Variance-Covariance Matrix

This lesson explains how to use matrix methods to generate a variance-covariance matrix from a matrix of raw data.

Variance

Variance is a measure of the variability or spread in a set of data. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance.

$$\mathrm{Var}(X) = \Sigma\,(X_i - \overline{X})^2 / N = \Sigma\, x_i^2 / N$$

where

- N is the number of scores in a set of scores
- $\overline{X}$ is the mean of the N scores.
- X_i is the ith raw score in the set of scores
- x_i is the ith deviation score in the set of scores
- $\mathrm{Var}(X)$ is the variance of all the scores in the set

Covariance

Covariance is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. We use the following formula to compute covariance.

$$\mathrm{Cov}(X, Y) = \Sigma\,(X_i - \overline{X})\,(Y_i - \overline{Y}) / N = \Sigma\, x_i y_i / N$$

where

- N is the number of scores in each set of data
- $\overline{X}$ is the mean of the N scores in the first data set
- X_i is the $i$the raw score in the first set of scores
- x_i is the ith deviation score in the first set of scores
- $\overline{Y}$ is the mean of the N scores in the second data set
- Y_i is the $i$the raw score in the second set of scores
- y_i is the ith deviation score in the second set of scores
- $\mathrm{Cov}(X, Y)$ is the covariance of corresponding scores in the two sets of data

Variance-Covariance Matrix

Variance and covariance are often displayed together in a variance-covariance matrix, (aka, a covariance matrix). The variances appear along the diagonal and covariances appear in the off-diagonal elements, as shown below.

$$\mathbf{V} = \begin{bmatrix} \Sigma\, x_1^2 / N & \Sigma\, x_1 x_2 / N & \dots & \Sigma\, x_1 x_c / N \\ \Sigma\, x_2 x_1 / N & \Sigma\, x_2^2 / N & \dots & \Sigma\, x_2 x_c / N \\ \dots & \dots & \dots & \dots \\ \Sigma\, x_c x_1 / N & \Sigma\, x_c x_2 / N & \dots & \Sigma\, x_c^2 / N \end{bmatrix}$$

where

V is a *c* x *c* variance-covariance matrix
N is the number of scores in each of the *c* data sets
x_i is a deviation score from the *ith* data set
$\Sigma x_i^2 / N$ is the variance of elements from the *i*th data set
$\Sigma x_i x_j / N$ is the covariance for elements from the *i*th and *j*th data sets

How to Create a Variance-Covariance Matrix

Suppose **X** is an *n* x *k* matrix holding ordered sets of raw data. For example, matrix **X** might display the scores on *k* tests for *n* students, as shown in Problem 1.

Starting with the raw data of matrix **X**, you can create a variance-covariance matrix to show the variance within each column and the covariance between columns. Here's how.

- Transform the raw scores from matrix **X** into deviation scores for matrix **x**.

$$\mathbf{x} = \mathbf{X} - \mathbf{11'X} (1 / n)$$

where

1 is an *n* x *1* column vector of ones
x is an *n* x *k* matrix of *deviation* scores: $x_{11}, x_{12}, \ldots, x_{nk}$
X is an *n* x *k* matrix of *raw* scores: $X_{11}, X_{12}, \ldots, X_{nk}$

- Compute **x'x**, the *k* x *k* deviation sums of squares and cross products matrix for **x**.
- Then, divide each term in the deviation sums of squares and cross product matrix by *n* to create the variance-covariance matrix. That is,

$$\mathbf{V} = \mathbf{x'x} (1 / n)$$

where

V is a *k* x *k* variance-covariance matrix
x'x is the deviation sums of squares and cross product matrix
n is the number of scores in each column of the original matrix **X**

In the next section, read Problem 1 for an example showing how to turn raw data into a variance-covariance matrix.

Test Your Understanding

Problem 1

The table below displays scores on math, English, and art tests for 5 students. Note that data from the table is represented in matrix **A**, where each column in the matrix shows scores on a test and each row shows scores for a student.

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \underset{\mathbf{A}}{\begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}}$$

Given the data represented in matrix **A**, compute the variance of each test and the covariance between the tests.

Solution

The solution involves a three-step process.

- First, we transform the *raw* scores in matrix **A** to *deviation* scores in matrix **a**, using the transformation formula described at how to transform raw scores to deviation scores.

$$\mathbf{a} = \mathbf{A} - \mathbf{11'A} (1 / n)$$

where

1 is an *5* x *1* column vector of ones
a is an *5* x *3* matrix of *deviation* scores: $a_{11}, a_{12}, \ldots, a_{53}$
A is an *5* x *3* matrix of *raw* scores: $A_{11}, A_{12}, \ldots, A_{53}$
n is the number of rows in matrix **A**

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} (1/5)$$

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \end{bmatrix} = \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}$$

- Then, to find the deviation score sums of squares matrix, we compute **a'a**, as shown below.

$$\mathbf{a'\,a} = \begin{bmatrix} 24 & 24 & -6 & -6 & -36 \\ 0 & 30 & 0 & 0 & -30 \\ 30 & -30 & 0 & 30 & -30 \end{bmatrix} \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix} = \begin{bmatrix} 2520 & 1800 & 900 \\ 1800 & 1800 & 0 \\ 900 & 0 & 3600 \end{bmatrix}$$

- And finally, to create the variance-covariance matrix, we divide each element in the deviation sum of squares matrix by *n*, as shown below.

$$\mathbf{V} = \mathbf{a'\,a} / n = \begin{bmatrix} 2520/5 & 1800/5 & 900/5 \\ 1800/5 & 1800/5 & 0/5 \\ 900/5 & 0/5 & 3600/5 \end{bmatrix} = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

We can interpret the variance and covariance statistics in matrix **V** to understand how the various test scores vary and covary.

- Shown in red along the diagonal, we see the variance of scores for each test. The art test has the biggest variance (720); and the English test, the smallest (360). So we can say that art test scores are more variable than English test scores.
- The covariance is displayed in black in the off-diagonal elements of matrix **V**.
  - The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.
  - The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

If the covariance between any tests had been negative, it would have meant that the test scores on those tests tend to move in opposite directions. That is, students with relatively high scores on the first test would tend to have relatively low scores on the second test.

< Previous lesson | Next lesson >