

# Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos  
Departamento de Matemáticas  
Pontificia Universidad Javeriana

Semana 4(Parte 2): 05/08/24 – 10/08/24



# Inferencia Sobre $\mu$

Considere el problema de evaluar la igualdad (o diferencia) de medias multidimensional de **diferentes poblaciones**. Para esto, debemos considerar los siguientes casos:

1. **Pareadas** o medidas repetidas.

*Las  $p$  variables se evalúan en la **misma unidad de muestreo en “dos” condiciones diferentes** (por ejemplo, *antes* y *después*).*

2. Dos poblaciones **independientes**.
3. Más de dos poblaciones **independientes**.



# Inferencia sobre $\mu$

Considere que se quiere estudiar un tratamiento para bajar de peso, y que además del peso ( $X_1$ ) se consideran la estatura ( $X_2$ ) y la edad ( $X_3$ ) de los participantes del estudio.

En este caso:

- ✓ hay tres variables,



# Inferencia sobre $\mu$

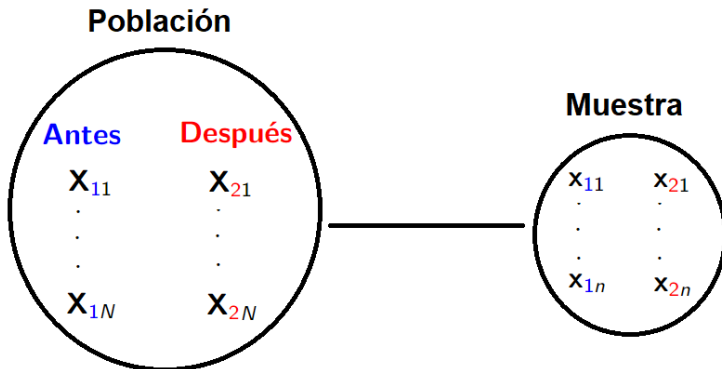
Considere que se quiere estudiar un tratamiento para bajar de peso, y que además del peso ( $X_1$ ) se consideran la estatura ( $X_2$ ) y la edad ( $X_3$ ) de los participantes del estudio.

En este caso:

- ✓ hay tres variables,
- ✓ las variables están involucradas en dos **grupos/situaciones** (antes y después del tratamiento) provienen de la **misma población**,
- ✓ habrá **correlación entre las variables de los dos grupos**, puesto que están medidas en la **misma unidad de muestreo** (participante),



# Inferencia sobre $\mu$



$$f_{\mathbf{X}} \left( \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)$$

# Inferencia sobre $\mu$

- ✓ el interés es **comparar** las **medias** de los dos grupos,

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

con

$$\mu_1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{bmatrix} ; \quad \mu_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix}$$

siendo

- $\mu_{11}$  y  $\mu_{21}$  la media de  $X_1$  **antes** y **después** del tratamiento, respectivamente.  $\mathbf{X}_{2M}$
- $\mu_{12}$  y  $\mu_{22}$  la media de  $X_2$  **antes** y **después** del tratamiento, respectivamente.
- $\mu_{13}$  y  $\mu_{23}$  la media de  $X_3$  **antes** y **después** del tratamiento, respectivamente.



# Inferencia sobre $\mu$

Los registros podrían ser representados por:

Uni. Muestreo	Antes			Después		
	$X_1$	$X_2$	$X_3$	$X_1$	$X_2$	$X_3$
1	$x_{111}$	$x_{112}$	$x_{113}$	$x_{211}$	$x_{212}$	$x_{213}$
2	$x_{121}$	$x_{122}$	$x_{123}$	$x_{221}$	$x_{222}$	$x_{223}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$j$	$x_{1j1}$	$x_{1j2}$	$x_{1j3}$	$x_{2j1}$	$x_{2j2}$	$x_{2j3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{1n1}$	$x_{1n2}$	$x_{1n3}$	$x_{2n1}$	$x_{2n2}$	$x_{2n3}$

Este es el contexto de **muestras pareadas**.

# Inferencia sobre $\mu$

Suponga que tres variables  $X_1, X_2, X_3$  fueron medidas en dos grupos: **Mujeres** y **Hombres** que se desean comparar

$$H_0 : \mu_{\text{Mujeres}} = \mu_{\text{Hombres}}$$

$$H_1 : \mu_{\text{Mujeres}} \neq \mu_{\text{Hombres}}$$

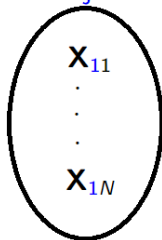
Si bien aquí también se midieran las **mismas tres variables en cada grupo**, éstos sería **independientes**. En este caso, las **muestras** son **independientes**.





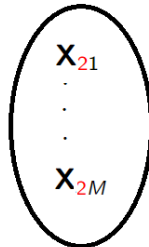
# Inferencia sobre $\mu$

Mujeres



$$f_{\mathbf{X}} \left( \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{bmatrix} \right)$$

Hombres



$$f_{\mathbf{X}} \left( \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} \right)$$

# Inferencia sobre $\mu$

## Muestras Pareadas:

- Sea  $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n}$  vectores aleatorios  $p \times 1$  de una población normal multivariada **antes de un tratamiento** con  $\mu_1 = E[\mathbf{X}_{1j}]$ ,  $j = 1, \dots, n$ .
- Sea  $\mathbf{X}_{21}, \dots, \mathbf{X}_{2n}$  vectores aleatorios  $p \times 1$  de una población normal multivariada **después de un tratamiento** con  $\mu_2 = E[\mathbf{X}_{2j}]$ ,  $j = 1, \dots, n$ .
- Asuma que  $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n}$  y  $\mathbf{X}_{21}, \dots, \mathbf{X}_{2n}$  son muestras aleatorias de una **misma población** en diferentes situaciones, donde  $\mathbf{X}_{1j}$  y  $\mathbf{X}_{2j}$  están **correlacionados**, por ejemplo, vectores aleatorios de mediciones **antes** ( $\mathbf{X}_{1j}$ ) y **después** de un tratamiento ( $\mathbf{X}_{2j}$ ).



# Inferencia sobre $\mu$

Etiquetamos las  $p$  mediciones dentro de la  $j$ -ésima unidad como:

$X_{1j1}$  = variable 1 dentro del tratamiento 1

$X_{1j2}$  = variable 2 dentro del tratamiento 1

$\vdots$

$X_{1jp}$  = variable  $p$  dentro del tratamiento 1

.....

$X_{2j1}$  = variable 1 dentro del tratamiento 2

$X_{2j2}$  = variable 2 dentro del tratamiento 2

$\vdots$

$X_{2jp}$  = variable  $p$  dentro del tratamiento 2



# Inferencia sobre $\mu$

Las observaciones multivariadas son evaluadas en la **misma unidad muestral** en **dos situaciones diferentes**:

$$\mathbf{X}_{1j} = \begin{bmatrix} X_{1j1} \\ X_{1j2} \\ \vdots \\ X_{1jp} \end{bmatrix} ; \quad \mathbf{X}_{2j} = \begin{bmatrix} X_{2j1} \\ X_{2j2} \\ \vdots \\ X_{2jp} \end{bmatrix}$$

Suponga que el interés es verificar que el **tratamiento no produce ningún efecto**:

$$H_0 : \mu_1 = \mu_2 \quad ; \quad H_0 : \mu_D = \mu_1 - \mu_2 = 0$$

# Inferencia sobre $\mu$

Se pueden plantear las hipótesis

$$H_0 : \mu_D = \delta_0$$

$$H_1 : \mu_D \neq \delta_0$$

siendo (en este caso)  $\delta_0 = \mathbf{0}$ , y cuya **estadística de prueba** está dada por

$$n \left( \bar{\mathbf{D}} - \delta_0 \right)^\top \mathbf{S}_D^{-1} \left( \bar{\mathbf{D}} - \delta_0 \right) \stackrel{H_0 \text{ verd.}}{\sim} \frac{(n-1)p}{n-p} F_{p, n-p}$$

donde

$$\bar{\mathbf{D}} = \frac{1}{n} \sum_{j=1}^n \mathbf{D}_j \quad ; \quad \mathbf{S}_D = \frac{1}{n-1} \sum_{j=1}^n \left( \mathbf{D}_j - \bar{\mathbf{D}} \right) \left( \mathbf{D}_j - \bar{\mathbf{D}} \right)^\top$$



# Inferencia sobre $\mu$

son respectivamente el **vector de medias** y la **matriz de varianzas y covarianzas muestrales** de las diferencias  $\mathbf{D}_j = \mathbf{X}_{1j} - \mathbf{X}_{2j}$ :

$$D_{j1} = X_{1j1} - X_{2j1}$$

$$D_{j2} = X_{1j2} - X_{2j2}$$

$$\vdots$$

$$D_{jp} = X_{1jp} - X_{2jp}$$

$$\mathbf{D}_j^\top = [D_{j1}, D_{j2}, \dots, D_{jp}]$$



# Inferencia sobre $\mu$

La **región**  $(1 - \alpha)100\%$  de **confianza**

$$\left\{ \delta : n (\bar{\mathbf{D}} - \delta)^{\top} \mathbf{S}_D^{-1} (\bar{\mathbf{D}} - \delta) \leq \frac{(n-1)p}{n-p} F_{p, n-p}(\alpha) \right\}$$

El **intervalo**  $(1 - \alpha)100\%$  de confianza **simultaneo** para diferencia de medias individuales  $\delta_k$ :

$$\bar{D}_k \pm \sqrt{\frac{(n-1)p}{n-p} F_{p, n-p}(\alpha)} \sqrt{\frac{S_{Dkk}}{n}}$$

donde  $S_{Dkk}$  es la varianza de la  $k$ -ésima diferencia.



# Inferencia sobre $\mu$

El **intervalo**  $(1 - \alpha)100\%$  de confianza **simultaneo Bonferroni** para diferencia de medias individuales  $\delta_k$ :

$$\bar{D}_k \pm t_{n-1} \left( \frac{\alpha}{2p} \right) \sqrt{\frac{S_{Dkk}}{n}}$$

donde  $S_{Dkk}$  es la varianza de la  $k$ -ésima diferencia.





# Inferencia sobre $\mu$

## Example

Las plantas de tratamiento de aguas residuales municipales están obligadas por ley a controlar periódicamente sus descargas en ríos y arroyos. La preocupación por la confiabilidad de los datos de uno de estos programas de autocontrol llevó a un estudio en el que se **dividieron muestras de efluentes** y se enviaron a **dos laboratorios** para su análisis. La **mitad de cada muestra** fue enviada a el **Laboratorio de Higiene del Estado de Wisconsin**, y la otra mitad a un **laboratorio comercial privado** utilizado habitualmente en el programa de seguimiento. Se obtuvieron **mediciones de la demanda bioquímica de oxígeno (DBO) y de sólidos suspendidos (SS)**, para  $n = 11$  divisiones de muestra, de los dos laboratorios.

*Observación: Example 6.1 de Johnson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 276*

# Inferencia sobre $\mu$

Table 6.1 Effluent Data				
Sample $j$	Commercial lab		State lab of hygiene	
	$x_{1j1}$ (BOD)	$x_{1j2}$ (SS)	$x_{2j1}$ (BOD)	$x_{2j2}$ (SS)
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Source: Data courtesy of S. Weber.

¿Coinciden los análisis químicos de los dos laboratorios? Si existen diferencias, ¿cuál es su naturaleza?

# Inferencia sobre $\mu$

*Observación: El **primer subíndice** hace referencia al **laboratorio** que en este caso son los **tratamientos o situaciones**.*

Observe que:

- se evalúan las mismas variables (BOD y SS) en cada condición (Commercial lab y State lab of hygiene).
- las muestras definen el emparejamiento o dependencia entre los dos conjuntos.
- El análisis se extiende a situaciones con dos conjuntos diferentes de variables.



# Inferencia sobre $\mu$

El interés es probar las hipótesis

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

donde

- $\mu_1$  es el vector de medias (media de BOD y media SS) del laboratorio comercial.
- $\mu_2$  es el vector de medias (media de BOD y media SS) del laboratorio de higiene del estado de Wisconsin.



# Inferencia sobre $\mu$

```
# ----- Muestras Pareadas ----- #  
X1j1<-c(6,6,18,8,11,34,28,71,43,33,20)      # X1 Lab 1  
X1j2<-c(27,23,64,44,30,75,26,124,54,30,14)  # X2 Lab 1  
X2j1<-c(25,28,36,35,15,44,42,54,34,29,39)   # X1 Lab 2  
X2j2<-c(15,13,22,29,31,64,30,64,56,20,21)   # X2 Lab 2  
  
X<-data.frame(X1j1,X1j2,X2j1,X2j2)  
  
Dj1<-X$X1j1-X$X2j1  # Dif. entre las X1 de los 2 lab.  
Dj2<-X$X1j2-X$X2j2  # Dif. entre las X2 de los 2 lab.  
Dbarra1<-mean(Dj1)  
Dbarra2<-mean(Dj2)  
Dbarra<-c(Dbarra1,Dbarra2)
```



# Inferencia sobre $\mu$

```
SD<-matrix(c(var(Dj1),cov(Dj1,Dj2),  
             cov(Dj1,Dj2),var(Dj2)),ncol=2)
```

```
T2<-11*t(Dbarra)%*%solve(SD)%*%Dbarra  
> T2
```

```
      [,1]  
[1,] 13.63931
```

```
Fc<-qf(0.05, 2,9,lower.tail = F)  
vc<-((10*2)/9)*Fc  
> vc  
[1] 9.458877
```

¿Cuál es la conclusión?



## Comparaciones de dos Poblaciones Normales Independientes

- Sea  $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1}$  vectores aleatorios  $p \times 1$  de una población normal multivariada con  $\mu_1 = E[\mathbf{X}_{1j}]$  y  $\Sigma_1$   $j = 1, \dots, n_1$ .
- Sea  $\mathbf{X}_{21}, \dots, \mathbf{X}_{2n}$  vectores aleatorios  $p \times 1$  de una población normal multivariada con  $\mu_2 = E[\mathbf{X}_{2j}]$  y  $\Sigma_2$ ,  $j = 1, \dots, n_2$ .
- Asuma que las poblaciones son independientes y que el interés es probar las hipótesis

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

$$H_1 : \mu_1 - \mu_2 \neq \delta_0$$

*Observación: Aquí se debe tener en cuenta si las matrices  $\Sigma_1$  y  $\Sigma_2$  son **iguales o diferentes**.*



# Inferencia sobre $\mu$

- Asumiendo que  $\Sigma_1 = \Sigma_2 = \Sigma$  (**homocedasticidad u homogeneidad**) y  $n_1$  y  $n_2$  **pequeños**, se **rechaza**  $H_0$  a un nivel de significancia  $\alpha$  si

$$T_{obs}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0)^\top \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0)$$
$$> \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

donde

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{x}_{1j} \qquad \bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{x}_{2j}$$





# Inferencia sobre $\mu$

$$\mathbf{S}_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} \mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1) (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^\top$$

$$\mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2) (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^\top$$



# Inferencia sobre $\mu$

Los **intervalos simultaneos** de  $T^2$

$$\overline{D_K} \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha)} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{KK, pooled}}$$

- Si las **matrices de covarianzas** de las dos poblaciones son **diferentes (heterocedasticidad o heterogeneidad)**,  $\Sigma_1 \neq \Sigma_2$ , y  $n_1 - p$  y  $n_2 - p$  **grandes**, se **rechaza**  $H_0$  a un nivel de significancia  $\alpha$  si

$$T_{obs}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0)^\top \left[ \frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0) > \chi_p^2(\alpha)$$



# Inferencia sobre $\mu$

Para tamaños de **muestra pequeños**, la estadística está dada por

$$T^{*2} = [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^\top \left[ \frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$
$$\underset{H_0 \text{ verd.}}{\sim} \frac{vp}{v - p + 1} F_{p, v-p+1}$$

donde

$$v = \frac{p + p^2}{\sum_{i=1}^2 \frac{1}{n_i} \left\{ \text{traza} \left[ \left( \frac{1}{n_i} \mathbf{S}_i \left( \frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right)^{-1} \right)^2 \right] + \left( \text{traza} \left[ \frac{1}{n_i} \mathbf{S}_i \left( \frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right)^{-1} \right] \right)^2 \right\}}$$

con  $\min\{n_1, n_2\} \leq v \leq n_1 + n_2$ . **Rechace**  $H_0$  a un nivel de significancia  $\alpha$  si

$$T^{*2} > \frac{vp}{v - p + 1} F_{p, v-p+1}(\alpha)$$



# Inferencia sobre $\mu$

## Example

Considere que se tienen mediciones de productividad y altura de plantas de dos variedades:

Variedad A		Variedad B	
$X_{11}$	$X_{12}$	$X_{21}$	$X_{22}$
5.7	2.1	4.4	1.8
8.9	1.9	7,5	1.75
6.2	1.98	5.4	1.78
5.8	1.92	4.6	1.89
6.8	2	5.9	1.9
6.2	2.01		

Pruebe la **igualdad** del **vector medio** de las **dos variedades**, bajo **homocedasticidad**.

# Inferencia sobre $\mu$

En este caso tenemos:

- **Dos muestras** de poblaciones **independientes**, variedad A y la variedad B.
- En cada muestra se miden **dos** características (**variables**)

	Variedad A		Variedad B	
	$X_{11}$	$X_{12}$	$X_{21}$	$X_{22}$
	5.7	2.1	4.4	1.8
	8.9	1.9	7,5	1.75
	6.2	1.98	5.4	1.78
	5.8	1.92	4.6	1.89
	6.8	2	5.9	1.9
	6.2	2.01		
Media	6.6	1.985	5.56	1.824
Varianza	1.42	0.005	1.543	0.0045

# Inferencia sobre $\mu$

Por lo tanto,

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 6.6 \\ 1.985 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 5.56 \\ 1.824 \end{bmatrix}$$

$$\mathbf{S}_1 = \begin{bmatrix} 1.42 & -0.0504 \\ -0.0504 & 0.005 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} 1.543 & -0.037 \\ -0.037 & 0.0045 \end{bmatrix}$$

$$\mathbf{S}_{pooled} = \frac{6 - 1}{6 + 5 - 2} \begin{bmatrix} 1.42 & -0.0504 \\ -0.0504 & 0.005 \end{bmatrix} + \frac{5 - 1}{6 + 5 - 2} \begin{bmatrix} 1.543 & -0.037 \\ -0.037 & 0.0045 \end{bmatrix}$$

$$= \begin{bmatrix} 1.4745 & -0.0442 \\ -0.0442 & 0.0049 \end{bmatrix}$$



# Inferencia sobre $\mu$

Luego,

$$T_{obs}^2 = \begin{bmatrix} 6.6 - 5.56 & 1.985 - 1.824 \end{bmatrix} \left[ \left( \frac{1}{6} + \frac{1}{5} \right) \mathbf{S} \right]^{-1} \begin{bmatrix} 6.6 - 5.56 \\ 1.985 - 1.824 \end{bmatrix} \\ = 24.91803$$

$$\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha) = \frac{(9)2}{8} 4.45897 = 10.03268$$

Como

$$T_{obs}^2 = 24.91803 > 10.03268 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

**Se rechaza  $H_0 : \mu_A = \mu_B$ .**



# Inferencia sobre $\mu$

Los **intervalos simultáneos**  $T^2$

$$6.6 - 5.56 \pm \sqrt{\frac{(9)^2}{8} F_{2,8}(0.05)} \sqrt{\left(\frac{1}{6} + \frac{1}{5}\right)} 1.47 = (-1.29, 3.37)$$

$$1.985 - 1.824 \pm \sqrt{\frac{(9)^2}{8} F_{2,8}(0.05)} \sqrt{\left(\frac{1}{6} + \frac{1}{5}\right)} 0.0049 = (0.0027, 0.295)$$





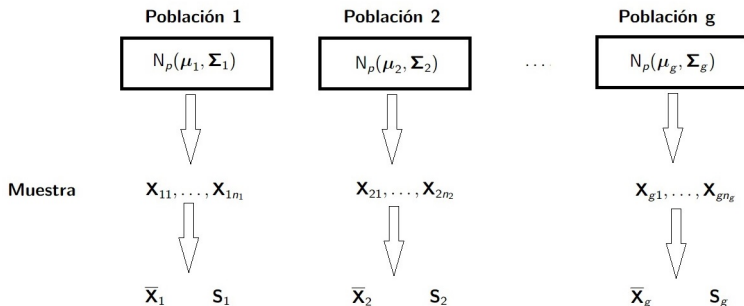
# Inferencia sobre $\mu$

Imagine que:

- se tienen **3 o más grupos** ( $g$ ),
- dentro de **cada grupo** tiene un número de observaciones ( $n_i$ ),
- en **cada observación** se tienen  $p$  **variables**.

Suponga que el interés es determinar si las medias poblacionales son iguales para todos los grupos.

# Inferencia sobre $\mu$



*Observación: Recuerde que la población está estratificada o clasificada (métodos, categorías, etapas, etc)*

## Comparaciones en más de dos Poblaciones Independientes

- $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1}$  una muestra aleatoria de una población normal multivariada con  $\mu_1 = E[\mathbf{X}_{1j}]$  y  $\Sigma = \text{Var}[\mathbf{X}_{1j}]$   $j = 1, \dots, n_1$ .
- $\mathbf{X}_{21}, \dots, \mathbf{X}_{2n_2}$  una muestra aleatoria de una población normal multivariada con  $\mu_2 = E[\mathbf{X}_{2j}]$  y  $\Sigma = \text{Var}[\mathbf{X}_{2j}]$   $j = 1, \dots, n_2$ .
- $\mathbf{X}_{g1}, \dots, \mathbf{X}_{gn_g}$  una muestra aleatoria de una población normal multivariada con  $\mu_g = E[\mathbf{X}_{gj}]$  y  $\Sigma = \text{Var}[\mathbf{X}_{gj}]$   $j = 1, \dots, n_g$ .
- Asuma que todas las poblaciones son independientes entre sí y que el interés es probar las hipótesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g = \mu$$

$$H_1 : \text{al menos un } \mu_i \text{ es diferente}$$

# Inferencia sobre $\mu$

Para probar esas hipótesis consideramos la **reparametrización**:

$$\mu_k = \mu + \tau_k \quad k = 1, \dots, g$$

Por lo que basta probar

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_g = 0$$

$$H_1 : \text{al menos un } \tau_k \text{ es diferente de } 0$$

*Observación:  $\tau_k$  es el efecto del grupo  $k$ .*



# Inferencia sobre $\mu$

El modelo **ANOVA multivariado (MANOVA)** está dado por:

$$\mathbf{X}_{kj} = \mu + \tau_k + \epsilon_{kj}$$

$$j = 1, \dots, n_k, k = 1, \dots, g.$$

**Supuestos** del modelo:

- $\epsilon_{kj} \stackrel{iid}{\sim} N(\mathbf{0}; \Sigma)$
- $\mu$  es la **media general**.
- $\tau_k$  es el **efecto** del  $k$ -ésimo **grupo**.
- $\sum_{k=1}^g n_k \tau_k = \mathbf{0}$  (para garantizar la **identificabilidad** del modelo)



# Inferencia sobre $\mu$

La **tabla Manova** queda dada por

Fuente de variación	Sumas de Cuadrados	Grados de Libertad
Tratamiento	$B$	$g - 1$
Residuo	$W$	$N - g$
Total	$T$	$N - 1$

donde

- $T$  es la **suma de cuadrados total**,

$$T = \sum_{k=1}^g \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \bar{\mathbf{x}}) (\mathbf{x}_{kj} - \bar{\mathbf{x}})^{\top}$$

- $B$  es la **suma de cuadrados entre los grupos**

$$B = \sum_{k=1}^g n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^{\top}$$



# Inferencia sobre $\mu$

- $W$  es la **suma de cuadrados dentro (intra) de los grupos**

$$W = \sum_{k=1}^g \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \bar{\mathbf{X}}_k) (\mathbf{x}_{kj} - \bar{\mathbf{X}}_k)^{\top}$$

- $N = \sum_{k=1}^g n_k$

Lo que **se desea verificar** es si  $W$  es **pequeño con respecto a  $T$** , esto es, **si todas las medias pueden ser consideradas iguales, sobraría poco para el residuo**. En consecuencia, para probar ésto, usamos la **estadística Lambda de Wilk** (Varianza Generalizada):



# Inferencia sobre $\mu$

$$\Lambda^* = \frac{|W|}{|W + B|} = \frac{|W|}{|T|}$$

La distribución de  $\Lambda^*$  es complicada, depende del número de variables y del número de grupos. Se estudiaron algunos casos para los cuales existen tablas.





# Inferencia sobre $\mu$

## Example

Los datos Skulls del paquete `heplots` de R, contiene 150 observaciones y **5 variables** de cráneos egipcios de **cinco épocas**:

- **epoch**: época en que el cráneo fue atribuido. Es un factor ordenado con **5 niveles**: c4000BC, c3300BC, c1850BC, c200BC, cAD150
- **mB**: largura máxima del cráneo.
- **bh**: altura de la base bregmática del cráneo.
- **bl**: longitud de la base alveolar del cráneo.
- **nh**: altura nasal del cráneo

El objetivo es **comparar medias** multidimensionales en los **5 grupos**.

# Inferencia sobre $\mu$

## Solución:

En este caso tenemos  $g = 5$  grupos.

Considerando

- Grupo 1 (c4000BC):  $\mathbf{X}_1 \sim N(\mu_1, \Sigma_1)$
- Grupo 2 (c3300BC):  $\mathbf{X}_2 \sim N(\mu_2, \Sigma_2)$
- Grupo 3 (c1850BC):  $\mathbf{X}_3 \sim N(\mu_3, \Sigma_3)$
- Grupo 4 (c200BC):  $\mathbf{X}_4 \sim N(\mu_4, \Sigma_4)$
- Grupo 5 (cAD150):  $\mathbf{X}_5 \sim N(\mu_5, \Sigma_5)$

el interés es probar

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu$$

$H_1$  : al menos un  $\mu_i$  es diferente



# Inferencia sobre $\mu$

En R:

```
# ----- Ex. Datos Skulls ----- #  
require(heplots)  
data("Skulls")  
require(mvShapiroTest)
```

```
# ----- Prueba de Normalidad Multivariada ----- #  
mvShapiro.Test(as.matrix(Skulls[,2:5]))
```

Generalized Shapiro-Wilk test for Multivariate  
Normality by Villasenor-Alva and Gonzalez-Estrada

```
data:  as.matrix(Skulls[, 2:5])  
MVW = 0.99212, p-value = 0.755
```



# Inferencia sobre $\mu$

```
# ----- Modelo para cada variable
mod.skulls<-lm(cbind( mb,  bh,  bl, nh) ~ epoch,
               data=Skulls)

summary(mod.skulls)

# ----- Manova
mod<-manova(mod.skulls)
summary(mod, test = "Wilks")
```

# Inferencia sobre $\mu$

```
          Df    Wilks approx F num Df den Df    Pr(>F)
epoch          4 0.66359    3.9009      16 434.45 7.01e-07 ***
Residuals 145
```

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Observe que  $p - \text{valor} = 7.01e - 07$  es altamente significativo, por lo tanto que **hay suficiente evidencia** en la muestra para **rechazar**  $H_0$ , así **almenos un  $\mu_i$  es diferente** de los otros,  $i = 1, 2, \dots, 5$ .

