Chapter 2
Intravariable Statistics

This chapter begins with expressing data sets by matrices. Then, we introduce two statistics (statistical indices): average and variance, where the *average* is an index value that represents scores and the *variance* stands for how widely scores disperse. Further, how the original scores are transformed into *centered* and *standard scores* using the average and variance is described.

As the statistics in this chapter summarize the scores *within* a variable, the chapter is named *intra*variable statistics, in contrast to the immediately following chapter entitled *inter*-variable statistics, where the statistics *between* variables would be treated.

2.1 Data Matrices

A *multivariate data* set refers to a set of values arranged in a table whose rows and columns are individuals and variables, respectively. This is illustrated in each panel of Table 2.1. Here, the term "*individuals*" implies the sources from which data are obtained; for example, individuals are participants, cities, and baseball teams, respectively, in panels (A), (B), and (C) of Table 2.1. On the other hand, the term "*variables*" refers to the indices or items for which individuals are measured; for example, variables are Japanese, mathematics, English, and sciences in Table 2.1 (A). By attaching "multi" to "variate," which is a synonym of "variable," we use the adjective "*multivariate*" for the data sets with multiple variables, as shown in Table 2.1. On the other hand, data with a single variable are called "*univariate data*".

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
DOI 10.1007/978-981-10-2341-5_2

Table 2.1 Three examples of multivariate data

(A) Test scores (Artificial example)

Participant	Item			
	Japan	Mathematics	English	Science
1	82	70	70	76
2	96	65	67	71
3	84	41	54	65
4	90	54	66	80
5	93	76	74	77
6	82	85	60	89

(B) Weather in cities in January (http://www2m.biglobe.ne.jp/ZenTech/world/kion/Japan/Japan.htm)

City	Weather		
	Min °C	Max °C	Precipitation
Sapporo	−7.7	−0.9	110.7
Tokyo	2.1	9.8	48.6
.	.	.	.
.	.	.	.
.	.	.	.
Naha	14.3	19.1	114.5

(C) Team scores (2005 in Japan) (http://npb.jp/bis/2005/stats/)

Team	Averages				
	Win %	Runs	HR	Avg.	ERA
Tigers	0.617	731	140	0.274	3.24
Dragons	0.545	680	139	0.269	4.13
BayStars	0.496	621	143	0.265	3.68
Swallows	0.493	591	128	0.276	4.00
Giants	0.437	617	186	0.260	4.80
Carp	0.408	615	184	0.275	4.80

Let us express a data set as an n-individuals × p-variables matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p\right], \tag{2.1}$$

whose jth column

Table 2.2 Raw, centered, and standard scores of tests with their averages, variances, and standard deviations (SD) (artificial example)

Student	(A) Raw		(B) Centered		(C) Standard	
	History	Mathematics	History	Mathematics	History	Mathematics
1	66	74	5	−3	0.52	−0.20
2	72	98	11	21	1.15	1.43
3	44	62	−17	−15	−1.78	−1.02
4	58	88	−3	11	−0.31	0.75
5	70	56	9	−21	0.94	−1.43
6	56	84	−5	7	−0.52	0.48
Average	61.0	77.0	0	0	0	0
Variance	91.67	214.33	91.67	214.33	1.00	1.00
SD	9.57	14.64	9.57	14.64	1.00	1.00

$$\mathbf{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} = \left[x_{1j}, \ldots, x_{nj}\right]' \tag{2.2}$$

stands for the jth *variable*. Examples of (2.1) are given in Table 2.1(A), (B), (C). A different example is presented in Table 2.2(A), where n-individuals and p-variables are 6 students and 2 items, respectively, with x_{ij} the score of individual i for item j and $\mathbf{x}_j$ the 6×1 vector containing the scores on the jth variable:

$$\underset{6\times 2}{\mathbf{X}} = \begin{bmatrix} 66 & 74 \\ 72 & 98 \\ \vdots & \vdots \\ 56 & 84 \end{bmatrix} \quad \text{with} \quad \mathbf{x}_1 = \begin{bmatrix} 66 \\ 72 \\ \vdots \\ 56 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 74 \\ 98 \\ \vdots \\ 84 \end{bmatrix}.$$

The scores in Table 2.2(B) and (C) will be explained later, in Sects. 2.4 and 2.6.

2.2 Distributions

The distribution of the six students' scores for each variable in Table 2.2(A) is graphically depicted in Fig. 2.1, where those scores are plotted on lines extending from 0 to 100. The distributions allow us to intuitively recognize that [1] their scores in history are lower on *average* than those in mathematics, and [2] the scores *disperse* more widely in mathematics than in history. The statistics related to [1] and [2] are introduced in Sects. 2.3 and 2.5, respectively.

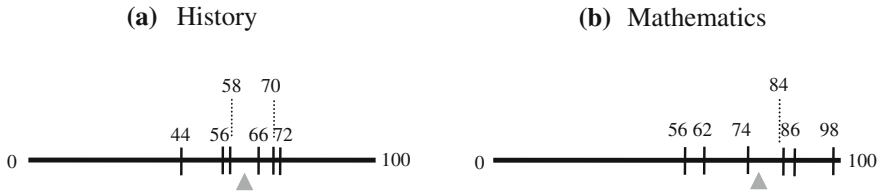

Fig. 2.1 Distributions of the test scores in Table 2.2(A)

2.3 Averages

Let us consider summarizing n scores into a single statistic. The most popular statistic for the summary is the *average*, which is defined as:

$$\bar{x}_j = \frac{1}{n}\left(x_{1j} + \cdots + x_{nj}\right) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{2.3}$$

for variable j, i.e., the jth column of $\mathbf{X}$. For example, the average score in mathematics ($j = 2$) in Table 2.2(A) is $\bar{x}_2 = (74 + 98 + 62 + 88 + 56 + 84)/6 = 77.0$. The average can be rewritten, using the $n \times 1$ *vector of ones* $\mathbf{1}_n = [1,1,\ldots, 1]'$ defined in (1.35): The *sum* $x_{1j} + \cdots + x_{nj}$ is expressed as:

$$\mathbf{1}_n'\mathbf{x}_j = [1, \ldots, 1]\begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}, \tag{2.4}$$

thus, the *average* (2.3) is also simply expressed as:

$$\bar{x}_j = \frac{1}{n}\mathbf{1}_n'\mathbf{x}_j, \tag{2.5}$$

without using the complicated "Sigma" symbol. For example, the average score in history ($j = 1$) in Table 2.2(A) is expressed as $6^{-1}\mathbf{1}_6'\mathbf{x}_1$ with $\mathbf{x}_1 = [66, 72, 44, 58, 70, 56]'$. The resulting average is $6^{-1}\mathbf{1}_6'\mathbf{x}_1 = 61.0$.

2.4 Centered Scores

The raw scores minus their average are called *centered scores* or *deviations from average*. Let the centered score vector for variable j be denoted as $\mathbf{y}_j = [y_{1j}, \ldots, y_{nj}]'$ ($n \times 1$), which is expressed as:

$$\mathbf{y}_j = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{bmatrix} = \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{x}_j - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix}. \tag{2.6}$$

In Table 2.2(B), the centered data for (A) are shown: The centered scores [5, 11, …, −5]′ for history are given by subtracting 61 from all elements of [66, 72, …, 56]′ and the centered scores for mathematics are given by subtracting 77 in a parallel manner.

Here, we rewrite (2.6) in a simpler form. First, let us note that all elements of the subtracted vector $[\bar{x}_j, \ldots, \bar{x}_j]'$ in (2.6) are equal to an average $\bar{x}_j$, and thus, that vector can be written as:

$$\begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \bar{x}_j \mathbf{1}_n = \mathbf{1}_n \times \bar{x}_j, \tag{2.7}$$

where we have used (1.20). Substituting (2.5) into $\bar{x}_j$ in (2.7), it is rewritten as:

$$\begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{1}_n \times \left(\frac{1}{n} \mathbf{1}_n' \mathbf{x}_j \right) = \frac{1}{n} \mathbf{1}_n \left(\mathbf{1}_n' \mathbf{x}_j \right) = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \mathbf{x}_j \tag{2.8}$$

Here, we have made use of the fact that "× scalar (n^{-1})" can be moved and $\mathbf{A}(\mathbf{BC}) = \mathbf{ABC}$ generally holds for matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, which implies $\mathbf{1}_n(\mathbf{1}_n'\mathbf{x}_j) = \mathbf{1}_n\mathbf{1}_n'\mathbf{x}_j$. Using (2.8) in (2.6) and noting property (1.44) for an identity matrix, the *centered score vector* (2.6) can be rewritten as:

$$\mathbf{y}_j = \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \mathbf{x}_j - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{I}_n \mathbf{x}_j - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \mathbf{x}_j = \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) \mathbf{x}_j = \mathbf{J} \mathbf{x}_j, \tag{2.9}$$

where $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ and we have made use of the fact that $\mathbf{BC} + \mathbf{EC} = (\mathbf{B} + \mathbf{E})\mathbf{C}$ holds for matrices $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{E}$. The matrix $\mathbf{J}$ has a special name and important properties:

Note 2.1. Centering Matrix
It is defined as

$$\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'. \tag{2.10}$$

The centering matrix has the following properties:

$$\mathbf{J} = \mathbf{J}' \text{ (symmetric)} \tag{2.11}$$

$$\mathbf{J}^2 = \mathbf{J}\mathbf{J} = \mathbf{J} \tag{2.12}$$

$$\mathbf{1}_n'\mathbf{J} = \mathbf{0}_n' \tag{2.13}$$

Equation (2.11) can easily be found. Equations (2.12) and (2.13) can be proved as:

$$\begin{aligned}\mathbf{J}\mathbf{J} &= (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n') = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n' - n^{-1}\mathbf{1}_n\mathbf{1}_n' + n^{-2}\mathbf{1}_n\mathbf{1}_n'\mathbf{1}_n\mathbf{1}_n' \\ &= \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n' - n^{-1}\mathbf{1}_n\mathbf{1}_n' + n^{-2}\mathbf{1}_n(n)\mathbf{1}_n' = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'\end{aligned}$$

and

$$\mathbf{1}_n'(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n') = \mathbf{1}_n' - n^{-1}\mathbf{1}_n'\mathbf{1}_n\mathbf{1}_n' = \mathbf{1}_n' - n^{-1}(n)\mathbf{1}_n' = \mathbf{0}_n',$$

respectively, where $\mathbf{1}_n'\mathbf{1}_n = n$ has been used.

Equations (2.12) and (2.13) further lead to the following important facts:

Note 2.2. Matrices Premultiplied by the Centering Matrix
A matrix $s\mathbf{J}\mathbf{A}$ with $\mathbf{A}$ an $n \times p$ matrix and s a scalar satisfies:

$$\mathbf{1}_n'(s\mathbf{J}\mathbf{A}) = s\mathbf{1}_n'\mathbf{J}\mathbf{A} = \mathbf{0}_p', \tag{2.14}$$

$$\mathbf{J}(s\mathbf{J}\mathbf{A}) = s\mathbf{J}\mathbf{J}\mathbf{A} = s\mathbf{J}\mathbf{A}. \tag{2.15}$$

When $\mathbf{A}$ is an $n \times 1$ vector $\mathbf{a}$, those equations are rewritten as $\mathbf{1}_n'(s\mathbf{J}\mathbf{a}) = 0$ and $\mathbf{J}(s\mathbf{J}\mathbf{a}) = s\mathbf{J}\mathbf{a}$, respectively.

Comparing (2.9) with (2.14), we can find that the sum and average of centered scores are always *zero*:

$$\mathbf{1}_n'\mathbf{y}_j = \frac{1}{n}\mathbf{1}_n'\mathbf{y}_j = 0. \tag{2.16}$$

This is shown in the row named "Average" in Table 2.1(B). Figure 2.3(B) (*on a later page*) illustrates (2.16); the centered scores are distributed with their average being the zero which is a *center between negative and positive values*. This property provides the name "centered scores," and the transformation of raw scores into centered ones is called *centering*. Comparing (2.9) with (2.15), we also find:

$$\mathbf{J}\mathbf{y}_j = \mathbf{y}_j. \tag{2.17}$$

The centered score vector, premultiplied by the centering matrix, remains unchanged.

2.5 Variance and Standard Deviation

The locations of averages in the distributions of scores are indicated by triangles in Fig. 2.1, which do not stand for how widely scores disperse. The most popular statistic for indicating dispersion is *variance*. It is defined using the *sum of squared distances* between *scores* and *their average*, which is illustrated in Fig. 2.2. Denoting the variance for variable j as v_{jj}, it is formally expressed as:

$$\begin{aligned} v_{jj} &= \frac{1}{n}\left\{|x_{1j} - \bar{x}_j|^2 + \cdots + |x_{nj} - \bar{x}_j|^2\right\} \\ &= \frac{1}{n}\left\{(x_{1j} - \bar{x}_j)^2 + \cdots + (x_{nj} - \bar{x}_j)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \end{aligned} \tag{2.18}$$

where the same subscript j is used twice as v_{jj}, for the sake of accordance with the related statistic introduced in the next chapter. The variance of the scores for mathematics in Table 2.2(A) is obtained as $6^{-1}\{(74 - 77)^2 + (98 - 77)^2 + \cdots + (84 - 77)^2\} = 214.33$, for example.

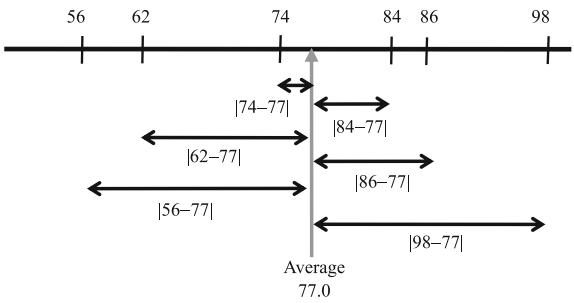

Fig. 2.2 Distances of scores to their average, which are squared, summed, and divided by n, to give the variance of the mathematics scores in Table 2.2(A)

To express (2.18) in *vector* form, we should note that it can be rewritten as:

$$v_{jj} = \frac{1}{n}\left[x_{1j} - \bar{x}_j, \ldots, x_{nj} - \bar{x}_j\right] \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix}. \tag{2.19}$$

Comparing (2.19) with $\begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \mathbf{J}\mathbf{x}_j$ in (2.9), the *variance* (2.18) or (2.19) is expressed as:

$$v_{jj} = \frac{1}{n}\left(\mathbf{J}\mathbf{x}_j\right)'\mathbf{J}\mathbf{x}_j = \frac{1}{n}||\mathbf{J}\mathbf{x}_j||^2 = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j, \tag{2.20}$$

where (1.12), (2.11), and (2.12) have been used. Further, we can use (2.9) in (2.20) to rewrite it as:

$$v_{jj} = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_j = \frac{1}{n}||\mathbf{y}_j||^2. \tag{2.21}$$

The variance of raw scores is expressed using their *centered score vector* simply as $n^{-1}\|\mathbf{y}_j\|^2$. We can also find in (2.20) and (2.21) that the variance is the squared length of vector $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ divided by n.

How is the variance of the centered scores (rather than raw scores) expressed? To find this, we substitute the centered score vector $\mathbf{y}_j$ for $\mathbf{x}_j$ in the variance (2.20). Then, we use (2.17) and (2.9) to get:

$$\frac{1}{n}\mathbf{y}_j'\mathbf{J}'\mathbf{J}\mathbf{y}_j = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j, \tag{2.22}$$

which is equal to (2.20); the variance of the centered scores equals that for their raw scores.

The square root of variance

$$\sqrt{v_{jj}} = \sqrt{\frac{1}{n}\mathbf{x}'_j\mathbf{J}\mathbf{x}_j} = \frac{1}{\sqrt{n}}||\mathbf{J}\mathbf{x}_j|| = \frac{1}{\sqrt{n}}||\mathbf{y}_j|| \tag{2.23}$$

is called the *standard deviation*, which is also used for reporting the dispersion of data. We can find in (2.23) that the standard deviation is the length of vector $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ divided by $n^{1/2}$.

2.6 Standard Scores

The centered scores (i.e., the raw scores minus their average) divided by their standard deviation are called *standard scores* or *z-scores*. Let the standard score vector for variable j be denoted by $\mathbf{z}_j = [z_{1j}, \ldots, z_{nj}]'$, which is expressed as:

$$\mathbf{z}_j = \begin{bmatrix} (x_{1j} - \bar{x}_j)/\sqrt{v_{jj}} \\ \vdots \\ (x_{nj} - \bar{x}_j)/\sqrt{v_{jj}} \end{bmatrix} = \frac{1}{\sqrt{v_{jj}}} \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \frac{1}{\sqrt{v_{jj}}} \mathbf{J}\mathbf{x}_j = \frac{1}{\sqrt{v_{jj}}} \mathbf{y}_j, \tag{2.24}$$

where we have used (2.9). In Table 2.2(C), the standard scores for (A) are shown; the standard scores $[-0.2, \ldots, 0.48]'$ for mathematics are given by dividing its centered scores (B) by 14.64. Transforming raw scores into standard ones is called *standardization*.

Standard scores have two important properties. One is that the sum and average of standard scores are always *zero*, as are those of centered scores:

$$\mathbf{1}_n' \mathbf{z}_j = \frac{1}{n} \mathbf{1}_n' \mathbf{z}_j = 0, \tag{2.25}$$

which follows from (2.16) and (2.24). The other property is that the variance of standard scores is always *one*, which is shown as follows: The substitution of $\mathbf{z}_j$ into $\mathbf{x}_j$ in (2.20) leads to the variance of standard scores being expressed as $n^{-1}\mathbf{z}_j'\mathbf{J}'\mathbf{J}\mathbf{z}_j = n^{-1}\mathbf{z}_j'\mathbf{z}_j$, where we have used $\mathbf{z}_j = \mathbf{J}\mathbf{z}_j$, following from the use of (2.17) in (2.24). Further, the variance can be rewritten, using (1.12), (2.21), and (2.24), as:

$$\frac{1}{n}\mathbf{z}_j'\mathbf{J}'\mathbf{J}\mathbf{z}_j = \frac{1}{n}\mathbf{z}_j'\mathbf{z}_j = \frac{1}{n}||\mathbf{z}_j||^2 = \frac{1}{nv_{jj}}\mathbf{y}_j'\mathbf{y}_j = \frac{n}{n\|\mathbf{y}_j\|^2}\mathbf{y}_j'\mathbf{y}_j = 1. \tag{2.26}$$

This also implies that the length of every standard score vector is always $\|\mathbf{z}_j\| = n^{1/2}$.

2.7 What Centering and Standardization Do for Distributions

The properties of centered and standard scores shown with (2.16), (2.22), (2.25), and (2.26) are summarized in Table 2.3.

Table 2.3 Averages and variances of centered and standard scores

	Average	Variance
Centered scores	0	Variance of raw scores
Standard scores	0	1

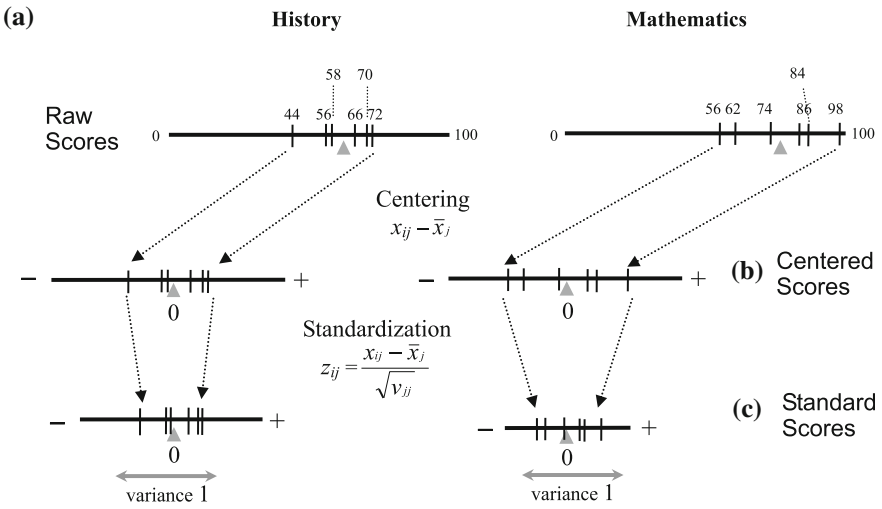

Fig. 2.3 Distributions of raw, centered, and standard scores in Table 2.2

Figure 2.3 illustrates the roles that centering and standardization (i.e., transforming raw scores into centered and standard ones) perform for the distributions of data: *Centering* simply *moves the distributions* of raw scores so that the average of the moved distributions is zero, and *standardization* further *accommodates the scale* of the moved distributions so that their variances are equal to one. The standard scores are unified among different variables so that the averages and variances are zero and one, respectively; thus, the greatness/smallness of the standard scores can be compared reasonably between variables.

2.8 Matrix Representation

We will now introduce a basic formula in matrix algebra:

Note 2.3. A Property of Matrix Product
If $\mathbf{A}$ is a matrix of $n \times m$ and $\mathbf{b}_1, \ldots, \mathbf{b}_K$ are $m \times 1$ vectors, then

$$[\mathbf{A}\mathbf{b}_1, \ldots, \mathbf{A}\mathbf{b}_K] = \mathbf{A}[\mathbf{b}_1, \ldots, \mathbf{b}_K]. \tag{2.27}$$

Using this and (2.5), the $1 \times p$ row vector containing the *averages* of p-variables is expressed as:

$$\left[\bar{x}_1, \ldots, \bar{x}_p\right] = \left[\frac{1}{n}\mathbf{1}'_n\mathbf{x}_1, \ldots, \frac{1}{n}\mathbf{1}'_n\mathbf{x}_p\right] = \frac{1}{n}\mathbf{1}'_n\left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] = \frac{1}{n}\mathbf{1}'_n\mathbf{X}. \tag{2.28}$$

For example, when $\mathbf{X}$ consists of the six students' scores in Table 2.2(A), $6^{-1}\mathbf{1}'_6\mathbf{X} = [61.0, 71.0]$.

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_p]$ denote the $n \times p$ matrix of *centered scores* whose jth column is defined as (2.9) for the corresponding column of $\mathbf{X}$. Then, we can use (2.9) and (2.27) to express $\mathbf{Y}$ as:

$$\mathbf{Y} = \left[\mathbf{J}\mathbf{x}_1, \ldots, \mathbf{J}\mathbf{x}_p\right] = \mathbf{J}\left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] = \mathbf{J}\mathbf{X}, \tag{2.29}$$

an example of which is presented in Table 2.2(B).

Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ denote the $n \times p$ matrix of *standard scores* whose jth column is defined as (2.24) for the corresponding columns of $\mathbf{X}$ and $\mathbf{Y}$. Then, $\mathbf{Z}$ is expressed as:

$$\mathbf{Z} = \left[\frac{1}{\sqrt{v_{11}}}\mathbf{y}_1, \ldots, \frac{1}{\sqrt{v_{pp}}}\mathbf{y}_p\right] = \left[\mathbf{y}_1, \ldots, \mathbf{y}_p\right]\begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{v_{pp}}} \end{bmatrix} = \mathbf{Y}\mathbf{D}^{-1}. \tag{2.30}$$

Here, the blanks in $\begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{v_{pp}}} \end{bmatrix}$ stand for the corresponding elements being zeros and $\mathbf{D} = \begin{bmatrix} \sqrt{v_{11}} & & \\ & \ddots & \\ & & \sqrt{v_{pp}} \end{bmatrix}$ is the $p \times p$ diagonal matrix whose diagonal elements are the standard deviations for p variables: We should recall (1.42) to notice that $\mathbf{D}^{-1}$ is the diagonal matrix whose diagonal elements are the reciprocals of the standard deviations. Those readers who have difficulties in understanding (2.30) should note the following simple example with $\mathbf{Y}$ being 3×2:

$$\mathbf{Y}\mathbf{D}^{-1} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & \\ & \frac{1}{\sqrt{v_{22}}} \end{bmatrix} = \begin{bmatrix} y_{11}/\sqrt{v_{11}} & y_{12}/\sqrt{v_{22}} \\ y_{21}/\sqrt{v_{11}} & y_{22}/\sqrt{v_{22}} \\ y_{31}/\sqrt{v_{11}} & y_{23}/\sqrt{v_{22}} \end{bmatrix}, \tag{2.31}$$

which illustrates the equalities in (2.30) in the reverse order. The standard score matrix $\mathbf{Z}$ can also be expressed as:

$$\mathbf{Z} = \mathbf{J}\mathbf{X}\mathbf{D}^{-1}, \tag{2.32}$$

using (2.29) in (2.30).

Table 2.4 Data matrix $\mathbf{X}$ of 5 persons × 3 variables

Person	Height	Weight	Sight
Bill	172	64	0.8
Brian	168	70	1.4
Charles	184	80	1.2
Keith	176	64	0.2
Michael	160	62	1.0

2.9 Bibliographical Notes

Carroll et al. (1997, Chap. 3) and Reyment and Jöreskog (1996, Chap. 2) are among the literature in which the matrix expressions of intravariable statistics are intelligibly treated.

Exercises

2.1. Compute $\mathbf{J} = \mathbf{I}_5 - 5^{-1}\mathbf{1}_5\mathbf{1}_5'$ and obtain the centered score matrix $\mathbf{Y} = \mathbf{JX}$ for the 5 × 3 matrix $\mathbf{X}$ in Table 2.4.

2.2. Compute the variance $v_{jj} = 5^{-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j$ $(j = 1, 2, 3)$, the diagonal matrix $\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \frac{1}{\sqrt{v_{22}}} & \\ & & \frac{1}{\sqrt{v_{33}}} \end{bmatrix}$, and the standard score matrix $\mathbf{Z} = \mathbf{JXD}^{-1}$ for $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$ (5 × 3) in Table 2.4.

2.3. Discuss the benefits of standardizing the data in Table 2.4.

2.4. If the average for each column of $\mathbf{Y}$ $(n \times p)$ is zero, show that the average for each column of $\mathbf{YA}$ is also zero.

2.5. Let $\mathbf{Z}$ be an n-individuals × p-variables matrix containing standard scores. Show that $\|\mathbf{Z}\|^2 = \mathrm{tr}\mathbf{Z}'\mathbf{Z} = \mathrm{tr}\mathbf{Z}\mathbf{Z}' = np$.

2.6. Let $\mathbf{x}$ be an $n \times 1$ vector with $v = n^{-1}\mathbf{x}'\mathbf{J}\mathbf{x}$ the variance of the elements in $\mathbf{x}$. Show that the variance of the elements in $b\mathbf{x} + c\mathbf{1}_n$ is b^2v.

2.7. Let $\mathbf{y} = [y_1, \ldots, y_n]'$ contain centered scores. Show that the average of the elements in $-\mathbf{y} + c\mathbf{1}_n = [-y_1 + c, \ldots, -y_n + c]'$ is c, and their variance is equivalent to that for $\mathbf{y}$.

2.8. Let $\mathbf{z} = [z_1, \ldots, z_n]'$ contain standard scores. Show that the average of the elements in $b\mathbf{z} + c\mathbf{1}_n = [bz_1 + c, \ldots, bz_n + c]'$ is c, and their standard deviation is b.

http://www.springer.com/978-981-10-2340-8

Matrix-Based Introduction to Multivariate Data Analysis

Adachi, K.

2016, XIII, 301 p. 55 illus., 8 illus. in color., Hardcover
ISBN: 978-981-10-2340-8