Random Vectors, Mean Vectors & Covariance Matrix

Module I (Lecture 2)

David Raj Micheal

August 2018

Contents

1 What is random vector?

Let us look at the data given below.

Note that each column of Table **??** represents a random variable. From this setup, one can identify the first observation as

$$(151, 53, 126, 110),$$

where the first, second, third and fourth coordinates represent the Height, weight, SBP and DBP of the oberservation respectively. So, any oservation can be viewed as a vector in the above way. The vector

$$\big(\text{Height}, \text{Weight}, \text{SBP}, \text{DBP}\big)$$

become a place holder and this vector is called as *Random Vector*.

For the convenience we write this vector as

$$x = \begin{pmatrix} \text{Height} \\ \text{Weight} \\ \text{SBP} \\ \text{DBP} \end{pmatrix}.$$

So, formally the definition of random vector goes as follows:

Table 1: Sample Multivariate Data having 4 variables

Height	Weight	SBP	DBP
152	52	105	113
167	55	114	135
156	62	115	124
174	53	107	127
175	51	124	130

Let $x_1, x_2, \dots, x_p$ be random variables. Then the vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

is called as *Random Vecotor*.

So, if a multivariate population is charcerized by

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix},$$

then we need to collect the data on these variables, which will look like

$$\begin{array}{c|cccc} & x_1 & x_2 & \dots & x_p \\ \hline 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 2 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & x_{n1} & x_{n2} & \dots & x_{np} \end{array}$$

Then the matrix from the above data

$$X_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

is called as *Data Matrix*. Note that, the ith row in the matrix represent the data on the ith observation in all the p variables. Similarly jth column in the data matrix represent the values of the random variable x_j for all the observations collected towards the random variable x_j. So, the (i, j)th entry in the data matrix,

$$x_{ij} = \text{the value of the random variable } x_j \text{ for the observation } i.$$

2 Population mean vector and Sample Mean Vector

In the case of univariate, the population mean of a random variable x is defined (informally) as the mean of all possible values of x and is denoted by μ. The mean is also referred to as the expected value of x, or $E(x)$. If the density $f(x)$ is unknown, the population mean(μ) remains unknown.

If a large random sample from the population represented by $f(x)$ is available, it is highly probable that the mean of the sample is close to μ.

The sample mean of a random sample of n observations is given by the ordinary arithmetic average of the n oberservations. That is, if x_j is the random variable then the sample mean of x_j, denoted by $\overline{(x_j)}$, is given by

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

Similar to the univariate case, we define the population mean and sample mean of a random vector as follows.

Population mean vector & Sample mean vector: If $x = (x_1, x_2, \ldots, x_p)^T$ is a random vector of p variables, then the population mean vector, μ, is given by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix},$$

and the sample mean vector, $\overline{x}$, is given by

$$\overline{x} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_p} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix}.$$

Lets caluculate the same vector using matrix algebra effectively which we will be using everywhere hereafter.

Note that, for any matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix},$$

$$X^T.e = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{pn} \end{pmatrix}_{p\times n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} = \begin{pmatrix} x_{11} + x_{21} + \cdots + x_{p1} \\ x_{21} + x_{22} + \cdots + x_{p2} \\ \vdots \\ x_{1n} + x_{2n} + \cdots + x_{pn} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i2} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{pmatrix}.$$

Hence the sample mean vector $\overline{x}$ is,

$$\overline{x} = \frac{1}{n} X^T e$$

where e is the vector of all ones.

Example: Consider the data matrix

$$X = \begin{array}{cc} \text{Age} & \text{Height} \\ \hline 10 & 100 \\ 12 & 110 \\ 11 & 105 \end{array}.$$

Find the mean vector.

3 Population Covariance and sample Covariance

If the random variable x_1 and x_2 simultaneously vary together then x_1 and x_2 are said to covary. Covariance between two random variables x_1 and x_2 is defined by

$$\text{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

Population (of size N) covariance between two randome variables x_k and x_j is, denoted by σ_{kj}, given by

$$\sigma_{kj} = \frac{1}{N}\sum_{i=1}^{N}(x_{ik} - \overline{x_k})(x_{ij} - \overline{x_j}).$$

Similarly the sample covariance between the random variables x_k and x_j is defined as,

$$s_{kj} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \overline{x_k})(x_{ij} - \overline{x_j}).$$

If there are p random variables, then the population covariance matrix is defined as follows, and denoted as Σ:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{np} \end{pmatrix},$$

Similarly the sample covariance matrix $S = (s_{jk})$ is the matrix of sample variances and covariances of the p variables:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{np} \end{pmatrix}.$$

Similar to the case of sample mean vector, we can also find the sample covariance matrix using matrix algebra as follows:

Note that, this matrix can be obtained by

$$S = \frac{1}{n-1} Y^T Y$$

where

$$Y = \begin{pmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & \dots & x_{1p} - \overline{x_p} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & \dots & x_{2p} - \overline{x_p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x_1} & x_{n2} - \overline{x_2} & \dots & x_{np} - \overline{x_p} \end{pmatrix} \tag{1}$$

Further you can simplify Y using the following observation.

$$\begin{aligned} Y &= \begin{pmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & \dots & x_{1p} - \overline{x_p} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & \dots & x_{2p} - \overline{x_p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x_1} & x_{n2} - \overline{x_2} & \dots & x_{np} - \overline{x_p} \end{pmatrix} & (2) \\ &= \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} - \begin{pmatrix} \overline{x_1} & \overline{x_2} & \dots & \overline{x_p} \\ \overline{x_1} & \overline{x_2} & \dots & \overline{x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_1} & \overline{x_2} & \dots & \overline{x_p} \end{pmatrix} & (3) \\ &= \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1} \begin{pmatrix} \overline{x_1} & \overline{x_2} & \dots & \overline{x_p} \end{pmatrix}_{1 \times p} & (4) \end{aligned}$$

Hence, the sample covariance matrix of the random variables $x_1, x_2, \ldots, x_p$ is

$$S = \frac{1}{n-1}(X - e\overline{x}^T)^T(X - e\overline{x}).$$