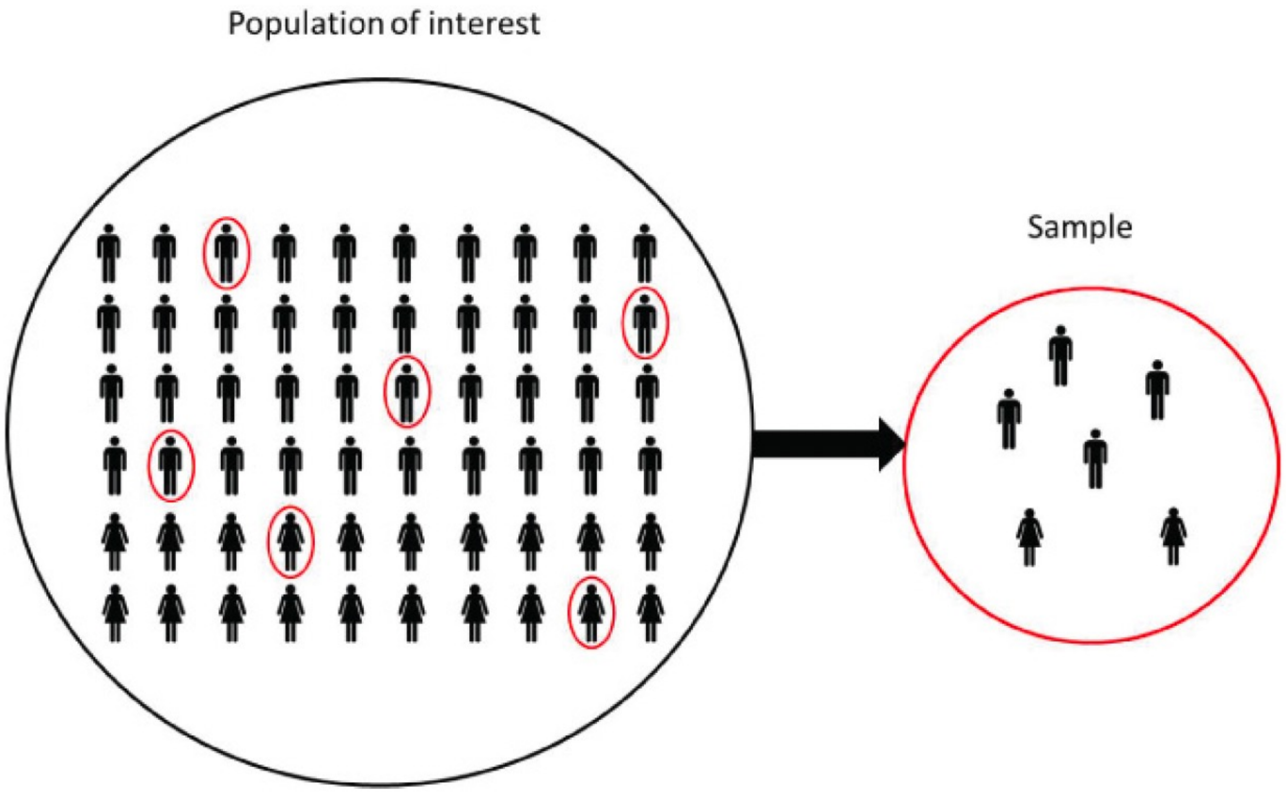
Random vector and matrices I

22/02/2022

Jing Qin

Re-cap: univariate population and sample



$$Y \longrightarrow \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} \longrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

Y_1, Y_2, Y_3, Y_4, Y_5 forms a *random sample* of Y

It takes a *vector* of (univariate) observations to estimate the distribution of a univariate random variable (r.v.).

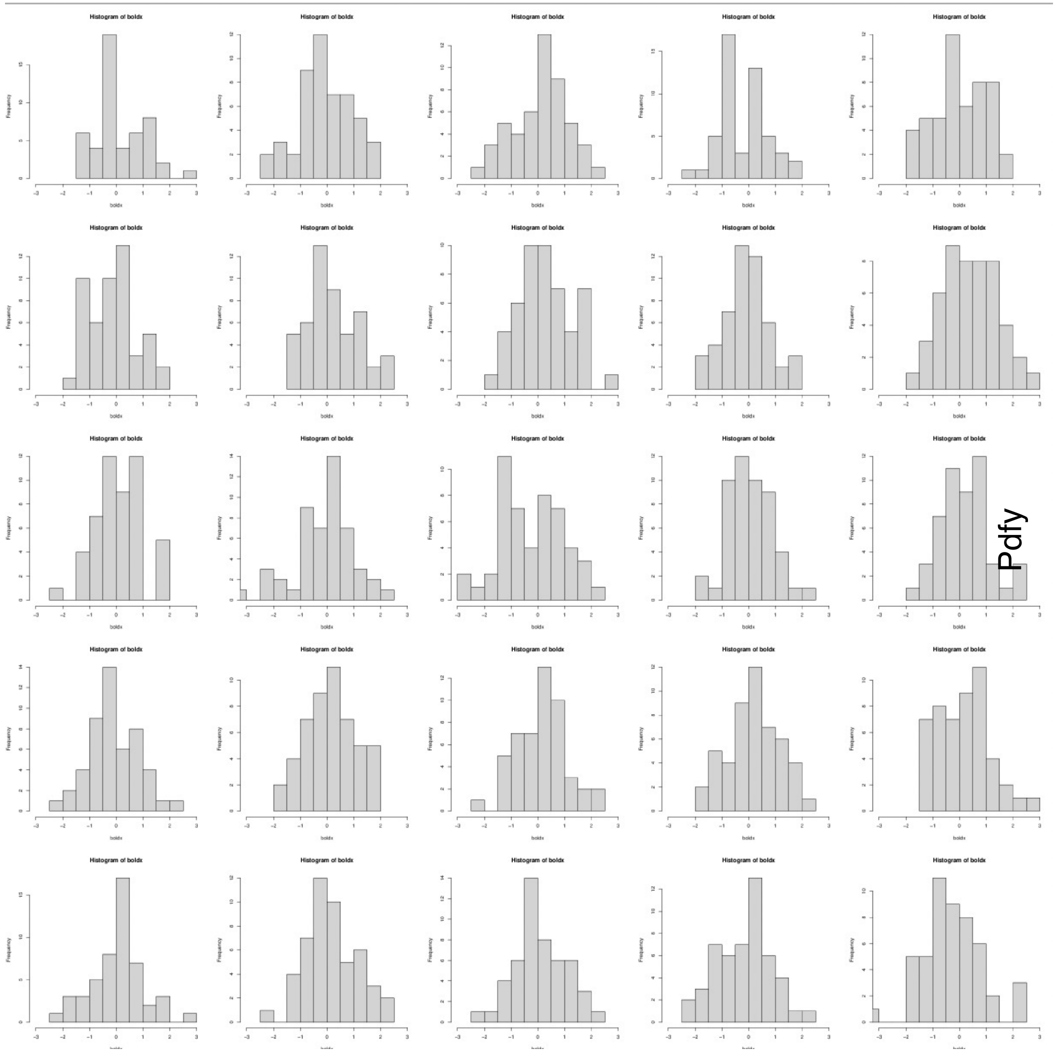
Re-cap: univariate Estimation

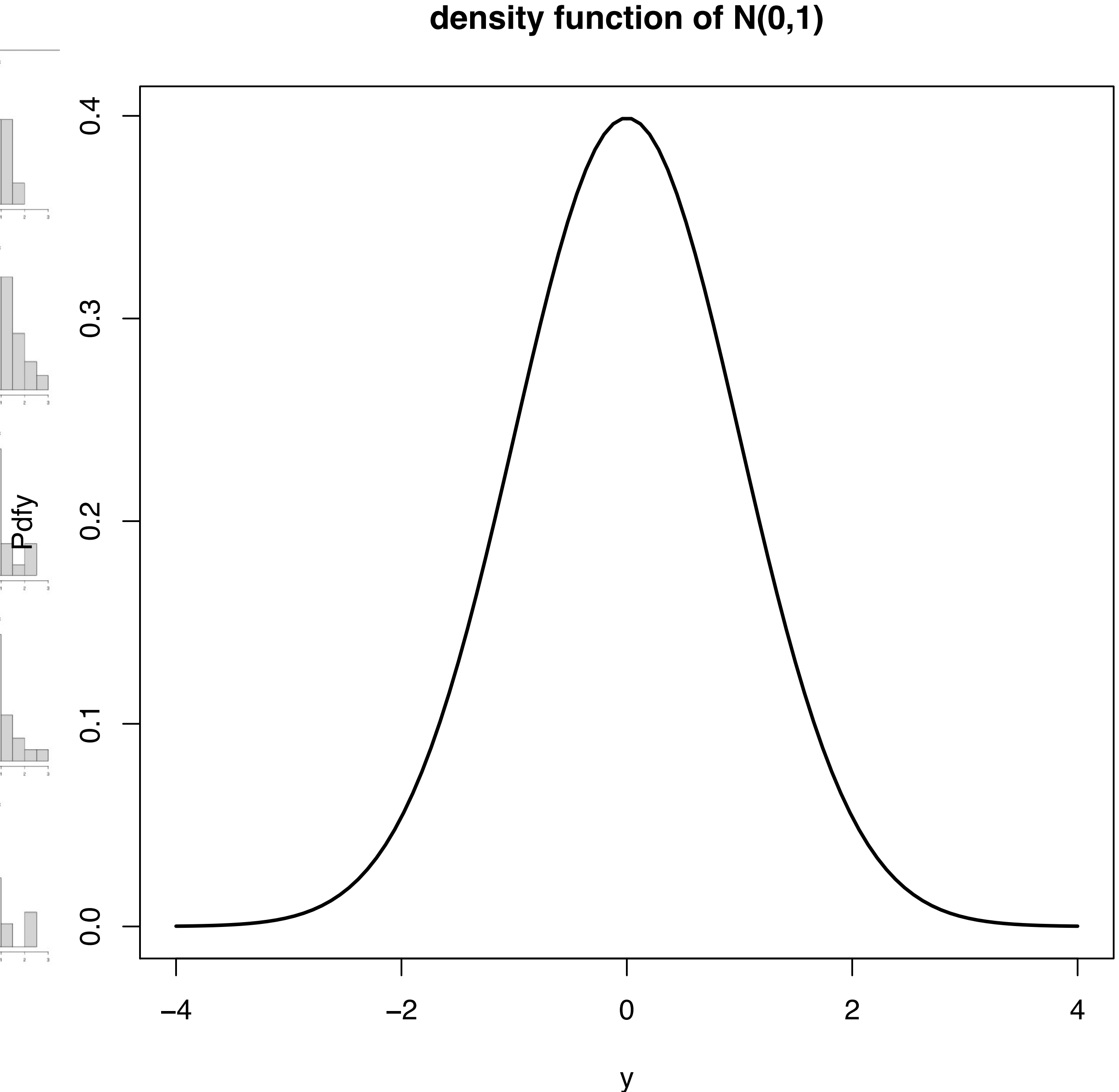
$$Y \sim Bernoulli(0.5) \qquad Y = \begin{cases} 1 & \text{if 'success'} \\ 0 & \text{otherwise} \end{cases}$$

```
> vec1 <- rbinom(5, 1, 0.5)
> vec1
[1] 1 1 1 1 1
> sum(vec1)/length(vec1)
[1] 1
> vec2 <- rbinom(100, 1, 0.5)
> vec2[1:20]
 [1] 1 1 1 1 1 1 1 0 1 0 1 0 1 0 1 1 0 0 1 0
> sum(vec2)/length(vec2)
[1] 0.55
```

Well, ’success’ can be very handy: e.g. ‘observing head of the coin’, ‘a drug is effective’ and ‘a person is vaccinated’ etc…

Re-cap: estimation is *not* a trivial task





Towards multivariate case A question

What if we want to know about the (joint) distribution of a vector of random variables, i.e., *random vector*?

$$\boldsymbol{X} = \left(X_1, X_2, X_3, \ldots, X_p\right)^T$$

What *kind* of data we will need then?

Towards multivariate case A question

What if we want to know about the (joint) distribution of a vector of random variables , i.e., *random vector*?

$$\boldsymbol{X} = \left(X_1, X_2, X_3, \ldots, X_p\right)^T$$

What *kind* of data we will need then?

1. Each random variable costs a vector of *univariate* observations

And?

Towards multivariate case A question

What if we want to know about the (joint) distribution of a vector of random variables, i.e., *random vector*?

$$\boldsymbol{X} = \left(X_1, X_2, X_3, \dots, X_p\right)^T$$

What *kind* of data we will need then?

1. Each random variable costs a vector of *univariate* observations
2. A vector of r.v.'s will cost a vector of *multivariate* observations

Towards multivariate case A question

What if we want to know about the (joint) distribution of a vector of random variables i.e., *random vector*?

$$\boldsymbol{X} = \left(X_1, X_2, X_3, \dots, X_p\right)^T$$

What *kind* of data we will need then?

1. Each random variable costs a vector of *univariate* observations
2. A vector of r.v.'s will cost a vector of *multivariate* observations

matrix

Finally, multivariate case

We need a $(n \times p)$ data matrix to estimate a p-dim random vector

$$\begin{array}{cccccc} X_1 & X_2 & \cdots & X_k & \cdots & X_p \\ \end{array}$$

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\ \vdots & \vdots & & \vdots & \cdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np} \end{bmatrix} \qquad \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix}$$

Compare to univariate

$$Y \longrightarrow \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} \longrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

Multivariate case

We need a $(n{\times}p)$ data matrix to estimate a p-dim random vector

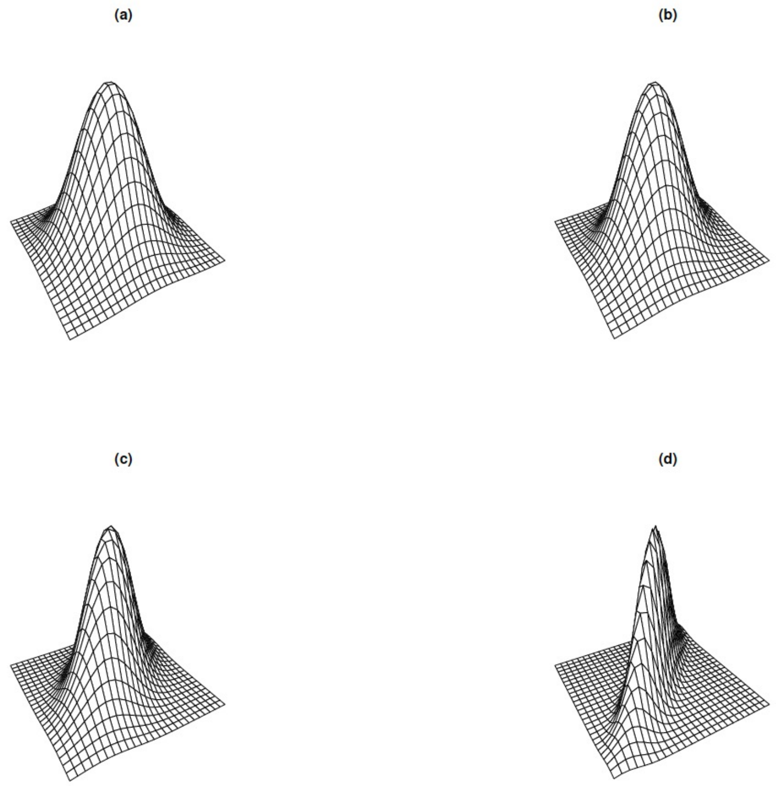
$$
\begin{array}{c c}
 & \begin{array}{cccccc} X_1 & X_2 & \cdots & X_k & \cdots & X_p \end{array} \\
\boldsymbol{X}_j' & \begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np}
\end{bmatrix}
\end{array}
\qquad
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{pmatrix}
$$

$\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ form a *random sample of size n* on $\boldsymbol{X}$. (3-8)

Not so long ago, "Y_1, Y_2, Y_3, Y_4, Y_5 forms a *random sample* of Y"

Multivariate distributions Take a peek

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[x_1^2 + x_2^2 - 2\rho_{12}x_1x_2\right]\right\}$$



Standard bivariate normal distributions' density functions for different values of ρ_{12}

Expected value of random matrix

- (2-23) Expected value of the random matrix $\boldsymbol{X}$ is a matrix as well:

$$
E(\boldsymbol{X}) = \begin{pmatrix}
E(X_{11}) & E(X_{12}) & \cdots & E(X_{1k}) & \cdots & E(X_{1p}) \\
E(X_{21}) & E(X_{22}) & \cdots & E(X_{2k}) & \cdots & E(X_{2p}) \\
\vdots & \vdots & & \vdots & & \vdots \\
E(X_{j1}) & E(X_{j2}) & \cdots & E(X_{jk}) & \cdots & E(X_{jp}) \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
E(X_{n1}) & E(X_{n2}) & \cdots & E(X_{nk}) & \cdots & E(X_{np})
\end{pmatrix}
$$

$$
E(\boldsymbol{X}) = (E(X_{j,k}))_{(n \times p)}
$$

- Exercise: P67 Exp(2.12)

Properties of expected values linearity

(2-24) $E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$ and $E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$

$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ and a numerical vector $\boldsymbol{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}$, we have $E(\mathbf{c}' \cdot \boldsymbol{X}) = \mathbf{c}' \cdot E(\boldsymbol{X})$.

Why previous result matters?

Example 2-1: Women's Health Survey (Linear Combinations)

The Women's Health Survey data contains observations for the following variables:

- X_1 calcium (mg)
- X_2 iron (mg)
- X_3 protein(g)
- X_4 vitamin A(μg)
- X_5 vitamin C(mg)

In addition to addressing questions about the individual nutritional component, we may wish to address questions about certain combinations of these components. For instance, we might want to ask what is the total intake of vitamins A and C (in mg). We note that in this case Vitamin A is measuring in micrograms while Vitamin C is measured in milligrams. There are a thousand micrograms per milligram so the total intake of the two vitamins, *Y*, can be expressed as the following:

$$Y = 0.001X_4 + X_5$$

In this case, our coefficients c_1 , c_2 and c_3 are all equal to 0 since the variables X_1, X_2 and X_3 do not appear in this expression. In addition, c_4 is equal to 0.001 since each microgram of vitamin A is equal to 0.001 milligrams of vitamin A. In summary, we have

$$c_1 = c_2 = c_3 = 0, c_4 = 0.001, c_5 = 1$$

When the rows are independent and identially distributed random vectors

$$
\begin{array}{c}
\begin{array}{cccccc} X_1 & X_2 & \cdots & X_k & \cdots & X_p \end{array} \\
\begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np}
\end{bmatrix}
\end{array}
$$

$$
\begin{pmatrix}
E(X_{11}) = \mu_1 & E(X_{12}) = \mu_2 & \cdots & E(X_{1p}) = \mu_p \\
E(X_{21}) = \mu_1 & E(X_{22}) = \mu_2 & \cdots & E(X_{2p}) = \mu_p \\
\vdots & \vdots & \vdots & \vdots \\
E(X_{j1}) = \mu_1 & E(X_{j2}) = \mu_2 & \cdots & E(X_{jp}) = \mu_p \\
\vdots & \vdots & \vdots & \cdots \\
E(X_{n1}) = \mu_1 & E(X_{n2}) = \mu_2 & \cdots & E(X_{np}) = \mu_p
\end{pmatrix}
$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$ is referred to as (population) mean vector

Population covariance matrix $\mathbf{\Sigma}$

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$.

- Its covariance matrix is defined in (2-31) and denoted by Σ or $\mathrm{Cov}(\boldsymbol{X})$.

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) = \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Cov}(X_p, X_p) \end{pmatrix}$$

- Recall that $\mathrm{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$, therefore

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$$

and thus Σ is symmetric.

Properties

- Useful results

  1. $\text{Cov}(X_j, X_j) = \text{Var}(X_j)$
  2. $\text{Cov}(X_i, X_j) = 0$ does not indicate two random variables are independent.
  3. $\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$

- (2-32) $\text{Cov}(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu}) \cdot (\boldsymbol{X} - \boldsymbol{\mu})']$

- (2-45) Let $\boldsymbol{C}$ be a numeric $(q \times p)$ matrix, then we have

$$E(\boldsymbol{C} \cdot \boldsymbol{X}) = \boldsymbol{C} \cdot \boldsymbol{\mu}$$

and

$$\text{Cov}(\boldsymbol{C} \cdot \boldsymbol{X}) = \boldsymbol{C} \cdot \Sigma \cdot \boldsymbol{C}'.$$