What to do today ?

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 Vectors (Chp 6.1-4, 6-7)

II.3.1 Introduction (Chp 6.1)

II.3.2 Comparing Mean Vectors from Two Populations (Chp 6.2-3)

II.3.3 Comparing Several Mean Vectors and Related (Chp 6.4, 6.6)

II.3.3 Two-Way Multivariate Analysis of Variance (Chp 6.7)

II.4 Multivariate Linear Regression (Chp 7)

II.3 Comparisons of Several Mean Vectors (Chp 6)

Consider g populations: $\mathbf{X}_1, \ldots, \mathbf{X}_g$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, \ldots, g$.

- how to compare $\mu_1, \ldots, \mu_g$? (Chp 6.4-6)
- how about to compare μ_1, μ_2 (i.e. $g = 2$)? (Chp 6.2-3)
- what if the g groups may be looked by two ways: (l, k) for $l = 1, \ldots, a$ and $k = 1, \ldots, b$? (Chp 6.7)

the analogues of those in the univariate situations!

II.3.2 Comparing Mean Vectors from Two Populations (Chp 6.2-3)

Consider 2 populations: $\mathbf{X}_1, \mathbf{X}_2$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, 2$.

Goal. to compare $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$

Data. $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are iid observations on $\mathbf{X}_1$, and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ are iid observations on $\mathbf{X}_2$.

The *key idea* is to use $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$.

- $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$
- $Var(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Var(\bar{\mathbf{X}}_1) + Var(\bar{\mathbf{X}}_2) - 2Cov(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$

Scenario A. $\mathbf{X}_1 \perp \mathbf{X}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

- $Var(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\Sigma}/n_1 + \boldsymbol{\Sigma}/n_2$
- Often $\mathbf{S}_{pooled} = \frac{n_1-1}{n_1+n_2-2}\mathbf{S}_1 + \frac{n_2-1}{n_1+n_2-2}\mathbf{S}_2$ is used to estimate $\boldsymbol{\Sigma}$.
- The T^2 statistic follows the Hotelling's T^2-distn $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$:

$$T^2 = \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]$$

II.3.2 Comparing Mean Vectors from Two Populations (Chp 6.2-3)

Consider 2 populations: $\mathbf{X}_1, \mathbf{X}_2$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, 2$.

Goal. to compare $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$

Data. $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are iid observations on $\mathbf{X}_1$, and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ are iid observations on $\mathbf{X}_2$.

The *key idea* is to use $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$.

- $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$
- $Var(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Var(\bar{\mathbf{X}}_1) + Var(\bar{\mathbf{X}}_2) - 2Cov(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$

<u>**Scenario B.**</u> $\mathbf{X}_1 \perp \mathbf{X}_2$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

- Use $\mathbf{S}_1, \mathbf{S}_2$ to estimate $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ correspondingly.
- T^2's distribution is complicate if n_1, n_2 are not large.*
- $T^2 \sim \chi^2_p$ approximately if n_1, n_2 are large.

$$T^2 = \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]' \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]$$

II.3.2 Comparing Mean Vectors from Two Populations (Chp 6.2-3)

Consider 2 populations: $\mathbf{X}_1, \mathbf{X}_2$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, 2$.

Goal. to compare $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$

Data. $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are iid observations on $\mathbf{X}_1$, and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ are iid observations on $\mathbf{X}_2$.

The *key idea* is to use $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$.

- $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$
- $Var(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Var(\bar{\mathbf{X}}_1) + Var(\bar{\mathbf{X}}_2) - 2Cov(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$

<u>**Scenario C.**</u> $\mathbf{X}_1 \not\perp \mathbf{X}_2$

- Given a good estimator for $Var(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$, denoted by $\hat{\boldsymbol{\Pi}}_{n_1,n_2}$, consider

$$T^2 = \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right]' \left[\hat{\boldsymbol{\Pi}}_{n_1,n_2}\right]^{-1} \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right]?$$

- If observations on the two populations $\mathbf{X}_1$ and $\mathbf{X}_2$ are in pairs: $(\mathbf{X}_{1i}, \mathbf{X}_{2i})$ for $i = 1, \ldots, n$,

  change the two-population problem into a one-population problem: $\mathbf{D} = \mathbf{X}_1 - \mathbf{X}_2$ with iid observations $\mathbf{D}_i = \mathbf{X}_{1i} - \mathbf{X}_{2i}$ for $i = 1, \ldots, n$.

II.3.3 Comparing Several Mean Vectors and Related (Chp 6.4, 6.6)

Consider g populations: $\mathbf{X}_1, \ldots, \mathbf{X}_g$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, \ldots, g$.

Goal. to compare $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$

Data. $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are iid observations on $\mathbf{X}_1$; $\ldots$; $\mathbf{X}_{g1}, \ldots, \mathbf{X}_{gn_g}$ are iid observations on $\mathbf{X}_g$.

Test on $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g = \boldsymbol{\mu}$ with type I error α?

Consider the ANOVA model: for $l = 1, \ldots, g$,

$$\mathbf{X}_{li} = \boldsymbol{\mu} + [\boldsymbol{\mu}_l - \boldsymbol{\mu}] + \boldsymbol{\epsilon}_{li}, \quad \boldsymbol{\epsilon}_{li} \sim MN_p(0, \boldsymbol{\Sigma}) \ \ iid \ \ i = 1, \ldots, n_l.$$

An analogous decomposition of the observations:

$$\underset{\text{(obstn)}}{\mathbf{x}_{li}} = \underset{\begin{pmatrix}\text{overall}\\ \text{sample mean}\end{pmatrix}}{\bar{\mathbf{x}}} + \underset{\begin{pmatrix}\text{estm}\\ \text{trt effect}\end{pmatrix}}{[\bar{\mathbf{x}}_l - \bar{\mathbf{x}}]} + \underset{\text{(residual)}}{[\mathbf{x}_{li} - \bar{\mathbf{x}}_l]}$$

$$\underset{(\mathbf{SS}_{cor})}{\sum_{l=1}^{g}\sum_{i=1}^{n_l}(\mathbf{x}_{li} - \bar{\mathbf{x}})(\mathbf{x}_{li} - \bar{\mathbf{x}})'} = \underset{(\mathbf{SS}_{tr})}{\sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'} + \underset{(\mathbf{SS}_{res})}{\sum_{l=1}^{g}\sum_{i=1}^{n_l}(\mathbf{x}_{li} - \bar{\mathbf{x}}_l)(\mathbf{x}_{li} - \bar{\mathbf{x}}_l)'}$$

Multivariate ANOVA Table ($n_T = \sum_{l=1}^{g} n_l$)

Source of Variation	df	SS
treatment	g-1	$\mathbf{SS}_{trt} = \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$
error	$n_T - g$	$\mathbf{SS}_{res} = \sum_{l=1}^{g} \sum_{i=1}^{n_l} (\mathbf{x}_{li} - \bar{\mathbf{x}}_l)(\mathbf{x}_{li} - \bar{\mathbf{x}}_l)'$
total	$n_T - 1$	$\mathbf{SS}_{cor} = \sum_{l=1}^{g} \sum_{i=1}^{n_l} (\mathbf{x}_{li} - \bar{\mathbf{x}})(\mathbf{x}_{li} - \bar{\mathbf{x}})'$

To test on $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$ using the Wilks' lambda statistic: $\Lambda^* = \frac{|\mathbf{SS}_{res}|}{|\mathbf{SS}_{cor}|}$.

- Reject H_0 if Λ^*_{obs} is small.
- Textbook Table 6.3 presents the distn of Λ^*.
- We use software to implement the test (e.g. *manova()* function in R).

II.3.3 Comparing Several Mean Vectors and Related (Chp 6.4, 6.6)

- The MNOVA model assumes the g populations have the same population variance: $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 1, \ldots, g$.
- It appears easier to handle in *Part II.3.2 Comparing Mean Vectors from Two Populations* when the two populations have the same variance.

Is there a way to test for equality of variance matrices?

Consider g populations: $\mathbf{X}_1, \ldots, \mathbf{X}_g$. Suppose $\mathbf{X}_j \sim MN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, \ldots, g$.

Data. $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are iid observations on $\mathbf{X}_1$; $\ldots$; $\mathbf{X}_{g1}, \ldots, \mathbf{X}_{gn_g}$ are iid observations on $\mathbf{X}_g$.

Test on $H_0 : \boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ with type I error α?

Box's M - Test.

For the multivariate normal populations with the given data, the likelihood ratio statistic for testing H_0 is

$$\Lambda = \prod_{l=1}^{g} \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{pooled}|} \right)^{(n_l-1)/2}$$

$\mathbf{S}_l$ is the lth group's sample variance, and
$\mathbf{S}_{pooled} = \frac{1}{\sum_{l=1}^{g}(n_l-1)} \{(n_1-1)\mathbf{S}_1 + \ldots + (n_g-1)\mathbf{S}_g\}$.

Box's M - statistic: $M = -2 \ln \Lambda$

$$C = (1-u)M \sim \chi^2(\nu) \text{ approximately under } H_0$$

$\nu = p(p+1)(g-1)/2$ and u is given in (6-51) of the textbook. Reject H_0 if $C_{obs} > \chi^2_\nu(\alpha)$.

- The approximation works well when $n_l > 20$ for $l = 1, \ldots, g$, and $p \leq 5$ and $g \leq 5$.

II.3.3 Two-Way Multivariate Analysis of Variance (Chp 6.7)

univariate 2-way ANOVA model: Suppose a study with two factors, one with g levels and the other with b levels: the rth observation from the group of (l, k)

$$X_{lkr} = \mu_{lk} + \epsilon_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + \epsilon_{lkr}$$

iid $\epsilon_{lkr} \sim N(0, \sigma^2)$ for $l = 1, \ldots, g$, $k = 1, \ldots, b$, and $r = 1, \ldots, n$, and $\sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{lk} = \sum_{k=1}^{b} \gamma_{lk} = 0$.

To test on $H_{01} : \tau_l = 0$, $H_{02} : \beta_k = 0$, and $H_{012} : \gamma_{lk} = 0$, consider the observation decomposition:

$$SS_{cor} = SS_{fac1} + SS_{fac2} + SS_{int} + SS_{res}$$

Source of Variation	df	SS	F-value
factor 1	g-1	SS_{fac1}	$F_1 = \frac{MSS_{fac1}}{MSS_{res}}$
factor 2	b-1	SS_{fac2}	$F_2 = \frac{MSS_{fac2}}{MSS_{res}}$
interaction	(g-1)(b-1)	SS_{int}	$F_{12} = \frac{MSS_{int}}{MSS_{res}}$
error	$gb(n-1)$	SS_{res}	
total	$gbn-1$	SS_{cor}	

Reject H_{01} if $F_{1,obs} > F_{g-1,gb(n-1)}(\alpha)$, reject H_{02} if $F_{2,obs} > F_{b-1,gb(n-1)}(\alpha)$, and reject H_{012} if $F_{12,obs} > F_{(g-1)(b-1),gb(n-1)}(\alpha)$.

II.3.3 Two-Way Multivariate Analysis of Variance (Chp 6.7)

Suppose a study with two factors, one with g levels and the other with b levels: the rth observation from the group of (l, k)

$$\mathbf{X}_{lkr} = \boldsymbol{\mu}_{lk} + \boldsymbol{\epsilon}_{lkr} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{lk} + \boldsymbol{\epsilon}_{lkr}$$

iid $\boldsymbol{\epsilon}_{lkr} \sim MN(0, \boldsymbol{\Sigma})$ for $l = 1, \ldots, g$, $k = 1, \ldots, b$, and $r = 1, \ldots, n$, and $\sum_{l=1}^{g} \boldsymbol{\tau}_l = \sum_{k=1}^{b} \boldsymbol{\beta}_k = \sum_{l=1}^{g} \boldsymbol{\gamma}_{lk} = \sum_{k=1}^{b} \boldsymbol{\gamma}_{lk} = \mathbf{0}$.

To test on $H_{01} : \boldsymbol{\tau}_l = 0$, $H_{02} : \boldsymbol{\beta}_k = 0$, and $H_{012} : \boldsymbol{\gamma}_{lk} = 0$, consider the observation decomposition:

$$\mathbf{SS}_{cor} = \mathbf{SS}_{fac1} + \mathbf{SS}_{fac2} + \mathbf{SS}_{int} + \mathbf{SS}_{res}$$

Source of Variation	df	SS	Wilks's lambda
factor 1	g-1	$\mathbf{SS}_{fac1}$	$\Lambda_1 = \frac{\lvert\mathbf{SS}_{res}\rvert}{\lvert\mathbf{SS}_{fac1}+\mathbf{SS}_{res}\rvert}$
factor 2	b-1	$\mathbf{SS}_{fac2}$	$\Lambda_2 = \frac{\lvert\mathbf{SS}_{res}\rvert}{\lvert\mathbf{SS}_{fac2}+\mathbf{SS}_{res}\rvert}$
interaction	(g-1)(b-1)	$\mathbf{SS}_{int}$	$\Lambda_{12} = \frac{\lvert\mathbf{SS}_{res}\rvert}{\lvert\mathbf{SS}_{int}+\mathbf{SS}_{res}\rvert}$
error	$gb(n-1)$	$\mathbf{SS}_{res}$	
total	$gbn-1$	$\mathbf{SS}_{cor}$	

Reject H_{01} if $\Lambda_{1,obs}$ is small, reject H_{02} if $\Lambda_{2,obs}$ is small, and reject H_{012} if $\Lambda_{12,obs}$ is small.

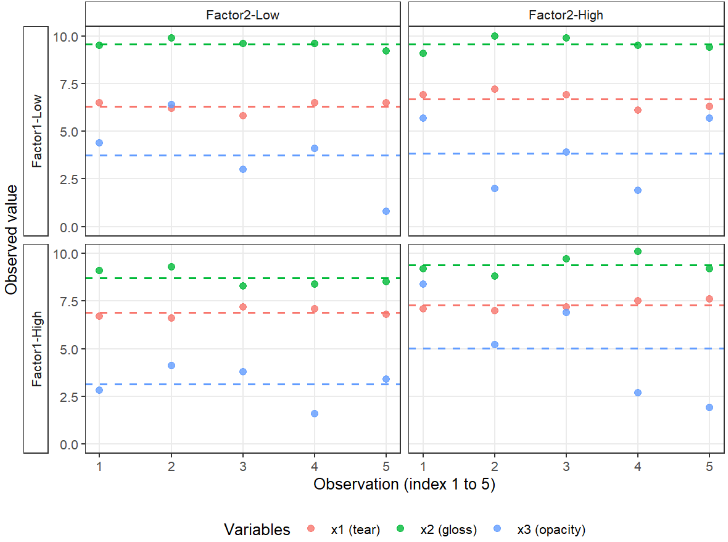
Example: Plastic film data (Textbook Example 6.13, p318)

- responses: X_1 = tear resistance, X_2 = gloss, X_3 = opacity. $n = 5$.
- factor 1: rate of extrusion – Low, High; factor 2: amount of an additive – Low, High.
-

		Factor 2: Amount of additive					
		Low			High		
Factor 1	Rep	x_1	x_2	x_3	x_1	x_2	x_3
Low	1	6.5	9.5	4.4	6.9	9.1	5.7
	2	6.2	9.9	6.4	7.2	10.0	2.0
	3	5.8	9.6	3.0	6.9	9.9	3.9
	4	6.5	9.6	4.1	6.1	9.5	1.9
	5	6.5	9.2	0.8	6.3	9.4	5.7
High	1	6.7	9.1	2.8	7.1	9.2	8.4
	2	6.6	9.3	4.1	7.0	8.8	5.2
	3	7.2	8.3	3.8	7.2	9.7	6.9
	4	7.1	8.4	1.6	7.5	10.1	2.7
	5	6.8	8.5	3.4	7.6	9.2	1.9

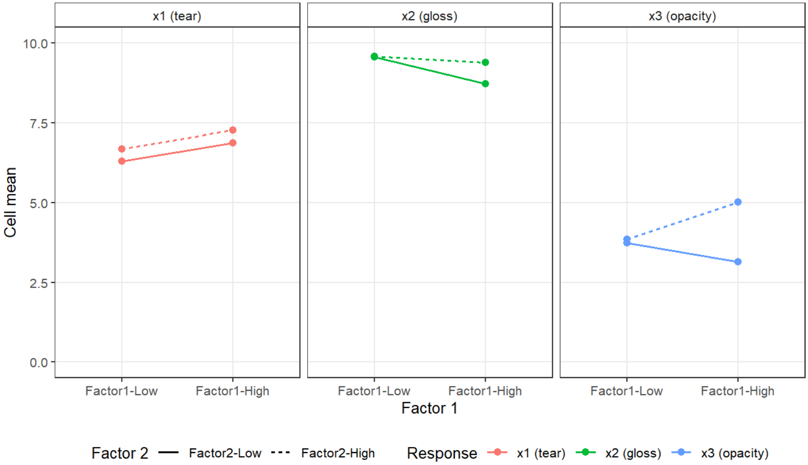
Example: Plastic film data (Textbook Example 6.13, p318)

- What we are comparing for $H_{01} : \boldsymbol{\tau}_l = 0$, $H_{02} : \boldsymbol{\beta}_k = 0$?



Example: Plastic film data (Textbook Example 6.13, p318)

- What we are comparing for $H_{012} : \gamma_{lk} = 0$?



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 Vectors (Chp 6)*
  - **II.4 Multivariate Linear Regression (Chp 7)**
    - **II.4.1 Introduction (Chp7.1)**
    - **II.4.2 Classical Linear Regression (Chp7.2-3)**
    - **II.4.3 Linear Regression Based Inference (Chp7.4-5)**
    - *II.4.4 Model Checking (Chp7.6)*
    - *II.4.3 Multivariate Multiple Regression (Chp7.7)*
- *Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-11)*