Random Sample

- Modeling of Multivariate Data
  - The data set
$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$
is usually regarded as a realization of a matrix of random variables
$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$
  - $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are said to form a *random sample* from $f(\mathbf{x})$ if
    - $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ represent *independent* observations
    - $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ are from a *common* joint distribution with density function
$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p)$$
$\Rightarrow$ measurements of the p variables in a single trial will usually be correlated
    - joint density function of $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$
$$f(\mathbf{x}_1)f(\mathbf{x}_2)\cdots f(\mathbf{x}_n)$$
where $f(\mathbf{x}_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$

- some examples
  - Example 1:
    - to design of a permit system for utilizing a wildness canoe area without overcrowding, a manager took a survey of users
    - total wilderness area was divided into subregions, and respondents were asked to give information on the regions visited, length of stay, and other variables
    - sampling method 1: persons were randomly selected from all those who entered the wilderness area during a particular week

      $\Rightarrow$ all person were equally likely to be in the sample
    - sampling method 2: sampler waited at a campsite and interviewed only canoeists who reached that spot
  - Example 2: a study concerns the gross weight of municipal solid waste generated per year, x_1 = paper and paperboard waste and x_2 = plastic waste

Table 3.1 Solid Waste

Year	1960	1970	1980	1990	1995	2000	2003
x_1 (paper)	29.2	44.3	55.2	72.7	81.7	87.7	83.1
x_2 (plastics)	.4	2.9	6.8	17.1	18.9	24.7	26.7

- **Q**: Should these measurements on $\mathbf{X}' = [X_1, X_2]$ be treated as a random sample?

- Some theoretical results under the modeling
  - **Result 3.1.** Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a joint distribution that has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then $\overline{\mathbf{X}}$ is an *unbiased* estimator of $\boldsymbol{\mu}$, and its covariance matrix is

$$\frac{1}{n}\boldsymbol{\Sigma}$$

That is,

$$E(\overline{\mathbf{X}}) = \boldsymbol{\mu} \qquad \text{(population mean vector)}$$

$$\text{Cov}(\overline{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma} \qquad \begin{pmatrix}\text{population variance-covariance matrix} \\ \text{divided by sample size}\end{pmatrix}$$

For the covariance matrix $\mathbf{S}_n$,

$$E(\mathbf{S}_n) = \frac{n-1}{n}\boldsymbol{\Sigma} = \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma}$$

Thus,

$$E\left(\frac{n}{n-1}\mathbf{S}_n\right) = \boldsymbol{\Sigma}$$

so $[n/(n-1)]\mathbf{S}_n$ is an *unbiased* estimator of $\boldsymbol{\Sigma}$, while $\mathbf{S}_n$ is a *biased* estimator with (bias) $= E(\mathbf{S}_n) - \boldsymbol{\Sigma} = -(1/n)\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_p\end{bmatrix} = \begin{bmatrix}E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p)\end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'$$

- In future lecture,

$$\mathbf{S} = \left(\frac{n}{n-1}\right)\mathbf{S}_n = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$$

will replace $\mathbf{S}_n$ as the sample covariance matrix in most of the material.

- Note: even though the (i, k)th entry of $\mathbf{S}$, s_{ik}, is an unbiased estimator of σ_{ik}

$$E(\sqrt{s_{ii}}) \neq \sqrt{\sigma_{ii}} \text{ and } E(r_{ik}) \neq \rho_{ik}$$

➢ linear combination of variables

- The linear combination $\mathbf{c}'\mathbf{X} = c_1X_1 + \cdots + c_pX_p$ has

$$\text{mean} = E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\mu}$$
$$\text{variance} = \text{Var}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$$

where $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$.

- The linear combinations $\mathbf{Z} = \mathbf{CX}$ have

$$\boldsymbol{\mu}_{\mathbf{Z}} = E(\mathbf{Z}) = E(\mathbf{CX}) = \mathbf{C}\boldsymbol{\mu}_{\mathbf{X}}$$
$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \text{Cov}(\mathbf{Z}) = \text{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are the mean vector and variance-covariance matrix of $\mathbf{X}$.

- sample values
  - **Result 3.5.** The linear combinations

$$\mathbf{b}'\mathbf{X} = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$
$$\mathbf{c}'\mathbf{X} = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

have sample means, variances, and covariances that are related to $\bar{\mathbf{x}}$ and $\mathbf{S}$ by

$$\text{Sample mean of } \mathbf{b}'\mathbf{X} = \mathbf{b}'\bar{\mathbf{x}}$$
$$\text{Sample mean of } \mathbf{c}'\mathbf{X} = \mathbf{c}'\bar{\mathbf{x}}$$
$$\text{Sample variance of } \mathbf{b}'\mathbf{X} = \mathbf{b}'\mathbf{S}\,\mathbf{b}$$
$$\text{Sample variance of } \mathbf{c}'\mathbf{X} = \mathbf{c}'\mathbf{S}\,\mathbf{c}$$
$$\text{Sample covariance of } \mathbf{b}'\mathbf{X} \text{ and } \mathbf{c}'\mathbf{X} = \mathbf{b}'\mathbf{S}\,\mathbf{c}$$

- The sample mean and covariance relations in Result 3.5 pertain to any number of linear combinations. Consider the q linear combinations

$$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, \qquad i = 1, 2, \ldots, q \tag{3-37}$$

These can be expressed in matrix notation as

$$\begin{bmatrix} a_{11}X_1 & + & a_{12}X_2 & + \cdots + & a_{1p}X_p \\ a_{21}X_1 & + & a_{22}X_2 & + \cdots + & a_{2p}X_p \\ \vdots & & \vdots & \vdots & \vdots \\ a_{q1}X_1 & + & a_{q2}X_2 & + \cdots + & a_{qp}X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathbf{AX}$$

Result 3.6. The q linear combinations $\mathbf{AX}$ have sample mean vector $\mathbf{A}\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{ASA}'$.

❖ **Reading**: Textbook, 3.3, 2.6, 3.6