Random Sample

- Modeling of Multivariate Data

  - The data set

$$\begin{array}{r} \\ \text{item 1} \\ \\ \\ \text{item } n \end{array} \begin{array}{c} \begin{array}{cccc} \text{var 1} & \text{var 2} & & \text{var } p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array}$$

(after data collected)
(empirical distribution)
$\boxed{\text{a good sample}} \xrightarrow{\text{ss}}$ population distribution

is usually regarded as a realization of a matrix of random variables

$\boxed{\text{contain assumptions}}$

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$

(before data collected)
(distribution of population)
↳ contain parameters.
unknown.

  - $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are said to form a *random sample* from $f(\mathbf{x})$ if

$f(x_2|x_1) = f(x_2)$

    - $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ 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

→ multivariate analysis (cov, cor)
time series
→ Linear model (GLM)
$(Y = \underline{X\beta} + \varepsilon)$

    - joint density function of $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$

relationship between variables

$$f(\mathbf{x}_1)f(\mathbf{x}_2)\cdots f(\mathbf{x}_n)$$

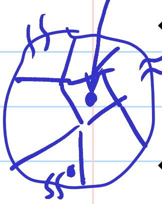
where $f(\mathbf{x}_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$

  - some examples

Data. user1 Region × LOS ×] sample 1.
user2 × ×

    - Example 1:
      - to design of a permit system for utilizing a wildness canoe area without overcrowding, a manager took a survey of users

sample 1.



      - 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] $\Rightarrow$ random 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

time index

Table 3.1 Solid Waste

	Year	1960 (item 1)	1970 (item 2)	1980	1990	1995	2000	2003 (item 7)
var 1	x_1 (paper)	29.2	44.3	55.2	72.7	81.7	87.7	83.1
var 2	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 ← i.i.d.

➢ **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

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

→ parameters

That is,

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

function of data

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

sample cov matrix

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, estimator / estimate

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

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

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

var 1 … var p.

proof: ① $E(\bar{X}) = E\left(\frac{1}{n}(X_1+X_2+\cdots+X_n)\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu$

② $\text{cov}(\bar{X}) = E[(\bar{X}-\mu)(\bar{X}-\mu)'] \rightarrow \left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right]\left[\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)\right]'$

$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu)(X_j-\mu)'$

$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E(X_i-\mu)(X_j-\mu)'$

∵ X_j, X_i indep. ∴ $E(X_i-\mu)(X_j-\mu)' = 0$, for $i \neq j$

$= \frac{1}{n^2}\sum_{i=1}^{n}E(X_i-\mu)(X_i-\mu)' = \frac{1}{n^2}(n\Sigma) = \frac{1}{n}\Sigma.$

③ $nS_n = X'(I - \frac{1}{n}11')X \quad \left(X = \begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}\right)$

$= X'X - \frac{1}{n}(X'1)(1'X) = \frac{1}{n}(n\bar{X})(n\bar{X})' = n\bar{X}\bar{X}'$

$X = [\ldots] \Rightarrow (X'X)_{jk} \Rightarrow X'X = \sum_{i=1}^{n}X_iX_i' \quad \left(\left(\sum_{i=1}^{n}X_iX_i'\right)_{jk} = ?\right)$

$E(nS_n) = E(X'X) - n\cdot E(\bar{X}\bar{X}')$

$= E\left(\sum_{i=1}^{n}X_iX_i'\right) - nE(\bar{X}\bar{X}') = \sum_{i=1}^{n}E(X_iX_i') - nE(\bar{X}\bar{X}')$

$\Sigma = E(X_i-\mu)(X_i-\mu)' = E(X_iX_i') - \mu\mu'$

$\Rightarrow E(X_iX_i') = \Sigma + \mu\mu'$

$= \sum_{i=1}^{n}(\Sigma+\mu\mu') - n\left(\frac{1}{n}\Sigma + \mu\mu'\right)$

$= n\Sigma + n\mu\mu' - \Sigma - n\mu\mu'$

$= (n-1)\Sigma$

- In future lecture,

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

d.f.

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}

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

$= s_{ik}/\sqrt{s_{ii}}\sqrt{s_{kk}}$

$Eg(X) \neq g(EX)$ in general. But $E(g(x)) \approx g(E(x))$ by δ-method.

➢ linear combination of variables

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

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

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

$E[(c'X-c'\mu)(c'X-c'\mu)']$
$=E[c'(X-\mu)(X-\mu)'c]$

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

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

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

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

$E[(CX-C\mu)(CX-C\mu)']$
$=E[C(X-\mu)(X-\mu)'C']$

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

unbiased est'or.

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

$\text{Cov}(b'X, c'X)$
$=E((b'X-b'\mu)(c'X-c'\mu)')$
$=E(b'(X-\mu)(X-\mu)'c)$
$=b'E(X-\mu)(X-\mu)'c$

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

vector

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

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

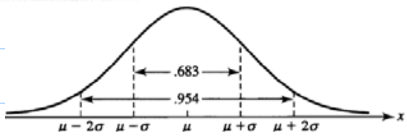
Multivariate Normal Distribution

for continuous variables.

- Density of Normal Distribution
  - univariate case: normal density with mean μ and variance σ^2

$$f(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{normalization}} e^{-\boxed{[(x-\mu)/\sigma]}^2/2} \qquad -\infty < x < \infty$$

distance. shape



  - multivariate case: a p-dimensional normal density for the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ has the form

quadratic form.

$$f(\mathbf{x}) = \underbrace{\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}}_{\text{normalization}} e^{-\underbrace{\boxed{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}/2}_{\text{shape}}}$$

where $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$. $\boldsymbol{\mu}$ is the mean of $\mathbf{X}$. $\mathbf{\Sigma}$ is the variance–covariance matrix of $\mathbf{X}$. (Note. $\mathbf{\Sigma}^{-1}$ must exist, i.e., $|\Sigma| \neq 0$)

symmetric positive definite

We shall denote this p-dimensional normal density by $\boxed{N_p(\boldsymbol{\mu}, \mathbf{\Sigma})}$

- the term $\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ | $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

  is a statistical distance measure

  - Contours of constant density for the p-dimensional normal distribution are ellipsoids defined by $\mathbf{x}$ such the that

$\Sigma \Rightarrow (\lambda_i, e_i)$

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

$\Sigma^{-1} \Rightarrow (\frac{1}{\lambda_i}, e_i)$

$1 = (\sqrt{c_1}\, y_1)^2 = (\sqrt{c_2}\, y_2)^2 = \cdots$

$y_1 : y_2 : \cdots : y_p = \frac{1}{\sqrt{c_1}} : \frac{1}{\sqrt{c_2}} : \cdots : \frac{1}{\sqrt{c_p}}$

These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$, where $\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$ for $i = 1, 2, \ldots, p$. $(\lambda, \mathbf{e})$ is an eigenvalue–eigenvector pair for $\mathbf{\Sigma}$

- Example: bivariate normal density → univariate normal

$$\mu_1 = E(X_1), \quad \mu_2 = E(X_2),$$

$$\sigma_{11} = \text{Var}(X_1), \sigma_{22} = \text{Var}(X_2), \text{ and } \rho_{12} = \sigma_{12}/(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}) = \text{Corr}(X_1, X_2).$$

- $$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \qquad \mathbf{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$
- by writing $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$
- $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

$$= [x_1-\mu_1, x_2-\mu_2]\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$$

$$= \frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$= \frac{1}{1-\rho_{12}^2}\left[\left(\underbrace{\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}}_{y_1}\right)^2 + \left(\underbrace{\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}}_{y_2}\right)^2 - 2\rho_{12}\left(\underbrace{\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}}_{y_1}\right)\left(\underbrace{\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}}_{y_2}\right)\right]$$

- written in terms of the standardized values $(x_1-\mu_1)/\sqrt{\sigma_{11}}$ and $(x_2-\mu_2)/\sqrt{\sigma_{22}}$.
- if X_1 and X_2 are uncorrelated, they are independent