = \sqrt{V_e} = \sqrt{38.9} = 6.24 \text{ cm} \tag{29.15}$$

In this way, the total variation of height among the 10 persons, S_T, is decomposed or separated into the variation of the difference between the Japanese and the Americans and the variation of the individual differences.

$$S_T = S_A + S_e \tag{29.16}$$

$$925 = 614 + 311 \qquad \text{(variation)}$$

$$9 = 1 + 8 \qquad \text{(degrees of freedom)} \tag{29.17}$$

Since there are 10 persons, there must be 10 individual differences; otherwise, the data are contradictory. In S_e, however, there are only eight individual differences. To make the total of 10, one difference is included in correction factor CF, and one difference in S_A.

In general, one portion of the individual differences is included in S_A, which equaled 614. Therefore, the real differences between nations must be calculated by subtracting a portion of the individual differences (which is the error variance) from S_A. The result after subtracting the error variance is called the *pure variation* between Japanese and Americans, distinguished from the original variation using a prime:

$$S'_A = S_A - V_e = 614 - 38.9 = 575.1 \tag{29.18}$$

Another portion of the individual differences is also in the correction factor. But in the case of the variation in height, this portion does not provide useful information. However, it will be shown in Example 2 where the correction factor does indicate some useful information. For this reason, the total quantity of variation is considered to have nine degrees of freedom.

Accordingly, the pure variation of error, S'_e, is calculated by adding the error variance, V_e (which is the unit portion subtracted from S_A), to S_e:

$$S'_e = S_e + V_e = 311 + 38.9 = 349.9 \tag{29.19}$$

Then the following relationship is concluded:

$$S_T = S'_A + S'_e \tag{29.20}$$

$$925 = 575.1 + 349.9 \tag{29.21}$$

Each term of equation (29.20) is divided by S_T and then multiplied by 100. This calculation is called the *decomposition* of the degrees of contribution.

In this case, the following relationship is concluded:

$$100 = \rho_A + \rho_e$$

$$= \frac{575.1}{925}(100) + \frac{349.9}{925}(100) = 62.2 + 37.8 \tag{29.22}$$

Equation (29.22) shows that the total varying quantity of 10 heights is equal to 925; the differences between the Japanese and American nationalities comprises 62.2% of the cause, and the individual differences consitutes 37.8%

In this way, the decomposition of variation is necessary to calculate when each data value changes. This is to express the total change by a total variation and then calculate the degrees of contribution, which is the influence of individual causes among the total variation, S_T.

To decompose variations at ease, one must master the calculation of the analysis of variance, which, mathematically, is the expression of variation in quadratic form. However, the most important thing is the cause of the variation. Possible causes must first be suggested by a researcher in the specialized field being studied. Without knowledge of the field, one cannot solve the problem of analysis of variance, no matter how well he or she knows the mathematics of quadratic equations.

It is important for a research worker to become acquainted with the analysis of variance, to have the ability to decompose variations for any type of data, and to be able to quickly calculate the magnitude of the influence of the cause being considered. Next, an example using an objective or theoretical value is shown.

❒ Example 2

To improve the antiabrasive properties of a product, two versions were manufactured on a trial basis.

A_1: product with no addictive in the raw material

A_2: product additive with a particular in the raw material

Six test pieces were prepared from the two versions, and wear tests were carried out under identical conditions. The quantities of wear (mg) were as follows:

A_1: 26, 18, 19, 21, 15, 29

A_2: 15, 8, 14, 13, 16, 9

In a case such as abrasion, quantity of deterioration, or percent of deterioration, the objective value would be zero. Accordingly, the correction factor is calculated from the original data:

$$A_1 = 26 + 18 + \cdots + 29 = 128 \tag{29.23}$$

$$A_2 = 15 + 8 + \cdots + 9 = 75 \tag{29.24}$$

$$T = A_1 + A_2 = 203 \tag{29.25}$$

Therefore, the correction factor would be

$$\mathrm{CF} = \frac{203^2}{12} = 3434 \tag{29.26}$$

When there is an objective value, the correction factor CF is also called the *variation of the general mean*, denoted by S_m. It signifies the magnitude of the average quantity of abrasion of all the data.

The difference between averages A_1 and A_2, denoted by L, is

$$L = \frac{A_1}{6} - \frac{A_2}{6} = \frac{A_1 - A_2}{6} \tag{29.27}$$

The variation of the differences between A_1 and A_2, or S_A, is calculated from the square of linear equation L divided by the sum of the coefficients of L squared.

$$\begin{aligned} S_A = S_L &= \frac{L^2}{\text{sum of coefficients squared}} \\ &= \frac{[(A_1 - A_2)/6]^2}{(1/6)^2(6) + (-1/6)^2(6)} = \frac{(A_1 - A_2)^2}{(1^2)(6) + (-1)^2(6)} = \frac{(A_1 - A_2)^2}{12} \\ &= \frac{(128 - 75)^2}{12} = \frac{53^2}{12} = \frac{2809}{12} = 234 \end{aligned} \tag{29.28}$$

The objective value in this case is zero, so the total variation is equal to the total sum squared.

$$S_T = 26^2 + 18^2 + \cdots + 9^2 = 3859 \tag{29.29}$$

Error variation is then

$$S_e = S_T - \text{CF} - S_A = 3859 - 3434 - 234 = 191 \tag{29.30}$$

Degrees of freedom are 12 for S_T, 1 for the correction factor (CF or general mean S_m), and $(12 - 2) = 10$ for the error variation S_e.

Accordingly, the decomposition of the variation as well as the degrees of freedom are concluded as follows:

$$S_T = S_m + S_A + S_e \tag{29.31}$$

$$3859 = 34.34 + 234 + 191 \qquad \text{(variation)} \tag{29.32}$$

$$12 = 1 + 1 + 10 \qquad \text{(degrees of freedom)} \tag{29.33}$$

Decomposition of the degrees of contribution is made as

$$S'_m = S_m - V_e = 3434 - \frac{S_e}{10} = 3434 - 19.1 = 3414.9 \tag{29.34}$$

$$S'_A = S_A - V_e = 234 - 19.1 = 214.9 \tag{29.35}$$

Also,

$$S'_e = S_e + 2V_e = 191 + (2)(19.1) = 229.2 \tag{37.36}$$

Degrees of contribution are

$$\rho_m = \frac{S'_m}{S_T} = \frac{3414.9}{3859} = 0.885 \tag{29.37}$$

$$\rho = \frac{S'_A}{S_T} = \frac{214.9}{3859} = 0.056 \tag{29.38}$$

$$\rho_e = \frac{S'_e}{S_T} = \frac{229.2}{3859} = 0.059 \tag{29.39}$$

Therefore,

$$100 = \rho_m + \rho_A + \rho_e = 88.5 + 5.6 + 5.9\% \tag{29.40}$$

29.3. Analysis-of-Variance Table

The results of the decomposing variation shown in Section 29.2 are arranged in an analysis of variance table. To affirm the existence of correction factor or the difference of A qualitatively, a *significance test* is made prior to calculation of the degrees of contribution in the traditional analysis of variance. In quality engineering, however, the significance is observed from the degrees of contribution. It is a basic rule to calculate the degrees of contribution only for those causes (called *significant factorial effects*) indicated by asterisks. However, we should not make light of insignificant factoral effects whose degrees of contribution are large. There is a high possibility that those factor effects will have a substantial influence on the result. An insignificant result is obtained in two situations: In one, the effect is really nonexistent; in the other, an effect does exist but there is insufficient evidence to affirm the significance associated with the small number of degrees of freedom of the error variance. Calculation of the degrees of contribution is shown

Table 29.1
ANOVA table

Source	Degrees of Freedom, f	Variation, S	Variance, V	Pure Variation, S'	Degrees of Contribution, ρ (%)
m	1	CF	V_m	S'_m	ρ_m
A	1	S_A	V_A	S'_A	ρ_A
e	$n - 2$	S_e	V_e	S'_e	ρ_e
Total	n	S_T		S_T	100.0

Table 29.2
ANOVA table for Example 2

Source	Degrees of Freedom, f	Variation, S	Variance, V	Pure Variation, S'	Degrees of Contribution, ρ (%)
m	1	3434	3434	3414.9	88.5
A	1	234	234	214.9	5.6
e	10	191	19.1	229.2	5.9
Total	12	3859		3859.0	100.0

in Table 29.1, an *analysis of variance* (*ANOVA*) *table*. The symbols in the table are defined as follows:

$$V_m = \frac{\text{CF}}{1} = \text{CF} \tag{29.41}$$

$$V_A = \frac{S_A}{1} = S_A \tag{29.42}$$

$$S'_m = \text{CF} - V_e \tag{29.43}$$

$$S'_A = S_A - (\text{degrees of freedom for } A)\,V_e = S_A - V_e \tag{29.44}$$

$$S'_e = S_e + 2V_e \tag{29.45}$$

$$\rho_m = \frac{S'_m}{S_T} \times 100 \tag{29.46}$$

$$\rho_A = \frac{S'_A}{S_T} \times 100 \tag{29.47}$$

$$\rho_e = \frac{S'_e}{S_T} \times 100 \tag{29.48}$$

The calculating procedures above constitute the analysis of variance. The results for Example 2 are shown in Table 29.2.