Reading: 2.5 & 2.6

A random variable is a variable that assigns a numerical value to each outcome in the sample space.

The values of a r.v are determined by the outcomes of a random experiment.

Ex

Toss a die once

RV: X = # of dots on the die

Possible values: $S = \{1, 2, 3, 4, 5, 6\}$.

* A function that maps each outcome in a sample space to a real number line.

Random Vector

- A vector whose elements are are random variables

Eg: ① Biology: (weight (X_1), height (Y_2), blood pressure (X_3)) of an individual

② Spatial data: Record the temperature at different locations at the same ~~in~~ of the a day for different days $\underset{\sim}{I} = (t_1, t_2, \dots, t_5)$

③ Survival data: Failure time (time 1, time 2, time 3) of different objects.

A random vector $\underset{\sim}{X} = (X_1, X_2, \dots, X_p)_{p\times 1}$ and each element of X_i is a r.v on its own with its own marginal distribution.

- Define the following marginal parameters:

① $E(X_i) = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous r.v with pdf } f_i(x_i) \\ \sum_{\text{all } x} x_i p_i(x_i) & \text{if } X_i \text{ " discrete " pmf } p_i(x_i) \end{cases}$

② $V(X_i) = E[(X_i - \mu_i)^2] = \sigma_{ii} = \sigma_i^2$

③ For 2 r.variables X_i, X_k, $\text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \cdot)] = \sigma_{ij}$; $i \neq j$ → A measure of linear association between the variables X_i, X_j.

- If X_i and X_j are independent, $\text{Cov}(X_i, X_j) =$. The converse is generally not true except in the case of Multivariate normal distribution
- $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$

  $\sigma_{ij} = \sigma_{ji}$
- The behaviour of any pair of random variables, such as X_i, X_j is described by their joint pdf.

Revise joint distributions.

*For a random vector $\underset{\sim}{X} = (X_1, X_2, \dots, X_p)_{p\times 1}$

① $\underset{\sim}{\mu} = E(\underset{\sim}{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$ = population mean vector . $\mu_{\underset{\sim}{X}}$ $\Sigma_{\underset{\sim}{X}}$

② Population Variance Covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \\ \sigma_{p1} & \cdots & & \sigma_{pp} \end{pmatrix}_{p\times p} = E\left((\underset{\sim}{X} - \underset{\sim}{\mu})^{p\times 1} (\underset{\sim}{X} - \underset{\sim}{\mu})^{T_{1\times p}} \right)$$

$$= E\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & & \\ \vdots & & \ddots & \\ (X_p-\mu_p)(X_1-\mu_1) & & & (X_p-\mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1-\mu_1)^2 & & \\ & & \\ & & \end{bmatrix}$$

- In a dataset of p variables, there will be p variances and $\frac{p(p-1)}{2}$ distinct covariances.

(3) From the Σ, we can obtain the correlation matrix as

$$\rho = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \vdots & \ddots & & \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \cdots & & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \vdots & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{p1} & \cdots & & 1 \end{bmatrix}$$

Note that if $V^{1/2} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\sigma_{pp}} \end{pmatrix}_{p\times p}$ then

$$\rho = V^{-1/2}\,\Sigma\,V^{-1/2}$$
$$= (V^{1/2})^{-1}\,\Sigma\,(V^{1/2})^{-1}$$

and

$$\Sigma = V^{1/2}\,\rho\,V^{1/2}$$

↑ Standard deviation matrix

Example 1

Find the covariance matrix for the random vector (X_1, X_2), when their pmf $p_{12}(x_1, x_2)$ is represented by

x_1 \ x_2	0	1	$p_1(x_1)$
−1	0.24	0.06	0.3
0	0.16	0.14	0.3
1	0.4	0.00	0.4
$p_2(x_2)$	0.8	0.2	1

Solution.

$E(X_1) = -1(0.3) + 0(0.3) + 1(0.4) = 0.1$

$E(X_2) = 0(0.8) + 1(0.2) = 0.2$

$\sigma_{11} = E(X_1-\mu_1)^2 = \sum_{\text{all } x_1} (x_1-0.1)^2 p_1(x_1) = (-1-0.1)^2(0.3) + (0-0.1)^2(0.3) + (1-0.1)^2(0.4) = 0.69$

$\sigma_{22} = E(X_2-\mu_2)^2 = \sum_{\text{all } x_2} (x_2-0.2)^2 p_2(x_2) = (0-0.2)^2(0.8) + (1-0.2)^2(0.2) = 0.16$

$\sigma_{12} = \sigma_{21} = E[(X_1-\mu_1)(X_2-\mu_2)] = \sum_{\substack{\text{all pairs} \\ (x_1,x_2)}} (x_1-0.1)(x_2-0.2)\,p_{12}(x_1,x_2)$

$= (-1-0.1)(0-0.2)(0.24) + (-1-0.1)(1-0.2)(0.06) + \ldots + (1-0.1)(1-0.2)(0.00)$

$= -0.08$

So $\underset{\sim}{\mu} = E(\underset{\sim}{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}$

$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 0.69 & -0.08 \\ -0.08 & 0.16 \end{pmatrix}$

Example 2

Given the covariance matrix below, find $V^{1/2}$ and ρ.

$$\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix}$$

Solution

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 \\ 0 & 0 & \sqrt{\sigma_{33}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

$$(V^{1/2})^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix}$$

$$\rho = (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}$$

$$= \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1/6 & 1/5 \\ 1/6 & 1 & -1/5 \\ 1/5 & -1/5 & 1 \end{bmatrix}$$

Random Matrix

A random matrix is a matrix whose elements are random variables

Define

$$\underset{n \times p}{\underline{X}} = \{X_{ij}\} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & & & \\ X_{n1} & \cdots & & X_{np} \end{pmatrix}_{n \times p}$$

a) $E(\underline{X}) = \{E\{X_{ij}\}\}$ if each expectation exists.

b) Let $\underline{X}$ and $\underline{Y}$ be random matrices of the same dimension and let $\underline{A}$ and $\underline{B}$ be conformable matrices of constants

$E(\underline{X} + \underline{Y}) = E(\underline{X}) + E(\underline{Y})$ - Linear property of expectation

$E(\underset{k \times n}{\underline{A}} \underset{n \times m}{\underline{X}} \underset{m \times}{\underline{B}}) = \underline{A} E(\underline{X}) \underline{B}$ ($k \times l$) (See example 2.39, 2.40)

Partitionin rando vectors

Given a large dataset, characteristics measured on individual trials could naturally fall into two or more roups.

Example 1: Consumpti n and inoome

Imagine a survey conducted on a group of individuals, where you collet data on both consumption habits - spending on groceries, entertainment and utilities, while income would encompass variables like salary, investments and other sources of earnings.

Categorize into 2 groups: consumption habits & income.

Example 2 : Personality traits and physical characteristics.

Variables measured : Extroversion, agreeableness, conscientiousness, height, weight, body mass index.

} 2 groups
a) Personality traits
b) Physical characteristics

To handle situations like these, where distinct groups of characteristics arise naturally, a useful approach is to consider these groups as subsets of a larger collection of characteristics.

- If the total collection is represented by a $p \times 1$ dimension random vector $\underset{\sim}{X}_{p\times 1}$. We can regard the subsets as components of $\underset{\sim}{X}$ and can be dealt with by partitioning

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \\ \hline X_{q+1} \\ \vdots \\ X_p \end{bmatrix} \begin{matrix} \} q\times 1 \\ \} (p-q)\times 1 \end{matrix} = \begin{bmatrix} \underset{\sim}{X}^{(1)} \\ \underset{\sim}{X}^{(2)} \end{bmatrix}$$

(a) Then $\underset{\sim}{\mu} = E(\underset{\sim}{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \\ \hline \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mu}^{(1)} \\ \underset{\sim}{\mu}^{(2)} \end{bmatrix}$

(b) $\underset{p\times p}{\Sigma} = E\left[(\underset{\sim}{X}-\underset{\sim}{\mu})(\underset{\sim}{X}-\underset{\sim}{\mu})^T\right] = E\begin{bmatrix} \begin{pmatrix} \underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)} \\ \underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)} \end{pmatrix} \left(\underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)} \quad \underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)}\right) \end{bmatrix}$

$$= E\begin{bmatrix} (\underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)})(\underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)})^T & (\underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)})(\underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)})^T \\ \cdots & (\underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)})(\underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)})^T \end{bmatrix}$$

$$= \begin{matrix} q \\ p-q \end{matrix}\begin{bmatrix} \overset{q}{\underset{q\times q}{\Sigma_{11}}} & \overset{p-q}{\underset{q\times(p-q)}{\Sigma_{12}}} \\ \underset{(p-q)\times q}{\Sigma_{21}} & \underset{(p-q)\times(p-q)}{\Sigma_{22}} \end{bmatrix}$$

Where Σ_{11} = Variance covariance matrix for $\underset{\sim}{X}^{(1)}$

Σ_{22} = " " $\underset{\sim}{X}^{(2)}$

$\Sigma_{12} = E\left[(\underset{\sim}{X}^{(1)}-\underset{\sim}{\mu}^{(1)})(\underset{\sim}{X}^{(2)}-\underset{\sim}{\mu}^{(2)})^T\right] = \Sigma_{21}^T$

Linear Combinations of random vectors

- Let X_1 be a single random variable and c be a constant with mean and variance μ_1 and σ_{11}
  - (a) $E(cX_1) = c\mu_1$
  - (b) $V(cX_1) = c^2\sigma_{11} = c^2 V(X_1)$

with $E(X_1)=\mu_1, E(X_2)=\mu_2$, $V(X_1)=\sigma_{11}, V(X_2)=\sigma_{22}$, $Cov(X_1,X_2)=\sigma_{12}$

- for 2 random variables X_1 and X_2, and for constants a and b,

  (a) $E(aX_1 + bX_2) = a\mu_1 + b\mu_2$ Note aX_1+bX_2 is a linear combination of X_1 and X_2

  (b) $Cov(aX_1, bX_2) = ab\,Cov(X_1,X_2) = ab\,\sigma_{12}$

Proof:

(c) $V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2ab\,Cov(X_1,X_2)$ Prove in HW 2

$= a^2\sigma_{11} + b^2\sigma_{22} + 2ab\,\sigma_{12}$

- $aX_1 + bX_2 = \underset{a\mu_1+b\mu_2=}{(a\ \ b)}\begin{pmatrix}X_1\\X_2\end{pmatrix} = \underset{\sim}{c}^T\underset{\sim}{X}$

  > $E(aX_1+bX_2) = E\left[(a\ b)\begin{pmatrix}X_1\\X_2\end{pmatrix}\right] = (a\ b)\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix} = \underset{\sim}{c}^T\underset{\sim}{\mu}$

  > If we let $\Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$

  then

$$V(aX_1+bX_2) = V(\underset{\sim}{c}^T\underset{\sim}{X})$$
$$= V\left[(a\ b)\begin{pmatrix}X_1\\X_2\end{pmatrix}\right]$$
$$= \underset{\sim}{c}^T\Sigma\,\underset{\sim}{c}$$
$$= (a\ b)\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix}$$
$$= a^2\sigma_{11} + 2ab\,\sigma_{12} + b^2\sigma_{22}$$

- We can extend this result for q linear combination of p variables

$$\underset{\sim}{c}^T\underset{\sim}{X} = c_1X_1 + c_2X_2 + \ldots + c_pX_p$$

Then $E(\underset{\sim}{c}^T\underset{\sim}{X}) = \underset{\sim}{c}^T\underset{\sim}{\mu}$; $\mu = E(X)$ and $\Sigma = Cov(X)$

$V(\underset{\sim}{c}^T\underset{\sim}{X}) = \underset{\sim}{c}^T\Sigma\,\underset{\sim}{c}$

- Even more generally, we can consider q linear combinations of p random variables $X_1, \ldots, X_p$

$$Z_1 = c_{11}X_1 + c_{12}X_2 + \ldots + c_{1p}X_p$$
$$Z_2 =$$
$$\vdots$$
$$Z_q = c_{q1}X_1 + c_{q2}X_2 + \ldots + c_{qp}X_p$$

So that

$$\underset{\sim}{Z} = \begin{bmatrix}c_{11} & c_{12} & \ldots & c_{1p}\\ c_{21} & c_{22} & \ldots & c_{2p}\\ \vdots & & & \\ c_{q1} & c_{q2} & \ldots & c_{qp}\end{bmatrix}\begin{pmatrix}X_1\\\vdots\\X_p\end{pmatrix} = \underline{C}\,\underset{\sim}{X}$$

Mean of $\underset{\sim}{Z} = E(\underset{\sim}{Z}) = E(\underset{\sim}{C}\underset{\sim}{X}) = \underset{\sim}{C}\underset{\sim}{\mu}$

$\Sigma_Z = Cov(\underset{\sim}{Z}) = Cov(\underset{\sim}{C}\underset{\sim}{X}) = \underset{\sim}{C}\Sigma_x\underset{\sim}{C}^T$

$(AB)^T = B^TA^T$

Proof

$$Cov(\underset{\sim}{C}\underset{\sim}{X}) = E\left[(\underset{\sim}{C}\underset{\sim}{X} - \underset{\sim}{C}\underset{\sim}{\mu})(\underset{\sim}{C}\underset{\sim}{X} - \underset{\sim}{C}\underset{\sim}{\mu})^T\right]$$

$$= E\left[\underset{\sim}{C}(\underset{\sim}{X}-\underset{\sim}{\mu})(\underset{\sim}{X}-\underset{\sim}{\mu})^T\underset{\sim}{C}^T\right] \qquad (C(\underset{\sim}{X}-\underset{\sim}{\mu}))^T$$

$$= \underset{\sim}{C}E\left[(\underset{\sim}{X}-\underset{\sim}{\mu})(\underset{\sim}{X}-\underset{\sim}{\mu})^T\right]\underset{\sim}{C}^T$$

$$= \underset{\sim}{C}\,\Sigma_x\,\underset{\sim}{C}^T$$

* We will rely on these results when discussing principal component analysis and factor analysis.

Example

Find μ_Z and Σ_Z for $\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ if

$Z_1 = X_1 - X_2$

$Z_2 = X_1 + X_2$

if $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\underset{\sim}{\mu}_x = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$,

$$\Sigma_x = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

Solution

$$\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$E(\underset{\sim}{Z}) = \underset{\underset{\sim}{C}}{\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}}\underset{\Sigma_x}{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}} = \underset{\underset{\sim}{C}^T}{\begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 + \mu_2 \end{pmatrix}}$$

$$V(\underset{\sim}{Z}) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11}-\sigma_{12} & \sigma_{12}-\sigma_{22} \\ \sigma_{11}+\sigma_{12} & \sigma_{12}+\sigma_{22} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11}-\sigma_{12}-\sigma_{12}+\sigma_{22} & \sigma_{11}-\sigma_{12}+\sigma_{12}-\sigma_{22} \\ \sigma_{11}+\sigma_{12}-\sigma_{12}-\sigma_{22} & \sigma_{11}+\sigma_{12}+\sigma_{12}+\sigma_{22} \end{pmatrix}$$

$$\begin{pmatrix} \sigma_{11}-2\sigma_{12}+\sigma_{22} & \sigma_{11}-\sigma_{22} \\ \sigma_{11}-\sigma_{22} & \sigma_{11}+2\sigma_{12}+\sigma_{22} \end{pmatrix}$$

Note

Just as we partitioned the population variance/covariance matrix and population mean vector, we can also partition the sample mean vector and sample covariance matrix in the same way to distinguish quantities corresponding to groups of variables. Thus

let $\underset{\sim}{\bar{X}}_{p\times 1} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_q \\ \bar{X}_{q+1} \\ \vdots \\ \bar{X}_p \end{pmatrix}_{q\times 1} = \begin{bmatrix} \bar{X}^{(1)} \\ \cdots \\ \bar{X}^{(2)} \end{bmatrix}_{p-q}^{q}$ be a vector of sample averages constructed from n observations on p variables $X_1, \ldots, X_p$.

and let

$$\underline{S}_n = \begin{pmatrix} S_{11} & S_{12} & S_{1,q+1} & \cdots & S_{1p} \\ \vdots & & & & \\ S_{q1} & \cdots & S_{qq} & S_{q,q+1} & \cdots & S_{qp} \\ S_{q+1,1} & & S_{q+1,q} & S_{q+1,q+1} & \cdots & S_{q+1,p} \\ \vdots & & & \vdots & & \\ S_{p1} & & S_{pq} & S_{p,q+1} & & S_{pp} \end{pmatrix}$$ be the corresponding sample variance covariance matrix.

Section 3.3 Sample Statistics.

Define $\underset{\sim}{X}_{n\times p} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} =$

- If the row vectors $\underset{\sim}{X}_1^T, \underset{\sim}{X}_2^T, \ldots, \underset{\sim}{X}_n^T$ represent independent observations from a common joint distribution with density $f(\underset{\sim}{x}) = f(x_1, \ldots, x_p)$, then $X_1, X_2, \ldots, X_n$ are said to form a random sample from $f(\underset{\sim}{x})$

  $f(\underset{\sim}{x}) = f(x_1)\, f(\,) \ldots f(x_n)$ where $f(x_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$ is the density function for the jth row

- Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ be a random sample from a joint distribution that has mean μ and variance cov matrix Σ

① Define $\underset{\sim}{\bar{X}}_{p\times 1} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X}_i$ = Sample mean vector

ⓐ Then $\underset{\sim}{\bar{X}}_{p\times 1}$ is an unbiased estimator of μ that is,

$$E(\underset{\sim}{\bar{X}}_{p\times 1}) = \underset{\sim}{\mu}_{p\times 1}$$

ⓑ $\mathrm{Cov}(\bar{x}_{p\times 1}) = \frac{1}{n}\Sigma_x$ (ie the population variance covariance matrix divided by sample size)

② Define the sample covariance matrix as S_n

$$S_n = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) \\ \vdots & & \\ \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_p)^2 \end{pmatrix}$$

ⓐ $E(S_n) = \frac{n-1}{n}\Sigma = \Sigma - \frac{1}{n}\Sigma$ so that

$E\left(\frac{n}{n-1}S_n\right) = \Sigma$, so $\frac{n}{n-1}S_n$ is an unbiased estimator of Σ

$\frac{n}{n-1}S_n$ = $\frac{1}{n-1}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p)$ for the nth row.

Sample mean vector, covariance and correlation as matrix operations

Define $\underset{n\times p}{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & & & \\ \vdots & & & \\ x_{n1} & & & x_{np} \end{pmatrix}$

$X^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & & & \\ \vdots & & & \\ x_{1p} & \cdots & & x_{np} \end{pmatrix}$

① $\underset{p\times 1}{\bar{X}} = \frac{1}{n} X^T \underset{n\times 1}{1}$

② Recall $S = \frac{n}{n-1} S_n$ ✓ → $x^T A x$ quadratic forms

$= \frac{1}{n-1} X^T \left(I - \frac{1}{n} 1 1^T \right) X$

where S_n is the biased sample covariance matrix.

$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & & & \\ \vdots & & & \\ \frac{1}{n} & & & \frac{1}{n} \end{pmatrix} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & & \\ \vdots & & \ddots & \\ -\frac{1}{n} & & & 1-\frac{1}{n} \end{pmatrix}$

$\frac{1}{n} \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & & & \\ \vdots & & & \\ x_{1p} & & & x_{np} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & & & \\ \vdots & & & \\ x_{n1} & \cdots & & x_{np} \end{pmatrix} = \begin{pmatrix} \frac{1}{n-1}\sum (x_i-\bar{x}_1)^2 & & & \\ & \ddots & & \\ \frac{1}{n-1}\sum (x_i-\bar{x}_1)(x_i-\bar{x}_p) & & & \frac{1}{n-1}(x_i-\bar{x}_p)^2 \end{pmatrix}$

③ $R = D^{-1/2} S_n D^{-1/2}$

$D^{1/2} = \begin{pmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & & \\ \vdots & & \ddots & \\ 0 & & & \sqrt{s_{nn}} \end{pmatrix}$

$\begin{matrix} & x_1 & x_2 & \cdots \end{matrix}$

$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & & & \\ \vdots & & & \\ x_{n1} & & & \end{pmatrix} - \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & & \bar{x}_p \\ \bar{x}_1 & \vdots & & \\ \vdots & & & \\ \bar{x}_1 & \bar{x}_2 & & \bar{x}_p \end{pmatrix} = \begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & \vdots & & \vdots \\ \vdots & & & \\ x_{n1}-\bar{x}_1 & & & x_{pp}-\bar{x}_p \end{pmatrix}$

$\underset{}{X}$ $\quad - \frac{1}{n} X^T 1 1^T$

$\underset{p\times n}{\begin{pmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ \vdots & & & \\ x_{1p}-\bar{x}_p & \cdots & & x_{np}-\bar{x}_p \end{pmatrix}} \quad \begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & & & \vdots \\ \vdots & & & \\ x_{n1}-\bar{x}_1 & & & x_{np}-\bar{x}_p \end{pmatrix}$

$\left(X - \frac{1}{n} X^T 1 1^T \right)^T \left(X - \frac{1}{n} X^T 1 1^T \right)$

$= X^T \left(I - \frac{1}{n} 1 1^T \right) \left(I - \frac{1}{n} 1 1^T \right) X$

$(n-1)\, S = X^T \left(I - \frac{1}{n} 1 1^T \right) X$

If $A^T A = A$, A is an idempotent matrix

Sample values of linear combinations of variables

Consider a linear combination of vectors:

$c' X = c_1 X_1 + c_2 X_2 + \cdots c_p X_p$ where $X_1, X_2, \ldots, X_p$ are independent realizations from the random vector.

Recall

$$E(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}'\underset{\sim}{\mu} \qquad Var(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}^T\Sigma\,\underset{\sim}{c}$$

Suppose we want to estimated these,

Let $Y_i = \underset{\sim}{c}^T\underset{\sim}{X}_i$

① An estimate of the mean $\underset{\sim}{c}^T\underset{\sim}{\mu}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{c}^T\underset{\sim}{X}_i = \underset{\sim}{c}^T\left(\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{X}_i\right) = \underset{\sim}{c}'\underset{\sim}{\bar{X}}_{p\times 1}$$

② An estimate of the variance $\underset{\sim}{c}^T\Sigma\,\underset{\sim}{c}$ is given by

$$\begin{aligned}\frac{1}{n}\sum(Y_i-\bar{Y})^2 &= \frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{c}^T\underset{\sim}{X}_i - \underset{\sim}{c}^T\underset{\sim}{\bar{X}})^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{c}^T(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}}))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\underset{\sim}{c}^T(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})^T\underset{\sim}{c}\right)\\ &= \underset{\sim}{c}^T\left[\frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})^T\right]\underset{\sim}{c}\\ &= \underset{\sim}{c}^T S_n\,\underset{\sim}{c}\end{aligned}$$

③ An estimate for the popupulation covariance matrix of $\underset{\sim}{b}'\underset{\sim}{X}$, $\underset{\sim}{c}'\underset{\sim}{X}$ which is $Cov(\underset{\sim}{b}^T\underset{\sim}{X}, \underset{\sim}{c}^T\underset{\sim}{X}) = \underset{\sim}{b}^T\Sigma\,\underset{\sim}{c}$ is given by $\underset{\sim}{b}^T S_n\,\underset{\sim}{c}$.

Example

Given $\underset{n\times p}{\underline{X}} = \underset{3\times 3}{\begin{pmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{pmatrix}}$ (X_1 X_2 X_3). Find the following:

(a) $\underset{3\times 1}{\bar{X}}$

(b) S_n

(c) Consider the linear combinations $2X_1+2X_2-X_3$ and $X_1-X_2+3X_3$

i) Find their means and variances respectively.

ii) Determine $Cov(2X_1+2X_2-X_3,\ X_1-X_2+3X_3)$

Solution

(a) $\underset{3\times 1}{\bar{X}} = \frac{1}{n}\underline{X}^T\underset{\sim}{1} = \frac{1}{3}\begin{pmatrix} 1 & 4 & 4 \\ 2 & 1 & 0 \\ 5 & 6 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 9 \\ 3 \\ 15 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix}$

(b) $S_n = \frac{n-1}{n}S = \frac{1}{n}\underline{X}^T\left(I-\frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right)\underline{X}$

$$= \frac{1}{3}\begin{pmatrix} 1 & 4 & 4 \\ 2 & 1 & 0 \\ 5 & 6 & 4 \end{pmatrix}\begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}\begin{pmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{pmatrix}$$

$$= \frac{1}{3}\begin{pmatrix} -2 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

$$= \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}$$

$$\begin{pmatrix} 1-\frac{1}{n} & - & - \\ - & - & - \\ - & - & - \end{pmatrix}$$

$$\begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

c) Let $\underset{\sim}{b}^T\underset{\sim}{X} = 2X_1 + 2X_2 - X_3 = \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ and $\underset{\sim}{c}^T\underset{\sim}{X} = X_1 - X_2 + 3X_3 = \begin{pmatrix} 1 & -1 & 3 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

Sample mean of $\underset{\sim}{b}^T\underset{\sim}{X} = \underset{\sim}{b}^T\underset{\sim}{\bar{x}}$

$= \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix} = 3$

" $\underset{\sim}{c}^T\underset{\sim}{X} = \underset{\sim}{c}^T\underset{\sim}{\bar{x}}$

$= \begin{pmatrix} 1 & -1 & 3 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} = 17$

Sample variance of $\underset{\sim}{b}^T\underset{\sim}{X} = \underset{\sim}{b}^T S_n \underset{\sim}{b}$

$= \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}$

$= \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} = 2$

Sample variance of $\underset{\sim}{c}^T\underset{\sim}{X} = \underset{\sim}{c}^T S_n \underset{\sim}{c}$

$= \begin{pmatrix} 1 & -1 & 3 \end{pmatrix}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 3 \end{pmatrix}\begin{pmatrix} 3 \\ -2/3 \\ 5/3 \end{pmatrix}$

$= 26/3$

Sample Cov of $\underset{\sim}{b}^1\underset{\sim}{X}$ and $\underset{\sim}{c}^1\underset{\sim}{X}$

$\text{Cov}(\underset{\sim}{b}^1\underset{\sim}{X}, \underset{\sim}{c}^1\underset{\sim}{X}) = \underset{\sim}{b}^1 S_n \underset{\sim}{c}$

$= \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}$

$= \begin{pmatrix} 2 & 2 & -1 \end{pmatrix}\begin{pmatrix} 3 \\ -2/3 \\ 5/3 \end{pmatrix}$

$= 3$

- The results disaussed for a single linear combination can be extended to any number of linear combinations.
- Consider q linear combinations which is expressed in matrix notation as

$$\begin{pmatrix} C_{11}\underset{\sim}{X}_1 + C_{12}\underset{\sim}{X}_2 + \cdots + C_{1p}\underset{\sim}{X}_p \\ C_{21}\underset{\sim}{X}_1 + \cdots \\ \vdots \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & & & \\ C_{q1} & \cdots & & C_{qp} \end{pmatrix}\begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \\ \vdots \\ \underset{\sim}{X}_p \end{pmatrix} = \underset{qxp}{\underline{C}}\ \underset{\sim}{X}_{p\times 1}$$

Let the ith row of $\underline{C}$ be $\underset{\sim}{c}_i^T$ and the jth row be $\underset{\sim}{c}_j^T$. From previous results, the ith row has sample mean $\underset{\sim}{c}_i$ and the i,jth element of the sample variance covariance, $\underset{\sim}{c}_i^T S_n \underset{\sim}{c}_j$.

- $\underline{C}\underset{\sim}{X}$ will have sample mean $\underline{C}\underset{\sim}{\mu}$ and sample covariance $\underline{C} S_n \underline{C}^T$

Ch 4
Univariate normal distribution

Density:

$$\phi(x) = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

CDF

$$P(X \le x) = \Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$$

Properties

It is symmetric so $\Phi(-x) = 1 - \Phi(x)$ for every value x.

Visualize normality:

Histogram, qqplot.

Tests for normality

Shapiro wilk, Kolmogorov smirnov

Ho: Data is normally distributed

Ha: Data is not normally distributed.

Multivariate normal

- Many multivariate techniques to be discussed in this class depend on the MVN assumption
- $p \ge 2$ dimensions
- Density function for a p-dimensional MVN random vector $\underset{\sim}{X} = (X_1, \cdots, X_p)$ has the form

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{ \frac{-1}{2} (\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu}) \right\}; \quad -\infty < x_i < \infty \text{ for } i = 1, \cdots, p$$

We say that $\underset{p\times 1}{\underset{\sim}{X}} \sim N_p(\underset{p\times 1}{\underset{\sim}{\mu}}, \underset{p\times p}{\Sigma})$

- When $p=2$, we have a bivariate normal distribution

Given a random vector $\underset{\sim}{X} = (X_1, X_2)$, let $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$. Suppose Σ is invertible ie,

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{pmatrix}$$. We know $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$ so $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$

$$= \frac{1}{\sigma_{11}\sigma_{22} - \rho_{12}^2\sigma_{11}\sigma_{22}}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{pmatrix}$$

We can show that the density function for the bivariate normal is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}\exp\left\{\frac{-1}{2(1-\rho_{12})^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

Homework:
Show that when X_1 and X_2 are uncorrelated, this joint density can be written as the product of 2 univariate normal densities.

Properties of MVN distribution

Suppose $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_{p\times 1}, \Sigma)$, then

1. Any linear combination of variables $\underset{\sim}{a}^T\underset{\sim}{X} = a_1X_1 + a_2X_2 + \dots a_pX_p$ is said to follow $N(\underset{\sim}{a}^T\underset{\sim}{\mu}, \underset{\sim}{a}^T\Sigma\underset{\sim}{a})$
2. If $\underset{\sim}{a}^T\underset{\sim}{X} \sim N(\underset{\sim}{a}^T\underset{\sim}{\mu}, \underset{\sim}{a}^T\Sigma\underset{\sim}{a})$ for every a, then $\underset{\sim}{X}$ must be $N_p(\underset{\sim}{\mu}, \Sigma)$
3. For q linear combinations $\underset{q\times p}{A}\underset{\sim}{X}_{p\times 1}$, if $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$,

$$\underline{A}\underset{\sim}{X} \sim N_p(\underline{A}\underset{\sim}{\mu}, \underline{A}\Sigma\underline{A}^T)$$

4. $\underset{\sim}{X}_{p\times 1} + \underset{\sim}{d}_{p\times 1} \sim N_p(\underset{\sim}{\mu}+\underset{\sim}{d}, \Sigma)$ where $\underset{\sim}{d}$ is is a vector of constants
5. All subsets of $\underset{\sim}{X}$ are multivariate normally distributed

Eg if we partition $\underset{\sim}{X}$, we know that its mean vector and covariance matrix will be

$$\underset{\sim}{X}_{p\times 1} = \begin{bmatrix}X^{(1)}\\ \hline X^{(2)}\end{bmatrix}\begin{matrix}q\\ p-q\end{matrix}, \quad \underset{\sim}{\mu} = \begin{bmatrix}\underset{\sim}{\mu}^{(1)}\\ \hline \underset{\sim}{\mu}^{(2)}\end{bmatrix}\begin{matrix}q\\ p-q\end{matrix}, \quad \Sigma = \begin{matrix}q\\ p-q\end{matrix}\begin{pmatrix}\underset{q\times q}{\Sigma_{11}} & \underset{q\times(p-q)}{\Sigma_{12}}\\ \underset{(p-q)\times q}{\Sigma_{21}} & \underset{(p-q)\times(p-q)}{\Sigma_{22}}\end{pmatrix}$$

then $X^{(1)} \sim N_q(\underset{\sim}{\mu}^{(1)}, \Sigma_{11})$

Independence:

⑥ If $\underset{(q_1\times 1)}{X_1}$ and $\underset{q_2\times 1}{X_2}$ are independent then $Cov(X_1, X_2) = 0_{q_1\times q_2}$

⑦ If $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \}q_1 \\ \}q_2 \end{matrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$ then X_1 and X_2 are independent iff $\Sigma_{12} = \Sigma_{21}^T = 0_{q_1\times q_2}$

⑧ If X_1 and X_2 are independent and distributed as $N_{q_1}(\mu_1, \Sigma_{11})$ and $N_{q_2}(\mu_2, \Sigma_{22})$ respectively, then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{q_1+q_2}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right]$$

⑨ If $X \sim N_p(\mu, \Sigma)$ with $|\Sigma| > 0$ and if Σ^{-1} exists then

$$(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_{(p)}$$

Proof

Recall: If $Z_1, Z_2, \ldots, Z_p$ are independent $N(0,1)$, then $\sum_{i=1}^{p} Z_i^2 \sim \chi^2_{(p)}$

So in general, if $Z \sim N_p(0, I)$ then $Z^T Z \sim \chi^2_{(p)}$

$X-\mu \sim N_p(0, \Sigma)$ and $Z = \Sigma^{-1/2}(X-\mu) \sim N(0, I)$

Since $Z^T Z = (\Sigma^{-1/2}(X-\mu))^T(\Sigma^{-1/2}(X-\mu))$

$= (X-\mu)^T\Sigma^{-1/2}\Sigma^{-1/2}(X-\mu)$

$= (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_{(p)}$

Sampling Distribution of $\bar{X}$ and S

① $X \sim N(\mu, \sigma^2)$, then $\bar{x} \sim N(\mu, \sigma^2/n)$

② $(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-1)} \Rightarrow (n-1)s^2 \sim \sigma^2\chi^2_{(n-1)}$ $\quad s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

③ If $X \sim N_p(\mu, \Sigma)$, then a) $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$

b) $(n-1)S \sim W_{(n-1)}$ → Wishart distribution

c) $\underset{\sim}{\bar{X}}$ and $\underset{\sim}{S}$ are independent [HW: Prove for the univariate case]

2 important theorems:

① Law of large numbers:

$Y_1, \ldots, Y_n$ are independent observation from a population with $E(Y_i) = \mu$. Then $\bar{Y} \xrightarrow{P} \mu$, that is $P(|\bar{Y} - \mu| < \varepsilon) = 0$ $\forall \varepsilon > 0$.

As a consequence of LLN we can also say

$$S^2 \xrightarrow{P} \sigma^2 \quad \text{or} \quad \frac{1}{n}\sum (Y_i - \bar{Y})^2 \xrightarrow{P} \sigma^2$$

for the multivariate case

a) Each $\bar{X}_i \xrightarrow{P} \mu_i$; $i = 1, \ldots, p$ so that

$$\underset{\sim}{\bar{X}} \xrightarrow{P} \underset{\sim}{\mu}$$

b) Each sample covariance $s_{ik} \xrightarrow{P} \sigma_{ik}$; $k = 1, 2, \ldots, p$ so that

$$\underline{S} \text{ (or } S_n) \xrightarrow{P} \Sigma$$

② Central Limit Theorem

Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ be independent observation from a population with mean $\underset{\sim}{\mu}$ and finite (non singular) covariance Σ, then $\sqrt{n}\,(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}) \sim N_p(\underset{\sim}{0}, \Sigma)$ when n is large.

It follows from here that

$$n(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu})^T \Sigma^{-1}(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu}) \sim \chi^2_{(p)}$$

Properties of Wishart Distribution

① If $\underset{\sim}{Z}_1, \underset{\sim}{Z}_2, \ldots, \underset{\sim}{Z}_m$ are iid $N_p(\underset{\sim}{0}, \Sigma)$

then $\sum_{j=1}^{m} \underset{\sim}{Z}_j \underset{\sim}{Z}_j^T \sim W_m(\Sigma)$

② If $A_1 \sim W_{m_1}(\Sigma)$ independent of $A_2 \sim W_{m_2}(\Sigma)$

then

$$A_1 + A_2 \sim W_{m_1+m_2}(\Sigma)$$

③ If $A \sim W_m(\underset{p\times p}{\Sigma})$, then $CAC^T \sim W_m(C\Sigma C^T)$

$k\times p \quad p\times p \quad p\times k \quad k\times k$

The Wishart distribution is a family of distributions for symmetric positive definite matrices.

Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ be independent $N_p(\underset{\sim}{0}, \Sigma)$, we can form a $n\times p$ data matrix $\underline{X} = (\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n)$. Then the distribution of the $p\times p$ $M = \underline{X}^T\underline{X} = \sum_{i=1}^{n} \underset{\sim}{X}_i^T \underset{\sim}{X}_i$ is said to follow the Wishart distribution.

Density function

For a random matrix M which follows a Wishart distribution with n degrees of freedom and covariance matrix Σ, $\underset{p\times p}{M} \sim W_n(\Sigma)$, if $n \geq p$

$W_p(n, \Sigma)$ ←

$$f(M) = \frac{1}{2^{np/2}\,\Gamma_p(n/2)\,|\Sigma|^{n/2}}\,|M|^{(n-p-1)/2}\exp\left[-\frac{1}{2}\operatorname{trace}(\Sigma^{-1}M)\right]$$

multivariate gama

The Wishart distribution is a multivariate extension of the χ^2 distribution so that

$$M \sim W_1(n, \sigma^2) \simeq \sigma^2 \chi^2_p(n)$$

Assessing the assumption of normality.

- Several of the multivariate statistical techniques to be discussed assume that each vector of observations

$$X_j \sim N_p(\mu, \Sigma)$$

- This assumption is not so important in cases of large n when we are making inferences related to $\bar{x}$.
- We would like to check the normality assumption for p-dimensions, but in practice, it suffices to investigate univariate and bivariate normality.
- Given a particular data set, we may be interested in addressing the following questions:

(1) Do the marginal distributions of the elements of X appear to be normal? What about a few linear combinations of the components X_i?

(2) Do the scatter plots of pairs of observations on different characteristics give elliptical appearances as expected from normal populations?

③ Are there any extreme observations that you should check for their accuracy.

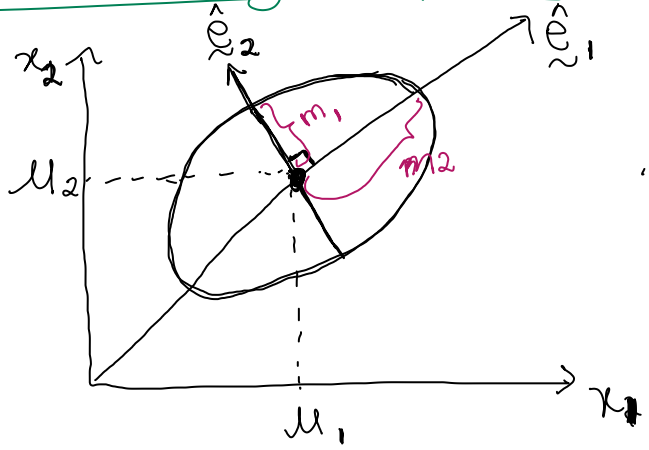
Evaluating normality in the univariate marginal distribution

* Check R file on Multivariate Normal.

Evaluating Bivariate Normality

Recall: If the observations were generated from a MV N distribution, each bivariate distribution would be normal and the contours of constant density would be ellipses.

> Understanding ellipses)



- $\hat{e}_1 \pm \hat{e}_2$; $\hat{e}_1$ and $\hat{e}_2$ are the the normalized eigen vectors

$$m_1 = \pm c\sqrt{\lambda_1}\,\hat{e}_1$$

$$m_2 = \pm c\sqrt{\lambda_2}\,\hat{e}_2$$

λ_1, λ_2 are eigen values

We can show that

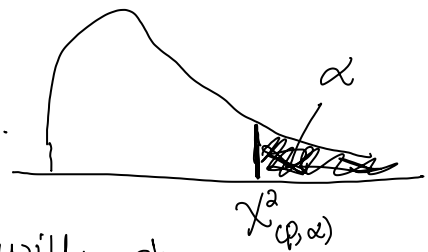
$$c^2 = (\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (x - \mu)$$

or

$$c^2 = (\underset{\sim}{x} - \bar{\underset{\sim}{x}})^T S^{-1} (\underset{\sim}{x} - \bar{\underset{\sim}{x}})$$

$$= \chi^2(p, \alpha)$$

ellipsoid centered at $\underset{\sim}{\mu}$



a) Center = mean vector. Change in mean will shift the contours without altering the shape.

b) The orientation of the ellipses are determined by Σ. If the ellipse is vertically or horizontally oriented, it suggest

little to no covariance (or correlation) between X_1, X_2.

If inclined at an angle, there is covariance / correlation between X_1, X_2

c) The size of the ellipse in terms of its major and minor axes provide information about the variance of the individual variables. A larger ellipse $\Rightarrow$ higher variance in both variables

Smaller ellipse $\Rightarrow$ lower variance - - - -

d) Aspect ratio (ratio of major axis to the minor axis length) gives an idea of how the variances compare between X_1, X_2.

$$\text{Aspect ratio} = \begin{cases} \text{Close to 1} & V(X_1) \text{ almost equal to } V(X_2) \\ \text{Large dev. from 1} & V(X_1) > V(X_2) \end{cases}$$

e) Change in Σ changes the shape.

f) if $\text{Cov}(X_1, X_2) = 0$, ellipse becomes circles indicates X_1 nd X_2 are independent

~~Matabranobis~~

---

① Given bivariate outcomes $\underset{\sim}{x}$, ~~such that~~ we expect at least 0.5 or 50% of he observations to lie within the ellipse given by

$$\{\underset{\sim}{x} : (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T S^{-1} (\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \leq \chi^2_2(0.5)\}$$

~~See~~ example document and R code for illustration.

② Chi square plot (Bivariate & Multivariate)

a) Order squared distances $d_j^2 = (\underset{\sim}{x} - \bar{x})^T S^{-1} (\underset{\sim}{x} - \bar{x})$; $j = 1, \cdots, n$ from the smallest ; that $d^2_{(1)} \leq d^2_{(2)} \leq \cdots d^2_{(n)}$

b) Graph the pairs $\left(q_{c,p}((j-1/2)/n), d^2_{(j)}\right)$

We expect the plot to resemble a straight line through the origin

Any systematic curve suggest a lack normality

Detecting outliers in your data..

- Outliers are best detected visually when possible.
- Univariate : dot plot, boxplot, $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$, qq plot etc.
- Bivariate : Scatterplot, chisquare plot
- Multivariate :
  a) Univariate plots as mentioned above
  b) Scatterplot of pairs of variables
  c) Obtain standardized values $z_{jk} = \frac{(x_{jk} - \bar{x}_k)}{\sqrt{s_{kk}}}$ for $j = 1, 2, \ldots, n$.

     and each column $k = 1, 2, \ldots, p$

     If $|z_{jk}| > 3.5$ observation is an outlier

  d) Obtain Generalized squared distance $d_j^2 = (\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})^T S^{-1} (\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})$ and observe the chi square plot. or compare each value to $\chi^2_p(0.005)$

     if $d_j^2 > \chi^2_p(0.005)$ observation is an outlier.

Transformations to near normality

- If normality assumption is violated we may need to transform our data and consider performing our analysis based on the transformed data (after it follows normality of course!)

Helpful transformations to near normality

Original Scale	Transformed Scale
① Count, y	$\sqrt{y}$
② Propoportions, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
③ Correlations, r	Fisher's z-transformation $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$

- In many instances, the choice of transformation to improve the approximation to normality is not obvious and so we could make use of a family of transformations known as Power transformations

For univariate data which deviate from the normality assumption, our goal is to find a value λ such that x^λ would be near normal. Some commonly used choices of λ are:

λ	transformation	
-1	$x^{-1} = 1/x$	} shrink large values of x
0	$\ln x$	
1	x	
$1/4$	$x^{1/4}$	

$\frac{1}{2}$	$x^{1/2}$
2	x^2
3	x^3

} increases large values of x.

We can use the Box Cox technique to analytically choose λ. In this technique, λ is chosen such that

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \quad ; x > 0$$

Given the observations $x_1, x_2, \ldots, x_n$, the Box-Cox solution for the choice of an appropriate power λ is the λ value that maximizes the expression

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j$$

where $\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_j^\lambda - 1}{\lambda} \right)$

In the multivariate case, let $\underline{\lambda} = (\lambda_1 \ \lambda_2 \ \ldots \ \lambda_p)$ be the power transformations for the p variables. Each λ_k; $k = 1, \ldots, p$ is obtained by maximizing

$$\ell_k(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_{jk}^{(\lambda_k)} - \overline{x_k^{(\lambda_k)}})^2 \right] + (\lambda_k - 1) \sum_{j=1}^{n} \ln x_{jk}$$

Check book examples 4.16 & 4.17 in the examples document and the R code.

Note: After transforming, one needs to check the transformed data to be sure that normality assumption is satisfied

Chapter 5: Inferences about the mean vector

Statistical inference involves reaching valid conclusions concerning a population based on information from a sample. This may involve hypothesis testing or confidence intervals.

Recall: In univariate data analysis; suppose $x_1, \dots, x_n$ is a random sample from a univariate $N(\mu, \sigma^2)$. Also suppose we want to test a hypothesis concerning the population mean such that

$H_0: \mu = \mu_0$ and $H_a: \mu \neq \mu_0$ (2-sided)

Test statistic:

Case 1: $$t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} \sim t(n-1) \text{ under } H_0$$

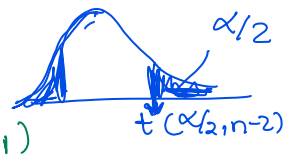
This test statistic has an equivalent alternative given by

Case 2: $$t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n} \sim F(1, n-1) \text{ under } H_0$$

Decision: Reject H_0 in favor of H_a if at a level of significance

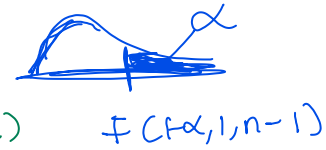
Case 1: Reject H_0 if $|t| > t(\alpha/2, n-1)$

in R = qt(1-α/2, n-1)



Case 2: Reject H_0 if $t^2 > F(1-\alpha, 1, n-1)$

in R = qf(1-α, 1, n-1)



Conclusion: There is enough evidence to conclude "H_a"

There is not enough evidence to conclude "H_a"

A $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - t(\alpha/2, n-1)\, s/\sqrt{n} \leq \mu \leq \bar{x} + t(\alpha/2, n-1)\, s/\sqrt{n}$$

Note that the confidence interval consists of all the values of μ_0 that will not be rejected by a level α test of $H_0: \mu = \mu_0$, that is if $\mu_0 \in$ Confidence Interval, fail to reject H_0.

Check R for t-test example for $\mu_{\text{sweat rate}} = 0$

Consider now the problem of determining whether a $p \times 1$ vector $\underset{\sim}{\mu}_0$ is plausible for the mean of the multivariate normal distribution.

Why consider them jointly and not separately?

a) overall Type I error becomes inflated when you test the variables individually. Type I error is the probability of falsely rejecting H_0. ie $P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$. For example, consider a dataset with 2 variables. Suppose we want an overall $\alpha = 0.05$. Performing each individual test at $\alpha = 0.5$ yields an overall $\alpha = 1 - (0.95)(0.95) = 0.1$. Of course, we can divide α by the number of tests (variables) and perform individual tests using α/p to yield an overall α (Dunn-Bonferroni adjustment). This may solve the problem of inflating α but then there is the 2nd reason:

b) The individual univariate t-tests ignores correlation and covariance among dependent variables.

c) It is possible that one may not find a single univariate mean difference between groups but a mean difference may exist when you consider the set of dependent variables jointly.

Note:

Situations may arise where all p variable means may not be jointly significant but some subsets could be, so check for subsets of dependent variables as well.

We may extend the univariate t-test to the multivariate case:

Given a random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ from an $N_p(\underset{\sim}{\mu}, \Sigma)$ population, where $\underset{\sim}{X}_j = (X_{j1}\ X_{j2}\ \ldots\ X_{jp})$ and both $\underset{\sim}{\mu}$ and Σ are unknown. Let

$$\underset{\sim}{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \underset{\sim}{X}_j \quad \text{and} \quad \underset{\sim}{S} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})^T$$

be the sample mean vector and sample covariance matrix respectively. Suppose we wish to test:

$$H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \quad \text{vs} \quad H_a: \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0 \quad \text{(two-sided)}$$

where $\underset{\sim}{\mu}_0$ is known.

Test statistic which is known as Hotelling's T^2 test statistic is given by

$$T^2 = n(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0)^T S^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0)$$

Under H_0,

$$T^{*2} = \frac{n-p}{p(n-1)} T^2 \sim F(p, n-p)$$

Decision: Reject H_0 if $T^{*2} > F(1-\alpha, p, n-p)$ 

Note: Hotelling's T^2 statistic reduces to the univariate test statistic t^2 when $p=1$

Refer to chapter 5 examples document corresponding to Book

examples 5.1 and 5.2. Check R code too

A confidence region for the mean vector $\underset{\sim}{\mu}$ is also an extension of the CI for the μ in the univariate case.

Recall the relationship between confidence interval/regions and hypothesis tests: A $100(1-\alpha)\%$ confidence interval/region can be derived when we invert the acceptance region of a level α test

So by inverting the acceptance region of the Hotelling's T^2 test, we can obtain the confidence region for the mean vector $\underset{\sim}{\mu}$.

The $100(1-\alpha)\%$ confidence region for the mean of p-dimensional normal distribution is the ellipsoid determined by all $\underset{\sim}{\mu}$ such that

$$n(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu})^T S^{-1}(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu}) \leq \frac{(n-1)p}{n-p} F(\alpha, p, n-p)$$

* If the squared distance $> \frac{(n-1)p}{n-p} F(\alpha, p, n-p)$, $\underset{\sim}{\mu}_0$ is not in the confidence region and this leads to rejecting H_0 for a hypothesis test for $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ vs $H_a: \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$.

Hypothesis test on q Linear combinations

More generally, suppose $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$ is a random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$. To test the hypothesis:

$$H_0: \underset{q\times p}{\bar{C}} \underset{p\times 1}{\underset{\sim}{\mu}} = \underset{q\times 1}{\underset{\sim}{\mu}_0} \quad \text{vs} \quad H_a: \bar{C}\underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$$

Let $\underset{\sim}{Y}_i = \bar{C}\underset{\sim}{X}_i - \underset{\sim}{\mu}_0$

Then $\underset{\sim}{Y}_i \sim N_p(\underset{\sim}{\mu}_Y, \Sigma_Y)$

where $\underset{\sim}{\mu}_Y = \bar{C}\underset{\sim}{\mu} - \underset{\sim}{\mu}_0$ and $S_Y = \bar{C} S_X C^T$

Test statistic:

$$T^{*2} = n(\bar{C}\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)^T (\bar{C} S_X \bar{C}^T)^{-1} (\bar{C}\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)$$

$$= n \underset{\sim}{\bar{Y}}^T (S_Y)^{-1} (\underset{\sim}{\bar{Y}})$$

Decision: Reject H_0 if $T^{*2} > \frac{(n-1)q}{n-q} F(1-\alpha, q, n-q)$

=

Back to joint confidence region for μ

- $p=2$: confidence ellipsoid
- $p=3$: something spherical (like a football)
- $p \geq 4$: hard to graph

Beginning from the center $\underset{\sim}{\bar{x}}$, the axes of the confidence ellipsoids (within each confidence region, there are ellipsoids are given by

$$\pm \sqrt{\lambda_i} \sqrt{\frac{p(n-1)}{n(n-p)} F(\alpha, p, n-p)} \; \underset{\sim}{e_i} \quad ; \; S e_i = \lambda e_i \quad i=1,2,\dots,p$$

length of the axes

$\underset{\sim}{e_i}$ - eigen vector corresponding to eigen value λ_i derived from the covariance matrix.

Example 5.3 from example document

Simultaneous Confidence Intervals

Suppose $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ and $\underset{\sim}{a} \in \mathbb{R}^p$ (a vector), we can form the linear combination

$$Z = a_1 X_1 + a_2 X_2 + \dots + a_p X_p = \underset{\sim}{a}^T \underset{\sim}{X}$$

We know $\mu_Z = E(Z) = \underset{\sim}{a}^T \underset{\sim}{\mu}$ and $\sigma_Z^2 = Var(Z) = \underset{\sim}{a}^T \Sigma \underset{\sim}{a}$

So $\underset{\sim}{Z} \sim N(\underset{\sim}{a}^T \underset{\sim}{\mu}, \underset{\sim}{a}^T \Sigma \underset{\sim}{a})$

For a random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ from $N_p(\underset{\sim}{\mu}, \Sigma)$, Consider a linear combination of the sample from the jth trial:

$$Z_j = a_1 X_{j1} + a_2 X_{j2} + \ldots + a_p X_{jp} = \underset{\sim}{a}' \underset{\sim}{X}_j \quad j = 1, 2, \ldots, n$$

Then for n trials, we have $Z_1, Z_2, \ldots, Z_n$ with

$$\bar{Z} = \underset{\sim}{a}^T \bar{\underset{\sim}{X}} \quad \text{and} \quad S_z^2 = \underset{\sim}{a}^T S_{\underset{\sim}{x}} \underset{\sim}{a} \longrightarrow \Sigma \text{ unknown.}$$

Then for a fixed $\underset{\sim}{a}$ and σ_z^2 unknown, the $100(1-\alpha)\%$ CI for $\mu_z = \underset{\sim}{a}' \underset{\sim}{\mu}$ will be given by

$$\bar{Z} \pm t(\alpha/2, n-1) \sqrt{\frac{\underset{\sim}{a}^T S \underset{\sim}{a}}{n}} \Big\} = \frac{1}{n} S_z$$

Fact:

Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ be a random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$ with Σ being positive definite. The $100(1-\alpha)\%$ Simultaneous confidence interval for all $\underset{\sim}{a}$ is given by

$$\underset{\sim}{a}' \bar{\underset{\sim}{X}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F(\alpha, p, n-p)\, \underset{\sim}{a}^T S \underset{\sim}{a}}$$ and this

interval will contain $a^T \mu$ with probability $1-\alpha$.

Conveniently, if we choose $\underset{\sim}{a}^T = (0\ 0\ 0 \ldots a_i\ 0\ 0\ 0)$ where $a_i = 1$, then the T^2 intervals allow us to conclude that

$$\bar{X}_i \pm \sqrt{\frac{p(n-1)}{n(n-p)} F(\alpha, p, n-p)\, S_{ii}} \quad ; \quad \underset{\sim}{a}^T \bar{\underset{\sim}{X}} = \bar{X}_i$$

- With the same confidence level $1-\alpha$, we can also make statements about mean differences, say $\mu_i - \mu_k$

In this case, choose $\underset{\sim}{a}' = (0\ 0 \dots a_i\ 0\ 0 \dots 0\ a_k\ 0\ 0 \dots 0)$

where $a_i = 1$ and $a_k = -1$.

Point estimator for $\mu_i - \mu_k = \bar{x}_i - \bar{x}_k$

$$\begin{aligned} \text{Var}(\bar{x}_i - \bar{x}_k) &= V(\bar{x}_i) + V(\bar{x}_k) - 2\,\text{Cov}(\bar{x}_i, \bar{x}_k) \\ &= \frac{1}{n}\left(V(x_i) + V(x_k) - 2\,\text{Cov}(x_i, x_k)\right) \\ &= \frac{1}{n}\left(s_{ii} + s_{kk} - 2s_{ik}\right) \end{aligned}$$

Hence a $100(1-\alpha)\%$ CI for $\mu_i - \mu_k$ will be given as

$$\bar{x}_i - \bar{x}_k \pm \sqrt{\frac{p(n-1)}{n(n-p)} F(\alpha, p, n-p)\,(s_{ii} + s_{kk} - 2s_{ik})}$$

Check examples 5.4 and 5.5 in examples document.

- Simultaneous confidence intervals are useful when one is interested in making specific inferences about each parameter separately, allowing for straightforward interpretation
  - Easier to visualize

- Joint confidence regions are useful when interested in the relationships among multiple parameters. They allow one to collectively assess the parameters and incorporate how changes

in one might be associated with changes in others.
- Complex or unfeasible to visualize, in higher dimensions.

More Precise Simultaneous Confidence Intervals

If p and or the number of linear combinations $\underset{\sim}{a}^T\underset{\sim}{\mu}$ is small, we may be able to form shorter simultaneous confidence intervals. One procedure is the Bonferroni correction method in which to control the familywise error rate, we need to adjust the error rates of the individual test (or confidence intervals).

- We set $\alpha_K = \alpha/q$ to control the familywise error rate at α when conducting q tests.
- The $100(1-\alpha)\%$ familywise confidence interval for q linear combinations, $\underset{\sim}{a}^T\underset{\sim}{\mu}$ is given as

$$\underset{\sim}{a}^T\underset{\sim}{\bar{x}} \pm t\left(\frac{\alpha_K}{2}, n-1\right)\sqrt{\frac{1}{n}\underset{\sim}{a}^T S \underset{\sim}{a}} \quad ; \alpha_K = \frac{\alpha}{q} \quad (\text{or } \frac{\alpha}{2q})$$

This is equivalent to

$$\mu_1 \in \bar{x}_1 \pm t\left(\frac{\alpha}{2q}, n-1\right)\sqrt{s_{11}/n}$$

$$\vdots$$

$$\mu_q \in \bar{x}_q \pm t\left(\alpha/2q, n-1\right)\sqrt{s_{qq}/n}$$

Note the Bonferroni procedure uses $t(\alpha/2q, n-1)$ quantiles

instead of $\sqrt{\frac{(n-1)p}{n-p} F(\alpha, p, n-p)}$ as in the previously discussed simultaneous CI.

Check Example 5.6 in examples document.

Fact:

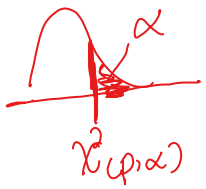
Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ be a random sample from a population with mean $\underset{\sim}{\mu}$ and positive definite covariance matrix Σ. When n-p is large (regardless of whether data is normal or not),

$$\sqrt{n}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \approx N(\underset{\sim}{0}, \Sigma) \quad \text{and}$$

$$n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu})^T S^{-1} (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}) \approx \chi^2_{(p)}$$

Hence in testing $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ against $H_1: \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$ at a level of significance of α, we reject H_0 if

$$n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu})^T S^{-1} (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}) > \chi^2_{(p,\alpha)}$$



Note: Using this $\chi^2_{(p,\alpha)}$ or $\frac{(n-1)p}{(n-p)} F(\alpha, p, n-p)$ should yield same conclusions since these 2 are approximately equal when n is large relative to p.

The $100(1-\alpha)\%$ confidence region when n-p is large is given by all μ such that:

$$n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu})^T S^{-1} (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}) \leq \chi^2_{p(\alpha)}$$

Fact:

$\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n \sim N_p(\underset{\sim}{\mu}, \Sigma)$; Σ - positive definite. for $n-p$ large,

$$\underset{\sim}{a}^T \bar{\underset{\sim}{X}} \pm \sqrt{\chi^2(p,\alpha) \frac{\underset{\sim}{a}^T S \underset{\sim}{a}}{n}}$$ will contain $\underset{\sim}{a}^T \underset{\sim}{\mu}$ for every $\underset{\sim}{a}$.

Consequently, the $100(1-\alpha)\%$ simultaneous confidence intervals is given by

$$\mu_1 \in \bar{x}_1 \pm \sqrt{\chi^2(p,\alpha)\, s_{11}/n}$$

$$\vdots$$

$$\mu_p \in \bar{x}_p \pm \sqrt{\chi^2(p,\alpha)\, s_{pp}/n}$$

The Bonferroni simultaneous confidence intervals for q intervals is given by:

$$\underset{\sim}{a}^T \bar{\underset{\sim}{X}} \pm z\left(\frac{\alpha}{2q}\right) \sqrt{\frac{1}{n} \underset{\sim}{a}^T S \underset{\sim}{a}}$$

So for example for individual means, $\bar{x}_i$, we have

$$\bar{x}_i \pm z(\alpha/2p) \sqrt{s_{ii}/n} \quad ; i = 1, 2, \ldots, p$$

* Note: one at a time CI do not use the familywise adjustments so for Bonferroni, that is

$\bar{x}_k \pm t(\alpha/2, n-1) \sqrt{\frac{s_{kk}}{n}}$ and $\bar{x}_k \pm z(\alpha/2) \sqrt{\frac{s_{kk}}{n}}$, respectively

Large sample

Comparison - Univariate and Multivariate tests

Univariate

① σ^2 is known

② σ^2 unknown

- (a) n large
- (b) n small

Multivariate

① Σ is known

② Σ is unknown

- (a) n-p large
- (b) n-p small

Chapter 6
Comparing Mean Vectors from Two Populations

Analysts may have data on samples from different populations and may be interested in answering questions as to whether the populations under consideration have different means (or mean vectors in the multivariate case)

Two independent mean differences

(1) When the two populations have equal variances

(A) Univariate case: (small n_1, n_2)

$X_{11}, X_{12}, X_{13}, \dots, X_{1n}$ is a random sample of size n_1 from $N(\mu_1, \sigma_1^2)$

$X_{21}, X_{22}, X_{23}, \dots, X_{2n}$ is a random sample of size n_2 from $N(\mu_2, \sigma_2^2)$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (Perform Levene's $\overset{F\text{ test}}{\wedge}$ to check this)

Test statistic for the test $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_a: \mu_1 - \mu_2 \neq \delta_0$ is given as

$$t^* = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{s^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t\left(\alpha/2, n_1 + n_2 - 2\right)$$

where $s^2_{pooled} = \dfrac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}$

with

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad \text{and} \quad S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

The $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is given by

$$\bar{x}_1-\bar{x}_2 \pm t(\alpha/2, n_1+n_2-2)\sqrt{s^2_{pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

(B) Multivariate case (small n_1, n_2): Assume $\underset{\sim}{x}_{11}, \underset{\sim}{x}_{12}, \ldots, \underset{\sim}{x}_{1n} \sim N_p(\mu_1, \Sigma_1)$
$\underset{\sim}{x}_{21}, \underset{\sim}{x}_{22}, \ldots, \underset{\sim}{x}_{2n} \sim N_p(\mu_2, \Sigma_2)$

To test equality of variance covariance matrices use the Box M test which test $H_0: \Sigma_1 = \Sigma_2$

Will be demonstrated in R.

If the test concludes that $\Sigma_1 = \Sigma_2 = \Sigma$, we have

$\sum_{j=1}^{n_1}(\underset{\sim}{x}_{1j}-\underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}_{1j}-\underset{\sim}{\bar{x}}_1)^T$ as an estimate of $(n_1-1)\Sigma$ and

$\sum_{j=1}^{n_2}(\underset{\sim}{x}_{2j}-\underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}_{2j}-\underset{\sim}{\bar{x}}_2)^T$ as an estimate of $(n_2-1)\Sigma$

so that the pooled covariance matrix is calculated as

$$S_p = S_{pooled} = \frac{\sum_{j=1}^{n_1} (\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)^T + \sum_{j=1}^{n_2} (\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)^T}{n_1 + n_2 - 2}$$

$$= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

where $S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)^T$ and

$S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)^T$

To test

$H_0: \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{\delta}_0$ vs $H_a: \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \neq \underset{\sim}{\delta}_0$

Test statistic:

$$T^2 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))$$

$$(T^*)^2 = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2 \sim F(\alpha, p, n_1 + n_2 - p - 1)$$

Reject H_0 if $(T^*)^2 > F(\alpha, p, n_1 + n_2 - p - 1)$

The $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ is given by all $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ satisfying:

$$(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} ((\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \leq$$

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F(\alpha, p, n_1 + n_2 - p - 1)$$

Note that this region is an ellipsoid centered at the observed $\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2$ and whose axes are determined by the eigen values and eigen vectors of S_{pooled}.

Refer to example 6.3 (Problem 4 in example document)

© Simultaneous confidence intervals for $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$

We construct these confidence intervals by considering all possible linear combinations of the differences in mean vectors.

The $100(1-\alpha)\%$ simultaneous confidence interval for $\underset{\sim}{a}^T(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ for all a is given as:

$$\underset{\sim}{a}^T(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2) \pm c\sqrt{\underset{\sim}{a}^T\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{pooled}\,\underset{\sim}{a}}$$

where $c^2 = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F(\alpha, p, n_1+n_2-p-1)$

$1-\alpha$ simultaneous

In particular, we have that the $100(1-\alpha)\%$ confidence interval for $\mu_{1i} - \mu_{2i}$ is given by:

$$\bar{X}_{1i} - \bar{X}_{2i} \pm c\sqrt{S_{ii,pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{for } i=1,2,\ldots,p$$

The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for the p population mean differences for $\mu_{1i} - \mu_{2i}$ is given as:

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm t\left(\frac{\alpha}{2p}, n_1+n_2-2\right)\sqrt{S_{ii,pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

② Independent samples with $\Sigma_1 \neq \Sigma_2$

Ⓐ Univariate case

X_{1i} $i=1,2,\dots,n$ $\sim N(\mu_1, \sigma_{11})$

X_{2i} $i=1,2,\dots,n$ $\sim N(\mu_2, \sigma_{22})$

$\sigma_{11} \neq \sigma_{22}$ (from Levene's test)

The test of $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_a: \mu_1 - \mu_2 \neq \delta_0$ has test statistic

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\alpha/2, \nu)$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$ (Round this down ALWAYS)

(B) Multivariate case: Independent samples, $\Sigma_1 \neq \Sigma_2$

Result: (When n_1 and n_2 are large)

Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are large. Then an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\mu_1 - \mu_2$ is given by all $\mu_1 - \mu_2$ satisfying

$$[\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)] \leq \chi^2(\alpha, p)$$

Test statistic. Reject H_0 if T.S $> \chi^2(\alpha, p)$

- $100(1-\alpha)\%$ simultaneous confidence intervals for all linear combinations $a^T(\mu_1 - \mu_2)$ is given by

$$a^T(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi^2(\alpha, p)} \sqrt{a^T\left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right) a}$$

Book example 6.5 (Problem 5 in example document)

When sample sizes are not large

2 populations need to be MVN

n_1 and n_2 each have to be greater than p

$H_0: \mu_1 - \mu_2 = \delta_0$ $\quad$ $H_a: \mu_1 - \mu_2 \neq \delta_0$

Test statistic

$$T^2 = (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))$$

$$(T^*)^2 = \frac{\nu - p + 1}{\nu p} T^2$$

Reject H_0 if $(T^*)^2 > F(\alpha, p, \nu - p + 1)$

ν can be calculated as

$$\frac{p + p^2}{\sum_{i=1}^{2} \frac{1}{n_i} \left\{ tr\left[\left(\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right)^2 \right] + \left(tr\left[\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right] \right)^2 \right\}}$$

where $\min(n_1, n_2) \leq \nu \leq n_1 + n_2$

The approximate $100(1-\alpha)\%$ confidence region is given by all $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ such that

$$(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \leq$$

$$\frac{\nu p}{\nu - p + 1} F(\alpha, p, \nu - p + 1)$$

where ν is as given before.

Book example 6.6 (Problem 6 in class example)

Dependent Populations

A) Two dependent populations:

- Previously, we conducted hypothesis testing about the difference between two population means when the two samples are drawn independently from two different populations
- Here we deal with inferences for the mean of dependent populations which are also called paired populations.
- Use this test when observations from the two populations of interest are collected in pairs.

Check examples document for an example on when you have paired data.

Let X_{j1} denote the response to treatment 1 (or the response before treatment), and let X_{j2} denote the response to treatment 2 (or response after treatment) for the jth trial.

Note: (X_{j1}, X_{j2}) are measurements recorded on the jth unit or jth pair of like units.

We analyze the n differences:

$$D_j = X_{j1} - X_{j2} \quad , j = 1, 2, \cdots, n$$

The differences D_j represent independent observations from $N(\delta, \sigma_d^2)$ where δ is the population mean difference between the 2 treatments, ie $E(X_{j1} - X_{j2})$

To test:

$H_0: \delta = 0$ (zero mean difference for treatments

$H_1: \delta \neq 0$

Test statistic: $t = \dfrac{\bar{D} - \delta}{S_d/\sqrt{n}} \sim t(\alpha/2, n-1)$

where $\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$ and $S_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})^2$

Reject H_0 if $|t| > t(\alpha/2, n-1)$

A $100(1-\alpha)\%$ confidence interval for the mean differences $\delta = E(X_{j1} - X_{j2})$ is given by:

$$\bar{D} \pm t(\alpha/2, n-1)\, S_d/\sqrt{n}$$

Check example (Problem 3) in example document

B) Multivariate Case

- p variables, n experimental units.
- The jth experimental unit has p observations given as

X_{1j1} = variable 1 under treatment 1
X_{1j2} = " 2 " "
:
X_{1jp} = " p under treatment 1

} Population 1

X_{2j1} = Variable 1 under treatment 2
X_{2j2} = " 2 " 2
⋮
X_{2jp} = " p under treatment 2

} Population 2

- The paired difference random variables becomes:

$$D_{j1} = X_{1j1} - X_{2j1}$$
$$D_{j2} = X_{1j2} - X_{2j2}$$
$$\vdots$$
$$D_{jp} = X_{1jp} - X_{2jp}$$

So for the jth experimental unit, we have

$$\underset{\sim}{D}_j^T = (D_{j1}\ D_{j2}\ \cdots\ D_{jp}),$$

$$E(\underset{\sim}{D}_j) = \underset{\sim}{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{pmatrix} \quad \text{and} \quad Cov(\underset{\sim}{D}_j) = \Sigma_d \quad j = 1, 2, \ldots, n$$

$\underset{\sim}{D}_1, \underset{\sim}{D}_2, \ldots, \underset{\sim}{D}_n$ are independent $N_p(\underset{\sim}{\delta}, \Sigma_d)$ random vectors.
Given the observed differences $\underset{\sim}{d}_j^T = (d_{j1}\ d_{j2}\ \cdots\ d_{jp})$, $j = 1, 2, \ldots, n$ from $\underset{\sim}{D}_1, \ldots, \underset{\sim}{D}_n$, to test:

$H_0: \underset{\sim}{\delta} = \underset{\sim}{0}$ vs $H_a: \underset{\sim}{\delta} \neq \underset{\sim}{0}$

Test statistic: $T^2 = n(\bar{\underset{\sim}{d}} - \underset{\sim}{\delta})^T S_d^{-1} (\bar{\underset{\sim}{d}} - \underset{\sim}{\delta})$

where

If $T^2 > \frac{(n-1)p}{n-p} f(\alpha, p, n-p)$, reject H_0.

A $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\delta}$ consists of all $\underset{\sim}{\delta}$ such that

$$(\bar{\underset{\sim}{D}} - \underset{\sim}{\delta})^T S_d^{-1} (\bar{\underset{\sim}{D}} - \underset{\sim}{\delta}) \leq \frac{(n-1)p}{n(n-p)} F(\alpha, p, n-p)$$

Also, $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences δ_i are given by

$$\bar{d}_i \pm \sqrt{\frac{(n-1)p}{(n-p)} F(\alpha, p, n-p)} \sqrt{\frac{s_{d_i}^2}{n}} \quad ; i = 1, 2, \ldots, p$$

where $\bar{d}_i$ is the ith element of $\bar{d}$ and $s_{d_i}^2$ is the ith diagonal element of S_d

- The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for q individual mean differences are

$$\bar{d}_i \pm t(\alpha/2q, n-1) \sqrt{\frac{s_{d_i}^2}{n}}$$

Note: (1) If n and $n-p$ are both large, T^2 is approximately $\chi^2(\alpha, p)$ regardless of the form of the underlying population of the differences.

(2) for $n-p$ large, $\frac{(n-1)p}{n-p} F(\alpha, p, n-p) \cong \chi^2(\alpha, p)$ and normality need not be assumed.

Refer to example document (Problem 7) - Book example 6.1

Question: What if we have more than 2 populations?

Comparing several population means

Univariate ANOVA

Suppose we are interested in comparing 3 or more population means. As usual, we can draw independent samples from each population, calculate their sample means and use them in the analyses.

Eg. ① We want to compare average weight of chickens from 4 different poultry farms [One variable of interest: weight, 4 populations: Location 1-4]

② Checking differences in average maths scores for high schools in the Madison county.

—

Now, suppose we wish to compare means of g different populations, we can test

$H_0: \mu_1 = \mu_2 = \cdots = \mu_g$ against

H_a: At least one group mean is different from the others.

Assumption: $X_{i1}, X_{i2}, \ldots, X_{n_i}$ is a random sample from $N(\mu_i, \sigma^2)$ (Constant variance) $i = 1, 2, \ldots, g$ and the random samples from each population are independent.

- Each population mean, μ_i can be decomposed into the overall mean (grand mean) and a component which is due to the specific population, that is:

$$\mu_i = \mu + \tau_i$$

(μ_i: ith population mean; μ: Grand mean; τ_i: ith population treatment effect)

H_0 and H_a can be written in terms of the individual treatment effects as

H_0: $\tau_1 = \tau_2 = \dots = \tau_g = 0$ (There is no treatment effect)

H_a: At least one of the $\tau_i \neq 0$ (There is a possible treatment effect)

- Each individual observation X_{ij} can be decomposed into

$$X_{ij} = \mu_i + e_{ij} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_i}_{\text{treatment effect}} + \underbrace{e_{ij}}_{\text{random error}} \qquad (\mu_i - \mu)$$

Subject to $\sum_{i=1}^{g} n_i \tau_i = 0$ · $e_{ij} \sim N(0, \sigma^2)$ (independent)

- We estimate the model from the sample as follows:

$$\underbrace{X_{ij}}_{\text{Observation}} = \underbrace{\bar{X}}_{\substack{\text{overall sample mean} \\ \text{- estimate of } \mu}} + \underbrace{(\bar{X}_i - \bar{X})}_{\substack{\text{estimated treatment effect} \\ \text{- estimate of } \tau_i}} + \underbrace{(X_{ij} - \bar{X})}_{\substack{\text{residual} \\ \text{estimate of } e_{ij}}}$$

- Question: How much of the total variation in the data comes from variation within the group or variation between the group?

Example (Check example ___ in examples document)

	Observations				
Population 1:	9	6	9	$n_1 = 3$	$\bar{X}_1 = 8$
Population 2:	0	2		$n_2 = 2$	$\bar{X}_2 = 1$
Population 3:	3	1	2	$n_3 = 3$	$\bar{X}_3 = 2$

Overall mean: $\frac{9+6+9+0+2+3+1+2}{8} = 4 = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1+n_2+n_3}$

We wish to find the 3 Sum of squares

a) SST = Sum of squares total
= Sum of squared deviation of each observation from the overall mean
$= (9-4)^2 + (6-4)^2 + (9-4)^2 + (0-4)^2 + (2-4)^2 + (3-4)^2 + (1-4)^2 + (2-4)^2$

$= 88$

$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$

= SS_{cor} (Total corrected SS)

$df(SS_{cor}) = \left(\sum_{i=1}^{g} n_i\right) - 1$ — Why?

(b) SSW = "Sum of squares within the individual groups"
= Sum of squared deviation of each observation from its corresponding population mean.
$= (9-8)^2 + (6-8)^2 + (9-8)^2 + (0-2)^2 + (2-2)^2 + (3-2)^2 + (1-2)^2 + (2-2)^2$

$= 10$

$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$

= SS_{res}

$df(SS_{res}) = \left(\sum_{i=1}^{g} n_i\right) - g$ — Why?

c) SSB = "Sum of squares between the groups"

$= (8-4)^2 + (8-4)^2 + (8-4)^2 + (2-4)^2 + (2-4)^2 + (3-4)^2 + (1-4)^2 + (2-4)^2$

$= 78$

$= \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$

$= SS_{tr}$

$df(SS_{tr}) = g-1$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}^2}_{(SS_{obs})} = \underbrace{SS_{tr}}_{(SS_{mean})} + \underbrace{\sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2}_{(SS_{tr})} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}_{(SS_{res})}$$

ANOVA Table

Source of Variation	Sum of squares (SS)	Degrees of freedom (df)
Treatments	$SS_{tr} = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$	$g-1$
Residual (error)	$SS_{res} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$\sum_{i=1}^{g} n_i - g$
Total	$SS_{cor} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$\sum_{i=1}^{g} n_i - 1$

$$F^* = \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{i=1}^{g} n_i - g)} \sim F(\alpha, g-1, \sum_{i=1}^{g} n_i - g)$$

Reject H_0 if $F^* > F(\alpha, g-1, \sum_{i=1}^{g} n_i - g)$

Back to our example (Testing at $\alpha = 0.01$)

$$F^* = \frac{78/2}{10/5} = 19.5$$

$F^* = 19.5 > F(0.01, 2, 5) = 13.27$ so reject H_0.

Multivariate ANOVA

- Used when comparing more than 2 populations.
- Random samples from each of g populations are arranged as

Population 1 : $\underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, \cdots, \underset{\sim}{X}_{1n_1}$

Population 2 : $\underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, \cdots, \underset{\sim}{X}_{2n_2}$

$\vdots$ $\vdots$

Population g : $\underset{\sim}{X}_{g1}, \underset{\sim}{X}_{g2}, \cdots, \underset{\sim}{X}_{gn_g}$

- We use MANOVA to investigate whether the population mean vectors are the same. If not, there are methods to investigate further which mean components differ significantly.

Assumptions about the structure of the data for one-way MANOVA

1. $\underset{\sim}{X}_{l1}, \underset{\sim}{X}_{l2}, \ldots, \underset{\sim}{X}_{ln_l}$ is a random sample of size n_l from a population with mean $\underset{\sim}{\mu}_l$, $l=1,2,\ldots,g$. The random samples from the different populations are independent.
2. All populations have a common covariance matrix, Σ.
3. Each population is multivariate normal. (Can be relaxed when sample sizes n_l are large.

Refer to problem 9 (Book example 6.9)

Model

$$\underset{\sim}{X}_{lj} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_l + \underset{\sim}{e}_{lj}, \quad j=1,2,\ldots,n_l \text{ and } l=1,2,\ldots,g$$

subject to $\sum_{l=1}^{g} n_l \underset{\sim}{\tau}_l = 0$

- $\underset{\sim}{e}_{lj} \sim N_p(0, \Sigma)$
- $\underset{\sim}{\mu}$ is the overall mean,
- $\underset{\sim}{\tau}_l$ is the lth treatment effect

The model is estimated from the sample as:

$$\underset{\sim}{X}_{lj} = \underset{\sim}{\bar{X}} + (\underset{\sim}{\bar{X}}_l - \underset{\sim}{\bar{X}}) + (\underset{\sim}{X}_{lj} - \underset{\sim}{\bar{X}}_l)$$

(Observation) (overall mean) + (estimated treatment effect, $\hat{\tau}_l$) + (residual $\hat{e}_{lj}$)

As in the univariate case, the total sum of squares is decomposed into the sum of squares between groups and the sum of squares within groups.

MANOVA Table for Comparing Population Mean Vectors

Source of variation	Matrix of sum of squares and cross products	Degrees of freedom
Treatment (Between)	$B = \sum_{i=1}^{g} n_i (\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})^T$ $= (n_1-1)S_1 + (n_2-1)S_2 + \cdots + (n_g-1)S_g$	$g-1$
Residual (Error)	$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)^T$	$\sum_{i=1}^{g} n_i - g$
Total	$B + W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})^T$	$\sum_{i=1}^{g} n_i - 1$

Testing $H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \cdots = \underset{\sim}{\mu}_g$

$H_a: \underset{\sim}{\mu}_i \neq \underset{\sim}{\mu}_j$ for at least one $i \neq j$

We calculate

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\left| \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{i})(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{i})^T \right|}{\left| \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}})^T \right|}$$

Λ^* is known as Wilk's lambda

- The exact distribution of Λ^* can be derived for the special cases as listed in the table below:

for any other case and large sample sizes, (i.e $\left(\sum_{i=1}^{g} n_i = n\right)$) a modification of Λ^* due to Bartlett can be used in the hypothesis test.

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda^* = -\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right)$$

$$\sim \chi^2(\alpha, p(g-1))$$

Reject H_0 if

$$-\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right) > \chi^2(\alpha, p(g-1))$$