Multivariate Data, Random Vector and Random Sample

Multivariate Data

Display n measurements on k variables:

	variable 1	variable 2	$\cdots$	variable j	$\cdots$	variable k
item 1:	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1k}
item 2:	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
item i:	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ik}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
item n:	x_{n1}	x_{n2}	$\cdots$	x_{nj}	$\cdots$	x_{nk}

Matrix representation:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nk} \end{pmatrix}$$

Sample mean:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Sample variance and covariance:

$$s_j^2 = s_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

$$s_{jl} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)$$

where $j = 1, 2, \ldots, k$ and $l = 1, 2, \ldots, k$.

Sample correlation coefficients:

$$r_{jl} = \frac{s_{jl}}{\sqrt{s_{jj}}\sqrt{s_{ll}}}$$

Basic descriptive statistics:

sample means

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_k \end{pmatrix}$$

Sample variance-covariance matrix:

$$S_n = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{pmatrix}$$

Sample correlation coefficient matrix:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}$$

Random vector

$$\vec{X}' = (X_1, X_2, \cdots, X_p) \text{ follows the distribution } f(x_1, x_2, \ldots, x_p)$$

Mean $\mu_i = E(X_i)$, variance $\sigma_i^2 = \sigma_{ii} = E(X_i - \mu_i)^2$

Covariance: $\sigma_{jl} = cov(X_j, X_l) = E(X_j - \mu_j)(X_l - \mu_l)$

correlation coefficient: $\rho_{jl} = \frac{\sigma_{jl}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}}}$

Population mean:

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

Population variance-covariance matrix:

$$\Sigma = E(\vec{X} - \mu)(\vec{X} - \mu)' = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

Population correlation coefficient matrix:

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix}.$$

Linear combinations of random vectors:

$$\vec{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = C\vec{X}$$

$$\mu_{\vec{Z}} = E(\vec{Z}) = E(C\vec{X}) = C\mu_{\vec{X}}$$

$$\Sigma_{\vec{Z}} = Cov(\vec{Z}) = E(\vec{Z} - \mu_{\vec{Z}})(\vec{Z} - \mu_{\vec{Z}})'$$

$$= E(C(\vec{X} - \mu_{\vec{X}})(\vec{X} - \mu_{\vec{X}})'C') = C\Sigma_{\vec{X}}C'$$

Random Sample

$$X_{n\times p} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} \vec{X}_1' \\ \vec{X}_2' \\ \vdots \\ \vec{X}_n' \end{pmatrix}$$

If the row vectors, $\vec{X}_1'$, $\vec{X}_2'$, …, $\vec{X}_n'$ represent independent observations from a common joint distribution with density function $f(x) = f(x_1, x_2, \ldots, x_p)$, then $\vec{X}_1'$, $\vec{X}_2'$, …, $\vec{X}_n'$ is said to form an independent and identically distributed (iid) random sample for $f(x)$. Mean, variance-covariance matrix and correlation-coefficient matrix can be defined for a random sample. The formulae are similar to those for multivariate data introduced earlier except that lowercase letters x_{ij} need be changed to uppercase letters.

Theorem: Let $X_1, X_2, \ldots, X_n$ be a random sample from a joint distribution that has mean vector $\vec{\mu}$ and variance-covariance matrix Σ. Then

$$E(\bar{X}) = \vec{\mu}$$
$$Cov(\bar{X}) = \frac{1}{n}\Sigma$$
$$E(S_n) = \frac{n-1}{n}\Sigma$$

Remark. unbiased sample variance-covariance matrix:

$$S = \frac{n}{n-1}S_n = \frac{1}{n-1}\sum_{i=1}^{n}(\vec{X}_i - \bar{X})(\vec{X}_j - \bar{X})'$$

Generalized sample variance $= \mid S \mid$