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 μ

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 oservacioes multivariadas 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.

Muestras Pareadas:

- Sea $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n}$ vectores aleatorios $p \times 1$ de una población normal multivariada antes de un tratamiento con $\boldsymbol{\mu}_1 = \mathrm{E}[\mathbf{X}_{1j}]$, $j = 1, \ldots, n$.
- Sea $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n}$ vectores aleatorios $p \times 1$ de una población normal multivariada después de un tratamiento con $\boldsymbol{\mu}_2 = \mathrm{E}[\mathbf{X}_{2j}]$, $j = 1, \ldots, n$.
- Asuma que $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n}$ y $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n}$ 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 μ

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

$$
\begin{aligned}
X_{1j1} &= \text{variable 1 dentro del tratamiento 1} \\
X_{1j2} &= \text{variable 2 dentro del tratamiento 1} \\
&\vdots \\
X_{1jp} &= \text{variable p dentro del tratamiento 1} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
X_{2j1} &= \text{variable 1 dentro del tratamiento 2} \\
X_{2j2} &= \text{variable 2 dentro del tratamiento 2} \\
&\vdots \\
X_{2jp} &= \text{variable p dentro del tratamiento 2}
\end{aligned}
$$

Inferencia Sobre $\boldsymbol{\mu}$

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

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

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

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad ; \quad \boldsymbol{\mu}_D = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0$$

Inferencia Sobre $\boldsymbol{\mu}$

Se pueden plantear las hipótesis

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\mu}_D = \delta_0 \\ \mathrm{H}_1 &: \boldsymbol{\mu}_D \neq \delta_0 \end{aligned}$$

siendo (en este caso) $\delta_0 = 0$, y cuya **estadistica de prueba** (T^2 de Hotelling) está dada por

$$n \left(\overline{\mathrm{D}} - \delta_0\right)^{\top} \mathbf{S}_D^{-1} \left(\overline{\mathrm{D}} - \delta_0\right) \overset{\mathrm{H}_0\ verd.}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p}$$

donde

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

Pontificia Universidad JAVERIANA Bogotá

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

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

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

$$\vdots$$

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

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

Inferencia Sobre $\boldsymbol{\mu}$

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

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

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

$$\overline{\mathrm{D}}_k\pm\sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{\mathbf{S}_{Dkk}}{n}}$$

donde $\mathbf{S}_{Dkk}$ es la varianza de la $k-$ésima diferencia.

Inferencia Sobre $\boldsymbol{\mu}$

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

$$\overline{\mathrm{D}}_k \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{\mathbf{S}_{Dkk}}{n}}$$

Inferencia Sobre μ

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 $\mathbf{n = 11}$ divisiones de muestra, de los dos laboratorios.

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

Inferencia Sobre μ

Table 6.1 Effluent Data

Sample j	Commercial lab x_{1j1} (BOD)	Commercial lab x_{1j2} (SS)	State lab of hygiene x_{2j1} (BOD)	State lab of hygiene 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 μ

*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 $\boldsymbol{\mu}$

El interés es probar las hipótesis

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \\ \mathrm{H}_1 &: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \end{aligned}$$

donde

- $\boldsymbol{\mu}_1$ es el vector de medias (media de BOD y media SS) del laboratorio comercial.
- $\boldsymbol{\mu}_2$ es el vector de medias (media de BOD y media SS) del laboratorio de higiene del estado de Wisconsin.

Inferencia Sobre μ

```
# -----  Muestras Pareadas ----- #
X1j1<-c(6,6,18,8,11,34,28,71,43,33,20)
X1j2<-c(27,23,64,44,30,75,26,124,54,30,14)
X2j1<-c(25,28,36,35,15,44,42,54,34,29,39)
X2j2<-c(15,13,22,29,31,64,30,64,56,20,21)

X<-data.frame(X1j1,X1j2,X2j1,X2j2)

Dj1<-X$X1j1-X$X2j1
Dj2<-X$X1j2-X$X2j2
Dbarra1<-mean(Dj1); Dbarra2<-mean(Dj2)
Dbarra<-c(Dbarra1,Dbarra2)
```

Inferencia Sobre μ

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

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

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

¿Cuál es la conclusión?

Inferencia Sobre $\boldsymbol{\mu}$

Comparaciones de dos Poblaciones Independientes

- Sea $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ vectores aleatorios $p \times 1$ de una población normal multivariada con $\boldsymbol{\mu}_1 = \mathrm{E}[\mathbf{X}_{1j}]$ y $\boldsymbol{\Sigma}_1$ $j = 1, \ldots, n_1$.
- Sea $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n}$ vectores aleatorios $p \times 1$ de una población normal multivariada con $\boldsymbol{\mu}_2 = \mathrm{E}[\mathbf{X}_{2j}]$ y $\boldsymbol{\Sigma}_2$, $j = 1, \ldots, n_2$.
- Asuma que las poblaciones son independientes y que el interés es probar las hipótesis

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \\ \mathrm{H}_1 &: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0 \end{aligned}$$

Inferencia Sobre $\boldsymbol{\mu}$

- Asumiendo que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (**homocedásticidad/homogeneidad**), se **rechaza** H_0 a un nivel de significancia α si

$$T^2_{obs} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)$$

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

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 $\boldsymbol{\mu}$

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

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

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

Inferencia Sobre $\boldsymbol{\mu}$

Los intervalos simultaneos de T^2

$$\overline{d_i} \pm \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{ii}}$$

- Si las **matrices de covarianzas** de las dos poblaciones son **diferentes (heterocedasticidad/heterogeneidad)**: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. En este caso se rechaza H_0 a un nivel de significancia α si

$$T^2_{obs} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)^\top \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)$$

$$> \chi^2_p$$

Inferencia Sobre $\boldsymbol{\mu}$

Siempre que $n_1 - p \longrightarrow \infty$ y $n_2 - p \longrightarrow \infty$.

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

$$T^{*2} = \left[\overline{X}_1 - \overline{X}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^\top \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} \left[\overline{X}_1 - \overline{X}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]$$

$$\overset{\mathrm{H}_0 \text{ verd.}}{\sim} \frac{vp}{v-p+1} F_{p, v-p+1}$$

donde

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

Rechace H_0 a un nivel de significancia α si

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

Inferencia Sobre μ

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 μ

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

Pontificia Universidad JAVERIANA Bogotá

Inferencia Sobre μ

Por lo tanto,

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 6.6 \\ 1.985 \end{bmatrix} \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} 5.56 \\ 1.824 \end{bmatrix}$$

$$\mathbf{S}_1 = \begin{bmatrix} 1.42 & -0.0504 \\ -0.0504 & 0.005 \end{bmatrix} \qquad \mathbf{S}_2 = \begin{bmatrix} 1.543 & -0.037 \\ -0.037 & 0.0045 \end{bmatrix}$$

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

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

Luego,

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

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

Como

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

Se rechaza $\mathrm{H}_0 : \boldsymbol{\mu}_A = \boldsymbol{\mu}_B$.

Inferencia Sobre $\boldsymbol{\mu}$

Los intervalos simultáneos T^2

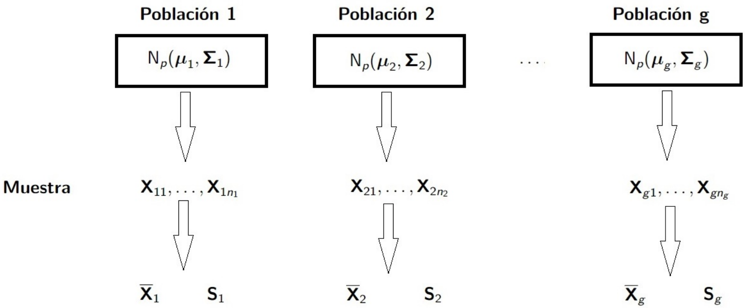
$$6.6 - 5.56 \pm \sqrt{\frac{(9)2}{8} F_{2,8}(0.05)} \sqrt{\left(\frac{1}{6} + \frac{1}{5}\right) 1.47} = (-1.29, 3.37)$$

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

Inferencia Sobre μ

Imagine que se tienen **3 o más grupos** (g), dentro de cada grupo tiene un número de observaciones (n_i), y 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 μ



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

Inferencia Sobre $\boldsymbol{\mu}$

Comparaciones en más de dos Poblaciones Independientes

- $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ una muestra aleatoria de una población normal multivariada con $\boldsymbol{\mu}_1 = \mathrm{E}[\mathbf{X}_{1j}]$ y $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}_{1j}]$ $j = 1, \ldots, n_1$.
- $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ una muestra aleatoria de una población normal multivariada con $\boldsymbol{\mu}_2 = \mathrm{E}[\mathbf{X}_{2j}]$ y $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}_{2j}]$ $j = 1, \ldots, n_g$.
- $\mathbf{X}_{g1}, \ldots, \mathbf{X}_{gn_g}$ una muestra aleatoria de una población normal multivariada con $\boldsymbol{\mu}_g = \mathrm{E}[\mathbf{X}_{gj}]$ y $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}_{gj}]$ $j = 1, \ldots, n_g$.
- Asuma que todas las poblaciones son independientes entre sí y que el interés es probar las hipótesis

$$\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g = \boldsymbol{\mu}$$
$$\mathrm{H}_1 : \text{al menos un } \boldsymbol{\mu}_i \text{ es diferente}$$

Inferencia Sobre $\boldsymbol{\mu}$

Para probar esas hipótesis consideramos la reparametrización:

$$\boldsymbol{\mu}_k = \boldsymbol{\mu} + \boldsymbol{\tau}_k \qquad k = 1, \ldots, g$$

Por lo que se debe probar

$$\begin{aligned} &\mathsf{H}_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \\ &\mathsf{H}_1 : \text{al menos un } \boldsymbol{\tau}_k \text{ es diferente de } \mathbf{0} \end{aligned}$$

Observación: $\boldsymbol{\tau}_k$ *es el efecto del grupo* k.

Inferencia Sobre $\boldsymbol{\mu}$

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

$$\mathbf{X}_{kj} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \boldsymbol{\epsilon}_{kj}$$

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

Supuestos del modelo:

- $\boldsymbol{\epsilon}_{kj} \overset{iid}{\sim} \mathrm{N}\left(\mathbf{0}; \boldsymbol{\Sigma}\right)$
- $\boldsymbol{\mu}$ es la media general.
- $\boldsymbol{\tau}_k$ es el efecto del $k-$ésimo grupo.
- $\sum_{k=1}^{g} n_k \boldsymbol{\tau}_k = \mathbf{0}$ (para garantizar la identificabilidad del modelo)

Inferencia Sobre μ

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

$$T = \sum_{k=1}^{g}\sum_{j=1}^{n_k}\left(\mathbf{X}_{kj} - \overline{\mathbf{X}}\right)\left(\mathbf{X}_{kj} - \overline{\mathbf{X}}\right)^\top$$

$$B = \sum_{k=1}^{g} n_k\left(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}\right)^\top$$

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

Inferencia Sobre μ

Observación: T es la suma de cuadrados total, B es la suma de cuadrados entre los grupos, y W es la suma de cuadrados dentro (intra) de los grupos.

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):

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

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

Inferencia Sobre μ

Example

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

- época: La 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 μ

```
require(heplots)
data("Skulls")            # dataset
require(mvShapiroTest)    # prueba de Norm. Mult.

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

# --- 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")
```