# MULTIVARIATE DATA

**Reference:** Johnson & Wichern (2002) *Applied Multivariate Statistical Analysis* Chapter 1 and 3.

## 1 Multivariate Random Sample

Let $X_{i1}, X_{i2}, \ldots, X_{in}$ be $n$ observations of the $i^{\text{th}}$ random variable $\boldsymbol{X}_i$ ($i = 1, 2, ..., p$), then $\boldsymbol{X}_j^T = (X_{1j} \quad X_{2j} \quad \ldots \quad X_{pj})$, $j = 1, 2, ..., n$, is the $j^{\text{th}}$ multivariate observation for random vector $\boldsymbol{X}$. Let $\boldsymbol{\mu}$ and $\Sigma$ be the mean and covariance of $\boldsymbol{X}$. That is, if

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

such that $\sigma_{ik} = \sigma_{ki}$ for all $i$ and $k$, then $\mu_i$ and $\sigma_{ii}$ are respectively the mean and variance of the random observation $X_i$ ($i = 1, 2, \ldots, p$) and

$$\mathbf{Cov}(X_i, X_k) = \sigma_{ik} = \sigma_{ki} \quad \text{for } i, k = 1, 2, \cdots, p.$$

Note : $n =$ sample size; $p =$ number of variables (p-dimension) in the random vector.

The entire data set can be placed in an $n \times p$ matrix:

variable = columns

obs = rows

sample data matrix

$$\mathcal{X} = \begin{pmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \\ \vdots \\ \boldsymbol{X}_n^T \end{pmatrix} = \begin{array}{c} \begin{array}{cccc} X_1 & \cdots & & X_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array}$$

## 2 Descriptive Statistics

(a) Sample Mean Vector $\overline{\boldsymbol{X}}_n$:

The sample mean of the random variable $X_i$ using the above random sample is given by

$$\boxed{\overline{X}_{in} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}.}$$

Note that $\boxed{\mathbf{E}}(\overline{X}_{in}) = \mu_i$, that is $\overline{X}_{in}$ is an unbiased estimator $\mu_i$.

expect

The sample mean vector of multivariate sample is given by

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j = \begin{pmatrix} \overline{\boldsymbol{X}}_{1n} \\ \overline{\boldsymbol{X}}_{2n} \\ \vdots \\ \overline{\boldsymbol{X}}_{pn} \end{pmatrix}.$$

p×1

Hence, $\mathbf{E}(\overline{\boldsymbol{X}}_n) = \boldsymbol{\mu}$, that is, $\overline{\boldsymbol{X}}_n$ is an unbiased estimator of $\boldsymbol{\mu}$.

(b) Sample Variances

- $S_i^{*2} = S_{ii}^*$ - bias estimator of $\sigma_{ii}$
- $S_i^2 = S_{ii}$ - unbiased estimator of $\sigma_{ii}$

Define

biased estimator of $\sigma_{ii}$

$$S_{ii}^* = S_i^{*2} = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \overline{X}_{in})^2$$

and

$$S_{ii} = S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \overline{X}_{in})^2$$

unbiased

Then

- $\mathbf{E}(S_i^{*2}) = \mathbf{E}(S_{ii}^*) = \frac{n-1}{n}\sigma_{ii}$ - implies $S_{ii}^*$ is a bias estimator of $\sigma_{ii}$.
- $\mathbf{E}(S_i^2) = \mathbf{E}(S_{ii}) = \sigma_{ii}$ - implies $S_{ii}$ is an unbiased estimator of $\sigma_{ii}$.

**Note:** $S_{ii} = \frac{n}{n-1}S_{ii}^*$. switch between two variance matrices

The square root of the sample variance is known as the sample standard deviation. $V^{1/2}$

(c) Sample Covariances $S^*_{ik}$ and $S_{ik}$: biased unbiased

The sample covariance gives a measure of association between two variables. A bias sample covariance between $X_i$ and $X_k$ is

pairs of variables

$$S^*_{ik} = \frac{1}{n}\sum_{j=1}^{n} \underbrace{\left(X_{ij} - \overline{X}_i\right)}_{\text{dev. var } i} \underbrace{\left(X_{kj} - \overline{X}_k\right)}_{\text{dev. var } k}.$$

and an unbiased sample covariance between $X_i$ and $X_k$ is

$$S_{ik} = \frac{1}{n-1}\sum_{j=1}^{n} \left(X_{ij} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_k\right).$$

**Note:** (1) $S_{ij} = \frac{n}{n-1}S^*_{ij}$. (2) $\mathbf{E}(S^*_{ij}) = \frac{n-1}{n}\sigma_{ij}$.
(3) $\mathbf{E}(S_{ij}) = \sigma_{ij}$ (4) When $i = k$, $S^*_{ik} = S^{*2}_i$ and $S_{ik} = S^2_i$.
(5) $S^*_{ik} = S^*_{ki}$ and (6) $S_{ik} = S_{ki}$ for all $i$ and $k$.

(d) Sample Covariance Matrices $\mathcal{S}_n$ and $\mathcal{S}$:

The bias sample covariance matrix $\mathcal{S}_n$ and unbiased sample covariance matrix $\mathcal{S}_n$ are given by

covariances symmetric / square

$$S^* = \mathcal{S}_n = \begin{pmatrix} S^*_{11} & S^*_{12} & \dots & S^*_{1p} \\ S^*_{21} & S^*_{22} & \dots & S^*_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^*_{p1} & S^*_{p2} & \dots & S^*_{pp} \end{pmatrix} = \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n\right)\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n\right)^T \quad \text{and}$$

variances

$$\mathcal{S} = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{pmatrix} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n\right)\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_n\right)^T.$$

Note that $\mathcal{S} = \frac{n}{n-1}\mathcal{S}_n$.

(e) Generalized variance Single value!

Generalized sample variance is determinant of the sample covariance matrix $|\mathcal{S}|$.

(f) Sample Correlation $R_{ij}$

Sample correlation coefficient is a measure of the *linear association* between two random variables. This does not depend on the unit of measurement.

Sample correlation coefficient between random variables $X_i$ and $X_j$ is defined as

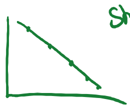
covariance

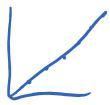
pairs

$$R_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{j=1}^{n}(X_{ij}-\overline{X}_i)(X_{kj}-\overline{X}_k)}{\sqrt{\sum_{j=1}^{n}(X_{ij}-\overline{X}_i)^2}\sqrt{\sum_{j=1}^{n}(X_{kj}-\overline{X}_k)^2}}$$

variance

for $i \neq k = 1, 2, \ldots, p$.

-1  Strong neg linear association

+1  Strong positive ass

**Properties of $R_{ik}$:**

(1) Sample correlation $R_{ik}$ must lie between $-1$ and $1$, that is, $-1 \leq R_{ik} \leq 1$ for all $i, k$.

(2) If $R_{ik} = 0$, there is no linear association between variables $X_i$ and $X_k$. Otherwise, the sign of $R_{ik}$ gives the direction of association.

(3) $R_{ik}$ remains unchanged if the random variables $X_i$ and $X_k$ are transformed to random variables $Y_i$ and $Y_k$ such that $Y_i = aX_i + b$, $Y_k = cX_k + d$ where $a, b, c$ and $d$ are constants and $a$ and $c$ have the same sign (that is, $ac > 0$).

(g) Sample Correlation Matrix $\mathcal{R}_n$

$$\mathcal{R}_n = \begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1p} \\ R_{21} & R_{22} & \ldots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p1} & R_{p2} & \ldots & R_{pp} \end{pmatrix}$$

=1

Example 1.1

Four receipts from a bookstore

| | $R_1$ | $R_2$ | $R_3$ | $R_4$ |
|---|---|---|---|---|
| Variable: sales ($) | $42 | $52 | $48 | $58 |
| Variables: number of books | 4 | 5 | 4 | 3 |

$$X_{n \times p} \atop (4 \times 2) = \begin{array}{cc} X_1 & X_2 \\ \end{array}$$

$$\underset{(4\times 2)}{X_{n\times p}} = \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix} \end{array}$$

first obs (from $R_1$)

Find the sample mean vector $\bar{X}$

first find the mean of each variable

$$\bar{x}_k = \frac{1}{n}\sum_1^n x_{jk}$$

$$\bar{x}_1 = \frac{1}{4}(42+52+48+58) = 50$$

$$\bar{x}_2 = \frac{1}{4}(4+5+4+3) = 4$$

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

Find the sample biased covariance matrix $S^*$

first find the variances

$$s^*_{kk} = \frac{1}{n}\sum_1^n (x_{jk} - \bar{x}_k)^2$$

$$s^*_{11} = \frac{1}{4}\left((42-50)^2 + (52-50)^2 + (48-50)^2 + (58-50)^2\right) = 34$$

$$s^*_{22} = \frac{1}{4}\left((4-4)^2 + (5-4)^2 + (4-4)^2 + (3-4)^2\right) = 0.5$$

then find the biased covariance

$$s^*_{ik} = \frac{1}{n}\sum_1^n (x_{ij} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

$$= \frac{1}{4}\Big((42-50)(4-4) + (52-50)(5-4) + (48-50)(4-4) +$$

$$(58-50)(3-4)\Big) = -1.5 = S^*_{12} = S^*_{21}$$

construct the matrix

$$S^* = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

To find the Generalised variance

$$|S| = \begin{vmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{vmatrix}$$

$$= ad - bc$$

$$= S_{11} S_{22} - S_{12} S_{21}$$

$$= 34 \times 0.5 - -1.5 \times -1.5$$

$$= 14.75$$

Find the sample correlation matrix

$R$

$$R_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{-1.5}{\sqrt{34}\sqrt{0.5}} = -0.36$$

$$R = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$