What to do today (Jan 17)?

Part I. Introduction and Preparation

I.1. General Introduction

I.2. Review on Matrix Algebra (Chp 2.1-4, Supplement 2A)

I.3. Introduction to R (More at the 1st tutorial)

I.4. Multivariate Random Variables and Distributions (Chp 1, 2.5-6, 3)

Part II. Inference under Multivariate Normal Distribution (Chp 4-7)

II.1 Multivariate Normal Distribution (Chp 4)

II.1.1 Multivariate Normal Distribution $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Chp 4.1-2)

Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-11)

Part IV. Other Topics (Chp 12)

Part I.4.1 Random Vectors and Matrices (Chp 2.5): Review

- **random vector (multivariate random variable).**

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = (X_1, X_2, \ldots, X_p)'$$

is a p-dim random vector if $X_1, \ldots, X_p$ are r.v.s.

- **distribution.** The cdf of $\mathbf{X}$ is the joint cumulative distribution function (joint cdf) of $X_1, \ldots, X_p$: for $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$,

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, \ldots, X_p \leq x_p).$$

- Suppose the study has n subjects with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ their observations on the p-dim r.v. $\mathbf{X} = (X_1, \ldots, X_p)'$. A $n \times p$ **random matrix**:

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

Part I.4.2 Mean Vectors and Covariance Matrices: Review

- **expectation (population mean)** of p-dim random vector $\mathbf{X} = (X_1, \ldots, X_p)^{'}$:

$$E(\mathbf{X}) = (E(X_1), E(X_2), \ldots, E(X_p))^{'}.$$

denoted by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{'}$.

- **(population) variance matrix.** p-dim r.v. $\mathbf{X}$'s variance:

$$V(\mathbf{X}) = E\left[\mathbf{X} - E(\mathbf{X})\right]\left[\mathbf{X} - E(\mathbf{X})\right]^{'} = E(\mathbf{X}\mathbf{X}^{'}) - (E\mathbf{X})(E\mathbf{X})^{'},$$

denoted by $\boldsymbol{\Sigma} = \left(\sigma_{ij}\right)$ with $\sigma_{ij} = \sigma_{ji} = Cov(X_i, X_j)$.

- **population correlation matrix.** A standardized variance-covariance matrix: $\boldsymbol{\rho} = \left(\rho_{ij}\right)$ with $\rho_{ij} = cor(X_i, X_j) = \sigma_{ij}\big/\sqrt{\sigma_{ii}}\big/\sqrt{\sigma_{jj}}$ and thus $\rho_{ii} = 1$.

With $\mathbf{V} = diag\left(\sigma_{11}, \ldots, \sigma_{pp}\right)$, $\boldsymbol{\rho} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2}$, $\boldsymbol{\Sigma} = \mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2}$

- The **covariance** of two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is $Cov(\mathbf{X}, \mathbf{Y}) = E\left[\mathbf{X} - E(\mathbf{X})\right]\left[\mathbf{Y} - E(\mathbf{Y})\right]^{'}$.

- **linear combinations of r.v.s.**
  - Suppose $Y = c_1 X_1 + c_2 X_2 + \ldots + c_p X_p = \mathbf{c}^{'}\mathbf{X}$.
    - $E(Y) = c_1 E(X_1) + c_2 E(X_2) + \ldots + c_p E(X_p) = \mathbf{c}^{'}\mathbf{X}$.
    - $V(Y) = \mathbf{c}^{'} V(\mathbf{X})\mathbf{c} = \mathbf{c}^{'}\boldsymbol{\Sigma}\mathbf{c}$.

    If $X_1, \ldots, X_p$ are indpt, $V(Y) = \mathbf{c}^{'} diag\left(\sigma_1^2, \ldots, \sigma_p^2\right)\mathbf{c} = \sum_{j=1}^{p} c_j^2 \sigma_j^2$.
  - Suppose $Z_j = c_{j1} X_1 + c_{j2} X_2 + \ldots + c_{jp} X_p$ for $j = 1, \ldots, q$, and $\mathbf{Z} = (Z_1, \ldots, Z_q)^{'} = \mathbf{C}_{q\times p}\mathbf{X}_{p\times 1}$.
    - $E(\mathbf{Z}) = \mathbf{C}E(\mathbf{X}) = \mathbf{C}\boldsymbol{\mu}$.
    - $V(\mathbf{Z}) = \mathbf{C}V(\mathbf{X})\mathbf{C}^{'} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{'}$.
  - Suppose $\mathbf{U} = \mathbf{A}\mathbf{X}$ and $\mathbf{W} = \mathbf{B}\mathbf{Y}$.
    - $Cov(\mathbf{U}, \mathbf{W}) = \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})\mathbf{B}^{'}$

Part I.4.3 Descriptive Multivariate Analysis: Review

Summary Statistics: Suppose a study has n iid observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ on a p-dim r.v. $\mathbf{X} = (X_1, \ldots, X_p)'$ with $E(\mathbf{X}) = \boldsymbol{\mu}$ and $V(\mathbf{X}) = \boldsymbol{\Sigma}$.

- $E(\mathbf{X}_i) = \boldsymbol{\mu}$ and $V(\mathbf{X}_i) = \boldsymbol{\Sigma}$ for $i = 1, \ldots, n$.
- **sample mean vector** $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)'$ with $\bar{x}_j = \sum_{i=1}^{n} x_{ij} / n$.
- **sample variance matrix**

$$\mathbf{S}_n = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{pmatrix}$$

  with $s_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) / n$.
- **sample correlation matrix**

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \ldots & \vdots & \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}$$

  with $r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$.

Part I.4.4 More on Descriptive Multivariate Analysis (Chp 3.3)

Consider $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from the population with population mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

- That is, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid observations on $\mathbf{X}$ with $E(\mathbf{X}) = \boldsymbol{\mu}$ and $V(\mathbf{X}) = \boldsymbol{\Sigma}$.
- The sample mean $\bar{\mathbf{X}} = \left(\mathbf{X}_1 + \ldots + \mathbf{X}_n\right)/n$ is an *unbiased* estimator of $\boldsymbol{\mu}$: $E\left(\bar{\mathbf{X}}\right) = \boldsymbol{\mu}$.
- The sample variance matrix $\mathbf{S}_n = \left(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i' - n\bar{\mathbf{X}}\bar{\mathbf{X}}'\right)/n$ is an *biased* estimator of $\boldsymbol{\Sigma}$: $E\left(\mathbf{S}_n\right) = \frac{n-1}{n}\boldsymbol{\Sigma}$.
- **(unbiased) sample variance-covariance matrix:**

$$\mathbf{S} = \frac{n}{n-1}\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

Part II.1.1 Multivariate Normal Distribution $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Chp 4.1-2)

The most important distribution in all of Statistics is the normal (Gaussian) distribution.

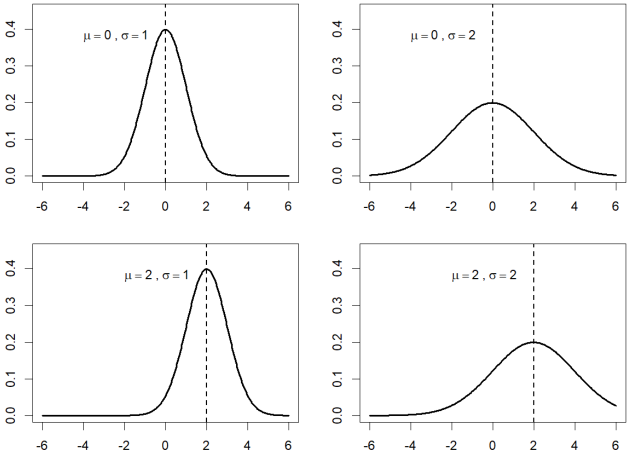
Review on univariate normal distribution $N(\mu, \sigma^2)$:

Definition. A r.v. X has a *normal* distribution if its pdf

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}, \quad -\infty < x < \infty,$$

where $\sigma > 0$. Denote it by $X \sim N(\mu, \sigma^2)$.

- If $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$ and $V(X) = \sigma^2$.
- $N(\mu, \sigma^2)$: a family of distributions.
  - e.g. $N(0, 1)$, *the standard normal distribution*. $F(x)$ of $N(0, 1)$ is often denoted by $\Phi(x)$ and the rv by Z.



More about the normal distributions ...

- The pdf is symmetric about μ.
- As μ changes, the mode of the pdf curve shifts accordingly.
- As σ increases, the spread of the pdf curve increases.

Part II.1.1 Multivariate Normal Distribution $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Chp 4.1-2)

What is a multivariate normal distribution?

Definition. A r.v. X has a *normal* distribution if its pdf

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-\mu)\sigma^{-2}(x-\mu)\right\}, \quad -\infty < x < \infty,$$

where $\sigma > 0$. Denote it by $X \sim N(\mu, \sigma^2)$.

Part II.1.1 Multivariate Normal Distribution $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Chp 4.1-2)

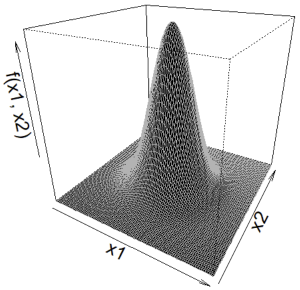
II.1.1A. Multivariate normal distribution

Definition. A p-dim r.v. $\mathbf{X}$ has a *normal* distribution if its pdf

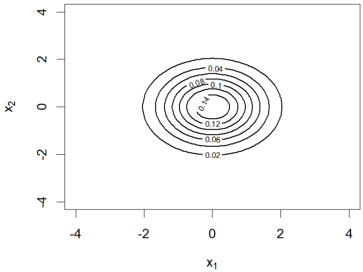
$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \ -\infty < \mathbf{x} < \infty,$$

where $\boldsymbol{\Sigma}$ is positive definite. Denote it by $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- If $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $E(\mathbf{X}) = \boldsymbol{\mu}$ and $V(\mathbf{X}) = \boldsymbol{\Sigma}$.
- $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: a family of (multivariate) distributions.
  - e.g. $MN_p(\mathbf{0}, \mathbf{I})$, *the standard (multivariate) normal distribution.*

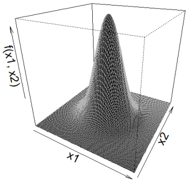


(a) BN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$)

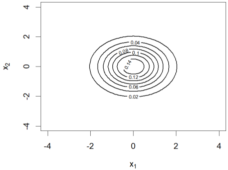


(b) Contour

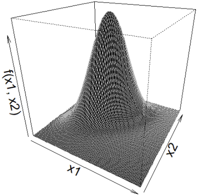
$\boldsymbol{\mu} = (0,0)^{'}$ and $\boldsymbol{\Sigma} = diag(1,1)$



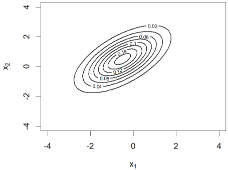
(a) $BN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$



(b) Contour



(c) $BN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

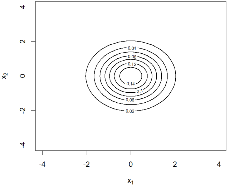


(d) Contour

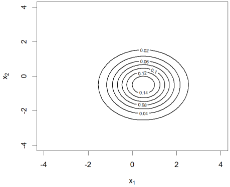
- (1) $\boldsymbol{\mu} = (0,0)'$ and $\boldsymbol{\Sigma} = diag(1,1)$; (2) $\boldsymbol{\mu} = (0.5, -0.5)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1.2 & 0.75 \\ 0.75 & 1.2 \end{pmatrix}$

More about the normal distributions ...

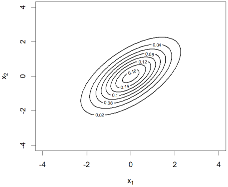
- The pdf is symmetric about $\boldsymbol{\mu}$.
- As $\boldsymbol{\mu}$ changes, the mode of the pdf curve shifts accordingly.
- As $\boldsymbol{\Sigma}$ changes, the spread and/or shape of the pdf surface change accordingly.



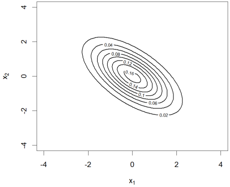
Contour of (1)



Contour of (2)



Contour of (3)



Contour of (4)

- (1) $\boldsymbol{\mu} = (0, 0)'$ and $\boldsymbol{\Sigma} = diag(1, 1)$; (2) $\boldsymbol{\mu} = (0.5, -0.5)'$ and $\boldsymbol{\Sigma} = diag(1, 1)$;
- (3) $\boldsymbol{\mu} = (0, 0)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1.2 & 0.75 \\ 0.75 & 1.2 \end{pmatrix}$; (4) $\boldsymbol{\mu} = (0, 0)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1.2 & -0.75 \\ -0.75 & 1.2 \end{pmatrix}$

II.1.1B. Shape of Multivariate Normal Density

Suppose $\mathbf{X} \sim BN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: its pdf

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \ -\infty < \mathbf{x} < \infty,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{\Sigma} = (\sigma_{ij}) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

- If $\rho = 0$, $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \phi(x_1; \mu_1, \sigma_1^2)\phi(x_2; \mu_2, \sigma_2^2)$: $\phi(x; \mu, \sigma^2)$ is the pdf of $X \sim N(\mu, \sigma^2)$.

- The density $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a constant for all $\mathbf{x} = (x_1, x_2)'$ satisfy $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$. (It defines an ellipse centered at $\boldsymbol{\mu} = (\mu_1, \mu_2)'$.)

Normal density contour:

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

- It defines an ellipse centered at $\boldsymbol{\mu} = (\mu_1, \mu_2)'$.
- The ellipse's axes are $\pm c\sqrt{\lambda_j}\mathbf{e}_j$ for $j = 1, 2$, where $\lambda_j, \mathbf{e}_j$ are *eigenvalue* and *eigenvector* pairs of $\boldsymbol{\Sigma}$.

II.1.1C. Important Properties of Multivariate Normal Distribution

- If $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{Y} = \mathbf{A}'\mathbf{X} + \mathbf{d} \sim MN_p(\mathbf{A}'\boldsymbol{\mu} + \mathbf{d}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})$.

  e.g. Using $\mathbf{d} = -\boldsymbol{\mu}$ and $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}$,
  $\mathbf{Y} = \mathbf{A}'\mathbf{X} + \mathbf{d} \sim MN_p(\mathbf{0}, \mathbf{I})$.

- $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff \mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ for any $\mathbf{a} \in \mathcal{R}^p$.

  More generally, $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff$
  $\mathbf{A}'\mathbf{X} \sim MN_q(\mathbf{A}'\boldsymbol{\mu}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})$ for any $\mathbf{A} \in \mathcal{R}^{p\times q}$.

The normality is preserved under any linear transformation.

- If $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

  $\mathbf{X}_1 \sim MN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{X}_2 \sim MN_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

- If $\mathbf{X}_1 \sim MN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{X}_2 \sim MN_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent $\Longleftrightarrow$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim MN_{p_1+p_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

- If $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim MN_{p_1+p_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$ with $\boldsymbol{\Sigma}_{22}$ invertible (i.e. $|\boldsymbol{\Sigma}_{22}| > 0$), the conditional distribution $\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2$ is

$$MN_{p_1}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

- If $\mathbf{X} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ invertible, $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p$, the Chi-square distribution with degree of freedom p.

What will we study next?

- *Part I. Introduction and Preparation*
- **Part II. Inference under Multivariate Normal Distribution (Chp 4-7)**
  - **II.1 Multivariate Normal Distribution (Chp 4)**
  - **II.2 Inferences on Mean Vector (Chp 5)**
  - *II.3 Comparisons of Several Mean Vector (Chp 6)*
  - *II.4 Multivariate Linear Regression (Chp 7)*
- *Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-11)*
- *Part IV. Other Topics (Chp 12)*