Chapter 2: Descriptive Statistics

Prerequisite: Chapter 1

2.1 *Review of Univariate Statistics*

The central tendency of a more or less symmetric distribution of a set of interval, or higher, scaled scores, is often summarized by the *arithmetic mean*, which is defined as

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \,. \tag{2.1}$$

We can use the mean to create a *deviation score*,

$$d_i = x_i - \bar{x}, \tag{2.2}$$

so named because it quantifies the deviation of the score from the mean.

Deviation is often measured by squaring, since it equates negative and positive deviations. The sum of squared deviations, usually just called the *sum of squares*, is given by

$$\begin{aligned} a &= \sum_{i}^{n} (x_i - \bar{x})^2 \text{ or} \\ &= \sum_{i}^{n} d_i^2 \,. \end{aligned} \tag{2.3}$$

Another method of calculating the sum of squares was frequently used during the era that preceded computers when students would work with calculating machines,

$$a = \sum_{i}^{n} x_i^2 - \frac{\left(\sum_{i}^{n} x_i\right)^2}{n}. \tag{2.4}$$

Regardless whether one uses Equation (2.3) or Equation (2.4), the amount of deviation that exists around the mean in a set of scores can be averaged using the *standard deviation*, or its square, the *variance*. The variance is just

$$s^2 = \frac{1}{n-1} a$$

with s being the positive square root of s^2.

We can take the deviation scores and standardize them, creating, well; *standardized scores*:

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{d_i}{s} \,. \tag{2.5}$$

Next, we define a very important concept, that of the *covariance* of two variables, in this case x and y. The covariance between x and y may be written Cov(x, y). We have

$$s_{xy} = \frac{1}{n-1}\left[\sum_i^n x_i y_{i-} \frac{\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)}{n}\right]$$

$$= \frac{1}{n-1}\sum_i^n d_{x_i} d_{y_i} \,, \tag{2.6}$$

where the d_{x_i} are the deviation scores for the x variable, and the d_{y_i} are defined analogously for y. Note that with a little semantic gamesmanship, we can say that the variance is the covariance of a variable with itself. The product $d_{x_i} d_{y_i}$ is usually called a *cross product*.

2.2 *Matrix Expressions for Descriptive Statistics*

In this section we will return to our data matrix, **X**, with n observations and m variables,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$= \{x_{ij}\}.$$

We now define the *mean vector* $\bar{\mathbf{x}}$, such that

$$\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_m \end{bmatrix}$$

$$= \frac{1}{n}\,{}_1\mathbf{1}'_n\,\mathbf{X} \tag{2.7}$$

$$= \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}.$$

You might note that here we are beginning to see some of the advantages of matrix notation. For example, look at the second line of the above equation. The piece **1'X** expresses the operation of adding each of the columns of the **X** matrix and putting them in a row vector. How many more symbols would it take to express this using scalar notation using the summation operator Σ?

The mean vector can then be used to create the deviation score matrix, as below.

$$\mathbf{D} = \mathbf{X} - {}_{n}\mathbf{1}_{1}\overline{\mathbf{X}}'$$

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} = \mathbf{X} - \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \end{bmatrix} \tag{2.8}$$

$$= \mathbf{X} - \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \cdots & \cdots & \cdots & \cdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \end{bmatrix}$$

We would say of the **D** matrix that it is *column-centered*, as we have used the column means to center each column around zero.

Now lets reconsider the matrix **X'X**. This matrix is known as the *raw*, or *uncorrected*, *sum of squares and cross products matrix*. Often the latter part of this name is abbreviated *SSCP*. We will use the symbol **B** for the raw SSCP matrix:

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{i1}x_{i1} & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{im} \\ \sum x_{i2}x_{i1} & \sum x_{i2}x_{i2} & \cdots & \sum x_{i2}x_{im} \\ \cdots & \cdots & \cdots & \cdots \\ \sum x_{im}x_{i1} & \sum x_{im}x_{i2} & \cdots & \sum x_{im}x_{im} \end{bmatrix}. \tag{2.9}$$

In addition, we have seen this matrix expressed row by row and column by column in Equations (1.26) and (1.27). The uncorrected SSCP matrix can be corrected for the mean of each variable in **X**. Of course, it is then called the *corrected SSCP* matrix at that point:

$$\mathbf{A} = \mathbf{D}'\mathbf{D} \tag{2.10}$$

$$\mathbf{A} = \mathbf{B} - \frac{1}{n}\begin{bmatrix} \left(\sum x_{i1}\right)^2 & \left(\sum x_{i1}\right)\left(\sum x_{i2}\right) & \cdots & \left(\sum x_{i1}\right)\left(\sum x_{im}\right) \\ \left(\sum x_{i2}\right)\left(\sum x_{i1}\right) & \left(\sum x_{i2}\right)^2 & \cdots & \left(\sum x_{i2}\right)\left(\sum x_{im}\right) \\ \cdots & \cdots & \cdots & \cdots \\ \left(\sum x_{im}\right)\left(\sum x_{i1}\right) & \left(\sum x_{im}\right)\left(\sum x_{i2}\right) & \cdots & \left(\sum x_{im}\right)^2 \end{bmatrix} \tag{2.11}$$

Note that Equation (2.10) is analogous to the classic statement of the sum of squares in Equation (2.3) while the second version in Equation (2.11) resembles the hand calculator formula found in Equation (2.4). The correction for the mean in the formula for the corrected SSCP matrix **A** can be expressed in a variety of other ways:

$$\mathbf{A} = \mathbf{B} - \frac{1}{n}\left({}_{m}\mathbf{X}'_{n\,n}\mathbf{1}_{1}\right)\left({}_{m}\mathbf{X}'_{n\,n}\mathbf{1}_{1}\right)'$$

$$= \mathbf{B} - \frac{1}{n}\left(\mathbf{X}'\mathbf{1}\right)\left(\mathbf{1}'\mathbf{X}\right)$$

$$= \mathbf{B} - \frac{1}{n}\mathbf{X}'(\mathbf{1}\mathbf{1}')\mathbf{X}$$

$$= \mathbf{B} - \frac{1}{n}\mathbf{X}'\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\mathbf{X}$$

$$= \mathbf{B} - \bar{\mathbf{x}}(\mathbf{1}'\mathbf{X}).$$

Now, we come to one of the most important matrices in all of statistics, namely the *variance-covariance matrix*, often just called the *variance matrix*. It is created by multiplying the scalar 1/(n-1) times **A**, i. e.

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A} \tag{2.12}$$

This is the unbiased formula for **S**. From time to time we might have occasion to see the maximum likelihood formula which uses n instead of n - 1. The covariance matrix is a symmetric matrix, square, with as many rows (and columns) as there are variables. We can think of it as summarizing the relationships between the variables. As such, we must remember that the covariance between variable 1 and variable 2 is the same as the covariance between variable 2 and variable 1. The matrix **S** has $m(m+1)/2$ unique elements and $m(m-1)/2$ unique off-diagonal elements (of course there are m diagonal elements). We should also point out that $m(m-1)/2$ is the number of m things taken two at a time.

Previously we had mean-centered **X** using its column means to create the matrix **D** of deviation scores. Now we will further standardize our variables by creating Z scores. Define $\boldsymbol{\Delta}$ as the matrix consisting of diagonal elements of **S**. We define the function *Diag*(·) for this purpose:

$$\begin{aligned} \boldsymbol{\Delta} &= \text{Diag}(\mathbf{S}) \\ &= \begin{bmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & s_m^2 \end{bmatrix} \end{aligned} \tag{2.13}$$

Next, we need to invert the $\boldsymbol{\Delta}$ matrix, and take the square root of the diagonal elements. We can use the following notation in this case:

$$\boldsymbol{\Delta}^{-1/2} = \begin{bmatrix} 1/\sqrt{s_1^2} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{s_2^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\sqrt{s_m^2} \end{bmatrix} \tag{2.14}$$

The notion of taking the square root does not exactly generalize to matrices [see Equation (3.38)]. However, with a diagonal matrix, one can create a unique square root by taking the square roots of all the diagonal elements. With non-diagonal matrices there is no unique way to decompose a matrix into two identical components. In any case, the matrix $\boldsymbol{\Delta}^{-1/2}$ will now prove useful to us in creating Z scores. When you postmultiply a matrix by a
diagonal matrix, you operate on the columns of the premultiplying matrix. That is what we will do to **D**:

$$\mathbf{Z} = \mathbf{D}\boldsymbol{\Delta}^{-1/2}$$

$$= \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \begin{bmatrix} 1/\sqrt{s_1^2} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{s_2^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\sqrt{s_m^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d_{11}}{\sqrt{s_1^2}} & \frac{d_{12}}{\sqrt{s_2^2}} & \cdots & \frac{d_{1m}}{\sqrt{s_m^2}} \\ \frac{d_{21}}{\sqrt{s_1^2}} & \frac{d_{22}}{\sqrt{s_2^2}} & \cdots & \frac{d_{2m}}{\sqrt{s_m^2}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{d_{n1}}{\sqrt{s_1^2}} & \frac{d_{n2}}{\sqrt{s_2^2}} & \cdots & \frac{d_{nm}}{\sqrt{s_m^2}} \end{bmatrix} \tag{2.15}$$

which creates a matrix full of z scores. Note that just as postmultiplication by a diagonal matrix operates on the columns of the premultiplying matrix, premultiplying by a diagonal matrix operates on the rows of the postmultiplying matrix.

Now we are ready to create the matrix of correlations, **R**. The correlation matrix is the covariance matrix of the z scores,

$$\mathbf{R} = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$$

$$= \boldsymbol{\Delta}^{-1/2}\mathbf{S}\boldsymbol{\Delta}^{-1/2} \tag{2.16}$$

$$= \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}$$

Since the correlation of x and y is the same as the correlation between y and x, **R**, like **S**, is a symmetric matrix. As such we will have occasion to write it like

$$\mathbf{R} = \begin{bmatrix} 1 & & & \\ r_{21} & 1 & - & \\ \cdots & \cdots & & \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}$$

leaving off the upper triangular part. We can also do this for **S**.