<u>**MULTIVARIATE DATA**</u>

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^{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^{th} multivariate observation for random vector $\boldsymbol{X}$. Let $\boldsymbol{\mu}$ and Σ 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 μ_i and σ_{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:

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

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

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

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

The sample mean vector of multivariate sample is given by

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

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 σ_{ii}
- $S_i^2 = S_{ii}$ - unbiased estimator of σ_{ii}

Define

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

and

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

Then

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

Note: $S_{ii} = \frac{n}{n-1}S_{ii}^*$.

The square root of the sample variance is known as the sample standard deviation.

(c) Sample Covariances S_{ik}^* and S_{ik}:

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

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

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 $i = k$, $S_{ik}^* = S_i^{*2}$ and $S_{ik} = S_i^2$.
(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

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

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

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

(e) Generalized variance

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

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

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

<u>Properties of R_{ik}:</u>

(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 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 $\mathcal{R}_n$

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