28 Fundamentals of Data Analysis

28.1. Introduction

In this chapter, the fundamentals of data analysis are illustrated by introducing the concepts of sum, mean, deviation, variation, and variance. This chapter is based on Genichi Taguchi et al., *Design of Experiments*. Tokyo: Japanese Standards Association, 1973.

28.2. Sum and Mean

The following data are the heights of 10 persons in centimeters:

163.2 171.6 156.3 159.2 160.0 158.7 167.5 175.4 172.8 153.5

Letting the total height of the 10 persons be T yeilds

$$T = 163.2 + 171.6 + \cdots + 153.5 = 1638.2 \tag{28.1}$$

Letting the mean be $\bar{y}$, we have

$$\bar{y} = \frac{1}{10} T = \frac{1638.2}{10} = 163.82 \tag{28.2}$$

To simplify the calculation, the sum and mean can be obtained for the data after subtracting a figure that is probably in the neighborhood of the mean. This number is called a *working mean*. Obviously, the numbers become smaller after

subtracting the working mean from each side; thus the calculation is simplified. Also for simplification, data after subtracting a working mean could be multiplied by 10, or 100, or 0.1, and so on.

This procedure is called *data transformation*. The height data after subtracting a working mean of 160.0 cm are:

$$3.2 \quad 11.6 \quad -3.7 \quad -0.8 \quad 0.0 \quad -1.3 \quad 7.5 \quad 15.4 \quad 12.8 \quad -6.5$$

The sum T and mean $\bar{y}$ are now calculated as follows:

$$T = (160.0)(10) + [3.2 + 11.6 + \cdots + (-6.5)] = 1600 + 38.2 = 1638.2 \quad (28.3)$$

$$\bar{y} = 160.0 + \frac{38.2}{10} = 163.82 \quad (28.4)$$

Letting n measurements by $y_1, y_2, \ldots, y_n$, their sum, T, is defined as

$$T = y_1 + y_2 + \cdots + y_n \quad (28.5)$$

The mean, $\bar{y}$, is

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \quad (28.6)$$

Letting the working mean be y_0, T and $\bar{y}$ may be written as

$$T = y_0 n + [(y_1 - y_0) + (y_2 - y_0) + \cdots + (y_n - y_0)] \quad (28.7)$$

$$\bar{y} = y_0 + \frac{1}{n}[(y_1 - y_0) + (y_2 - y_0) + \cdots + (y_n - y_0)] \quad (28.8)$$

28.3. Deviation

There are two types of deviations: deviation from an objective value and the deviation from the mean, $\bar{y}$. First, deviation from an objective value is explained. Let n measurements be $y_1, y_2, \ldots, y_n$. When there is a definite objective value, y_0, the differences of $y_1, y_2, \ldots, y_n$ from the objective value, y_0, are called the deviations from an objective value.

The following is an example to illustrate calculation of the deviation from an objective value. In order to produce carbon resistors valued at 10 kΩ, 12 resistors were produced for trial. Their resistance was measured to obtain the following values in kilohoms:

$$10.3 \quad 9.9 \quad 10.5 \quad 11.0 \quad 10.0 \quad 10.3 \quad 10.2 \quad 10.3 \quad 9.8 \quad 9.5 \quad 10.1 \quad 10.6$$

The deviations from an objective value of 10 kΩ are

$$0.3 \quad -0.1 \quad 0.5 \quad 1.0 \quad 0.0 \quad 0.3 \quad 0.2 \quad 0.3 \quad -0.2 \quad -0.5 \quad 0.1 \quad 0.6$$

An objective value is not always limited to the value that a person would assign arbitrarily. Sometimes standard values are used for comparison, such as nominal values (specifications or indications, such as 100 g per container), theoretical values (values calculated from theories or from a standard method used by a

company), forecast values, and values concerning the product used by other companies or other countries.

Table 28.1 shows the results of measuring the electric current of an electric circuit when both the voltage (E: volts) and resistance (R: ohms) were varied.

From Ohm's law, the theoretical values for each voltage and resistance are calculated as 5.00, 1.00, 2.00, 0.75, and 1.50. The differences between each observational value, y, and the theoretical value, y_0, are shown in the right-hand column of the table.

When the heights of sixth-grade children in a primary school of a certain school district are measured, deviations from the mean height of sixth-grade children in the entire country are usually calculated. In this case, the mean value of the country is an objective value; therefore, deviations from the objective value must be calculated.

Next, let's discuss the deviation from a mean value. When there is neither an objective value nor a theoretical value, it is important to calculate the deviation from a mean value. For example, the average height of the 10 persons described earlier was $\bar{y} = 163.82$. Deviations from the mean are

$$\begin{aligned} 163.2 - 163.82 &= -0.62 \\ 171.6 - 163.82 &= 7.78 \\ 156.3 - 163.82 &= -7.52 \\ &\vdots \\ 153.5 - 163.82 &= -10.32 \end{aligned} \tag{28.9}$$

The total of these deviations from the mean is equal to zero. In many cases, the word *deviation* signifies the deviation from an arithmetic mean. It is important to distinguish in each case whether a deviation is a deviation from a mean or from an objective value.

Again, using the 10 height data values, if these heights are data from 10 young men who live in a remote place, a deviation from an objective value should be calculated in addition to the deviation from the mean. The problem is to determine whether the mean value of the young men living in this remote place differs from the mean value of the entire country. Therefore, y_0, the deviation from the mean value of the country, is calculated. (*Note:* this is not the same as calculating the deviation from the mean or from an objective value.)

Table 28.1
Observational values y and theoretical values y_0

R	E	y	y_0	Deviation
10	50	5.10	5.00	0.10
20	20	0.98	1.00	−0.02
20	40	2.04	2.00	0.04
40	30	0.75	0.75	0.00
40	60	1.52	1.50	0.02

28.4. Variation and Variance

No matter whether a deviation is from a mean or from an objective value, it is plus, minus, or zero. When there are many deviations with different magnitudes, the magnitude as a whole is expressed by the sum of the deviations squared or by the mean of the sum squared.

The sum of the deviations squared is called *variation*, and its mean is called *variance*. In some books, the sum of the deviations squared is called the *sum of squares*. In using either variation or variance, the magnitude of the data change is obtained quantitatively.

Letting n measurements be $y_1, y_2, \ldots, y_n$ and their objective value be y_0, the sum of squares of $(y_1 - y_0)$, $(y_2 - y_0)$, is called the *total sum squared* (or *total variation*), denoted by S_T:

$$S_T = (y_1 - y_0)^2 + (y_2 - y_0)^2 + \cdots + (y_n - y_0)^2 \tag{28.10}$$

The number of independent squares in the total variation, S_T, is called the number of *degrees of freedom*. The number of degrees of freedom in equation (28.10) is therefore n.

The deviation of the resistance of carbon in Section 28.2 varied from -0.5 to $+1.0$. To express these different values by a simple figure, the sum of these deviations squared is calculated:

$$\begin{aligned} S_T &= (0.3)^2 + (-0.1)^2 + (0.5)^2 + \cdots + (0.6)^2 \\ &= 0.09 + 0.01 + 0.25 + \cdots + 0.36 \\ &= 2.23 \end{aligned} \tag{28.11}$$

The deviations in equation (28.11) are from an objective value of 10 kΩ; its number of degrees of freedom is 12. The total variation, S_T, of Table 28.1 is

$$\begin{aligned} S_T &= (0.10)^2 + (-0.02)^2 + (0.04)^2 + (0.00)^2 + (0.02)^2 \\ &= 0.0124 \end{aligned} \tag{28.12}$$

Its number of degrees of freedom is 5.

Let the mean of n measurements be $\bar{y}$. The sum of the deviation squared from the mean would correctly be called the sum of deviations squared from the mean, but it is usually called the *total variation* (or the *total sum squared*).

$$S_T = (\bar{y}_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 \tag{28.13}$$

There are n squares in equation (28.13); however, its number of degrees of freedom is $n - 1$. This is because there exist the following linear relationships among n deviations:

$$(y_1 - \bar{y}) + (y_2 - \bar{y}) + \cdots + (y_n - \bar{y}) = 0 \tag{28.14}$$

Generally, the number of degrees of freedom for the sum of n number of squares is given by n minus the number of relational equations among the deviations before these deviations are squared.

There are n squares in the S_T equation (28.13), but there is a relational equation shown by (28.14); therefore, its number of degrees of freedom is $n - 1$.

When there are n measurements, the number of degrees of freedom of the sum of the deviations squared either from an objective value or from a theoretical value is n, and the number of degrees of freedom of the sum of the deviations squared from the mean, $\bar{y}$, is $n - 1$.

The sum of the deviations in equation (28.9) squared, denoted by S_T,

$$\begin{aligned} S_T &= (-0.62)^2 + 7.78^2 + (-7.52)^2 + \cdots + (-10.32)^2 \\ &= 0.3844 + 60.5284 + 56.5504 + \cdots + 106.5204 \\ &= 514.9360 \end{aligned} \tag{28.15}$$

is the sum of deviations (from the mean $\bar{y}$) squared with nine degrees of freedom.

To obtain a variation from the arithmetic mean, $\bar{y}$, it is easier to calculate by way of subtracting a working mean from $y_1, y_2, \ldots, y_n$, calculating their sum squared, then subtracting the correction factor:

$$\begin{aligned} S_T &= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 \\ &= (y_1')^2 + (y_2')^2 + \cdots \\ &\quad + (y_n')^2 - \frac{(y_1' + y_2' + \cdots + y_n')^2}{n} \end{aligned} \tag{28.16}$$

The measurements $y_1', y_2', \ldots, y_n'$ are the values subtracted by a working mean. It is relatively easy to prove equation (28.16). In equation (28.16),

$$\text{CF} = \frac{(y_1' + y_2' + \cdots + y_n')^2}{n} \tag{28.17}$$

is called the *correction factor*, usually denoted by CF or CT.

The total variation of the heights is calculated using equation (28.16) with a working mean of 160.0 cm:

$$\begin{aligned} S_T &= (\text{sum of the data after subtracting the working mean squared}) \\ &\quad - (\text{sum of the data after subtracting the working mean})^2 \\ &= (163.2 - 160.0)^2 + (171.6 - 160.0)^2 + \cdots + (153.5 - 160.0)^2 \\ &\quad - \frac{[(163.2 - 160.0) + (171.6 - 160.0) + \cdots + (153.5 - 160.0)]^2}{10} \\ &= 3.2^2 + 11.6^2 + \cdots + (-6.5)^2 - \frac{[3.2 + 11.6 + \cdots + (-6.5)]^2}{10} \\ &= 660.32 - \frac{38.2^2}{10} \\ &= 660.32 - 145.924 \\ &= 514.396 \end{aligned} \tag{28.18}$$

Usually, a correction factor or a variation is calculated as far as the lowest unit after squaring its original data. In this case, the calculation is made to the second decimal place, so the third decimal place and the lower units are rounded.

$$S_T = 660.32 - 145.924 = 514.40 \qquad (f = 9) \tag{28.19}$$

The correction factor is

$$\text{CF} = \frac{38.2^2}{10} = 145.92 \tag{28.20}$$

Next, we discuss *variance*, the value of variation divided by the degrees of freedom:

$$\text{variance} = \frac{\text{variation}}{\text{degrees of freedom}} \tag{28.21}$$

Variance signifies the mean of deviation per unit degree of freedom. It really means the magnitude of error, or mistake, or dispersion, or change per unit. This measure corresponds to work or energy per unit time in physics.

Letting $y_1, y_2, \ldots, y_n$ be the results of n measurements, and letting an objective value be y_0, the variance of these measurements is

$$V = \frac{(y_1 - y_0)^2 + (y_2 - y_0)^2 + \cdots + (y_n - y_0)^2}{n} \tag{28.22}$$

The number of degrees of freedom of total variation S_T in equation (28.11) is 12; then its variance is

$$\begin{aligned} V &= \frac{S_T}{\text{degrees of freedom}} \\ &= \frac{2.23}{12} \\ &= 0.186 \end{aligned} \tag{28.23}$$

Sometimes the value above is called the *error variance* with 12 degrees of freedom. It is recommended that one carry out this calculation to one more decimal place. This is shown in equation (28.23), which is calculated to the third decimal place.

When the variation is calculated as the sum of the deviations squared from the arithmetical mean, as in the case of height, variance is obtained as variation divided by the degrees of freedom. Letting the arithmetical mean of $y_1, y_2, \ldots, y_n$ be $\bar{y}$, the variance, V, is then

$$\begin{aligned} V &= \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{(n - 1)} \\ &= \frac{S_T}{n - 1} \end{aligned} \tag{28.24}$$

In cases when there is an objective value of a theoretical value, deviations of each datum from the objective or from the theoretical value are calculated. For example, some carbon resistances were produced with an objective value of y_0 kilohms. In this case, $(y_1 - y_0)$, $(y_2 - y_0)$, ... , $(y_n - y_0)$ are discussed. Another example would be the differences between the time indicated by a certain watch and the actual time. The data would then show the time that the watch either gained or lost.

Letting the deviations from an objective or a theoretical value be $y_1, y_2, \ldots, y_n$, and the mean be $\bar{y}$,

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} \tag{28.25}$$

This would then be called *deviation,* or more correctly, the *estimate of deviation.* When the initial observational values before subtracting an objective or a theoretical value are always larger than the objective value, deviation $\bar{y}$ in (28.25) would be a large positive value, or vice versa.

In the case of carbon resistance discussed earlier, the deviations from the objective value are

$$0.3 \quad -0.1 \quad 0.5 \quad 1.0 \quad 1.0 \quad 0.0 \quad 0.3 \quad 0.2 \quad 0.3 \quad -0.2 \quad -0.5 \quad 0.1 \quad 0.6$$

The deviation $\bar{y}$ is

$$\bar{y} = \frac{1}{12}(0.3 - 0.1 + 0.5 + \cdots + 0.6) = 2.5$$
$$= 0.208 \tag{28.26}$$

It is also recommended that one calculate $\bar{y}$ to one or two more decimal places.

A deviation that shows a plus value indicates that the resistance values of carbon are generally larger than the objective value. To express the absolute magnitude of a deviation, no matter whether it is plus or minus, $\bar{y}$ is squared and then multiplied by n:

$$\text{magnitude of deviation} = n(\bar{y})^2$$
$$= n\frac{(y_1 + y_2 + \cdots + y_n^2}{n} \tag{28.27}$$

This value signifies the variation of the mean, m, of $y_1, y_2, \ldots, y_n$, which are the deviations from an objective value or from a theoretical value.

The following decomposition is therefore derived:

$$y_1^2 + y_2^2 + \cdots + y_n^2 = \frac{(y_1 + y_2 + \cdots + y_n)^2}{n}$$
$$+ [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2] \tag{28.28}$$

The relational equation above is written as

$$\text{sum of data squared} = (\text{variation of the mean of deviations}) + (\text{sum of the deviations from the mean squared}) \tag{28.29}$$

(*Note:* When the differences between an objective value or from a theoretical value are discussed, and letting these differences be $y_1, y_2, \ldots, y_n$, the variation of the mean of deviations must not be calculated from the data that were subtracted by a working mean. The variation of the mean of deviations must be calculated from the differences between the original data and either an objective value or a theoretical value.)

If the sum of the carbon resistances, S_T, in (28.11) is 2.23, then S_m, the variation of the mean (m), is

$$S_m = \frac{(0.3 - 0.1 + \cdots + 0.6)^2}{12}$$

$$= \frac{2.5^2}{12}$$

$$= 0.52 \tag{28.30}$$

Therefore, the equation

$$2.23 = 0.52 + (\text{sum of deviations from mean } \bar{y} \text{ squared}) \tag{28.31}$$

is calculated. The sum of deviations from the mean squared, S_e, is given as

$$S_e = S_T - S_m = 2.23 - 0.52 = 1.71 \tag{28.32}$$

S_e is called the *error variation*. The number of degrees of freedom of error variation is $12 - 1 = 11$.

Error variation divided by its degrees of freedom is denoted by V_e. Using the example of resistance,

$$V_e = \frac{S_e}{\text{degrees of freedom}} = \frac{1.71}{11} = 0.155 \tag{28.33}$$

where V_e would signify the extent of dispersion of individual carbon resistances, a large V_e value means a large difference between individual products. On the other hand, a large variation of mean, or a large S_m, shows that the mean value of the products is considerable apart from the objective value.

It is easy in many cases to reduce S_m, the variation of a mean, when it is large. In the case of carbon resistance, for example, it is possible to reduce the magnitude, either by prolonging the time of carbonation or by adjusting the cutting method for carbon.

It is the same for some other dimensional characteristics; such deviations are easily adjustable. However, many contrivances are required in order to reduce dispersion. In the case of height, the total variation, S_T, is calculated as the variation of deviations from the mean, since there is neither objective value nor theoretical value in height. Sometimes zero is taken as a theoretical value. In such case, let $y_1, y_2, \ldots, y_n$ be the differences from zero. The magnitude of deviations, S_m, is calculated from $y_1, y_2, \ldots, y_n$.

The relationship in equation (28.29) corresponds to a concept such as the electric current of the telephone, where direct current and alternating current are mixed. This would be expressed as follows:

$$\text{power of the total current} = (\text{square of the dc voltage}) + (\text{ac power})$$

$\bar{y}$ is the dc voltage (48 V in the case of a telephone), and the ac power is equal to the square of the ac voltage. It is very important to recognize that there is no additivity in voltage, but there is additivity in power, which is the square of voltage. Such decomposition is called *power spectrum analysis* in engineering.

Such concepts become easier to understand as one uses such calculations. In the case of telephone current, the dc voltage has nothing to do with the transmission of voice, whereas the ac voltage does.

When we discuss deviations from an objective or a theoretical value, the variation of mean (S_m) and the remaining part of variation are entirely different, both in technical meaning and in countermeasures that are to be taken. It is very important to distinguish between the two.