# Estadística Multivariada: Semana 6

Andrés García Medina

CIMAT-MTY
andres.garcia@cimat.mx

Enero-Junio de 2020

1. Comparaciones del vector de medias para dos o más poblaciones.
2. Comparando multiples medias poblacionales
3. Suplemento: Análisis multivariado

# Comparando vector de medias de dos poblaciones

La estadística $T^2$ es apropiada para comparar las respuestas de un conjunto de experimentos con ciertas especificaciones (población 1) con las respuestas de otro conjunto de experimentos con distintas especificaciones (población 2).

considere una muestra aleatoria de dimensión $n_1$ proveniente de la población 1, y una muestra de dimensión $n_2$ de la población 2. Las observaciones sobre $p$ variables se pueden arreglar de la siguiente manera:

| población 1 | estadística suficiente | |
|---|---|---|
| $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ | $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$ | $\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$ |
| $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ | $\bar{\mathbf{x}}_2 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{2j}$ | $\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$ |

Nos interesa realizar inferencia sobre:
(vector promedi de población 1) - (vector promedio de población 2) $= \mu_1$ - $\mu_2$

Por ejemplo, nos gustaría responder las preguntas:
¿Es $\mu_1 = \mu_2$?
Si $\mu_1 - \mu_2 \neq 0$, ¿cuáles componentes son diferentes?

## Suposiciones sobre la estructura de los datos

1. La muestra $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ es una muestra aleatoria de dimensión $n_1$ proveniente de una población $p$-variada con media $\mu_1$ y matriz de covarianza $\mathbf{\Sigma}_1$
2. La muestra $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ es una muestra aleatoria de dimensión $n_2$ proveniente de una población $p$-variada con media $\mu_2$ y matriz de covarianza $\mathbf{\Sigma}_2$
3. $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$, son independientes de $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$.

Además, si las muestras $n_1$ y $n_2$ no son suficientemente grande, se necesita considerar también las siguientes suposiciones:

1. Ambas poblaciones son normales multivariadas.
2. $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$

- Cuando $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma}$, $\sum_{j=1}^{n_1}(\mathbf{x_{1j}} - \bar{\mathbf{x}}_\mathbf{1})(\mathbf{x_{1j}} - \bar{\mathbf{x}}_\mathbf{1})'$ es una estimación de $(n_1 - 1)\mathbf{\Sigma}$, y $\sum_{j=1}^{n_1}(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)'$ es una estimación de $(n_2 - 1)\mathbf{\Sigma}$.

- Por lo que podemos agrupar la información de ambas muestras con el proposito de estimar la matriz de covarianza común $\mathbf{\Sigma}$.

- $\mathbf{S}_{pooled} = \frac{\sum_{j=1}^{n_1}(\mathbf{x_{1j}} - \bar{\mathbf{x}}_1)(\mathbf{x_{1j}} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_1}(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)'}{n_1 + n_2 - 2}$
$= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$

- Puesto que $\sum_{j=1}^{n_1}(\mathbf{x_{1j}} - \bar{\mathbf{x}}_\mathbf{1})(\mathbf{x_{1j}} - \bar{\mathbf{x}}_\mathbf{1})'$ tiene $n_1 - 1$ d.f., y $\sum_{j=1}^{n_1}(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)(\mathbf{x_{2j}} - \bar{\mathbf{x}}_2)'$ tiene $n_2 - 1$ d.f., el divisor de la ec. anterior se obtiene al combinar los grados de libertad de cada componente.

Para probar la hipótesis $\mu_{\mathbf{1}} - \mu_{\mathbf{2}} = \delta_0$ (arbitrario), consideramos la distancia de $\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_{\mathbf{2}}$ a $\delta_0$.

Sabemos que $E(\bar{\mathbf{X}}_{\mathbf{1}} - \bar{\mathbf{X}}_{\mathbf{2}}) = \mu_{\mathbf{1}} - \mu_{\mathbf{2}}$, además por la suposición de independencia se tiene que

$$Cov(\bar{\mathbf{X}}_{\mathbf{1}} - \bar{\mathbf{X}}_{\mathbf{2}}) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{\Sigma} \tag{1}$$

Ahora, dado que $\mathbf{S}_{pooled}$ estima $\mathbf{\Sigma}$, vemos que $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}$

es un estimador de $Cov(\bar{\mathbf{X}}_{\mathbf{1}} - \bar{\mathbf{X}}_{\mathbf{2}})$

Por último, prueba de la razón de verosimilitud de $H_0 : \mu_{\mathbf{1}} - \mu_{\mathbf{2}} = \delta_{\mathbf{0}}$, se basa en el cuadrado de la distancia estadística $T^2$, y esta dada por

$$T^2 = (\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_{\mathbf{2}} - \delta_0)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} (\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_{\mathbf{2}} - \delta_0) > c^2, \tag{2}$$

donde el valor crítico de $c^2$ se determina mediante la distribución de dos muestras del estadístico $T^2$

# Distribución de dos muestras del estadístico $T^2$

## Resultado

*Si* $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ *es una muestra aleatoria de dimensión* $n_1$ *proveniente de* $N_p(\mu_1, \mathbf{\Sigma})$, *y* $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ *es una muestra aleatoria de dimensión* $n_2$ *proveniente de* $N_p(\mu_2, \mathbf{\Sigma})$, *entonces*

$$T^2 = [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)] \quad (3)$$

*se distribuye como* $\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}$

*Por lo que*

$P[\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2) \leq c^2] = 1 - \alpha$, *donde*

$$c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p,n_1+n_2-p-1}(\alpha) \quad (4)$$

# Distribución de dos muestras del estadístico $T^2$

- Nos interesa principalmente la región de confianza para $\mu_1 - \mu_2$.
- Del resultado anterior podemos concluir que todas las $\mu_1 - \mu_2$ dentro de una distancia estadística $c^2$ de $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ constituye una región de confianza.
- Esta región es un elipsoide centrado en la diferencia muestral $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, y cuyos ejes son determinados por los eigenvalores y eigenvectores de $\mathbf{S}_{pooled}$

# Ejercicio 1: Construcción de una región de confianza para la diferencia de dos vectores de medias

Se producen 50 barras de jabón mediante dos procedimientos distintos. Se mide las características $X_1 =$ espuma, y $X_2 =$suavidad. El resumen estadístico para ambos tipos de barras esta dado por:

$$\begin{aligned} \bar{\mathbf{x}}_1 &= \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix} \\ \bar{\mathbf{x}}_2 &= \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \end{aligned} \tag{5}$$

Obtenga la región de confianza de $\mu_1 - \mu_2$ al 95 %.

## Intervalos de confianza simultaneos

### Resultado

*Sea* $c^2 = [(n_1 + n_2 - 2)p/(n_1 + n_2 - p - 1)]F_{p,n_1+n_2-p-1}(\alpha)$ *con probabilidad* $1 - \alpha$.

$$\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm c\sqrt{\mathbf{a}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\mathbf{pooled}}\mathbf{a}} \tag{6}$$

*cubrirá* $\mathbf{a}'(\mu_1 - \mu_2)$ *para toda* $\mathbf{a}$. *En particular,* $\mu_{1i} - \mu_{2i}$ *estará cubierto por:*

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii,pooled}}, \quad para \quad i = 1, 2 \ldots, p \tag{7}$$

# Observación

- Para testear $H_0 : \boldsymbol{\mu_1} - \boldsymbol{\mu_2} = \mathbf{0}$, la combinación lineal $\hat{\mathbf{a}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, con vector de coeficientes $\hat{\mathbf{a}}' \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, mide la diferencia más grande en la población.
- Esto es, si $T^2$ rechaza $H_0$, entonces $\hat{\mathbf{a}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ no tendrá media igual a cero.
- En la práctica, se interpretan los componenetes de la combinación lineal mediante su importancia estadística, así como conocimiento del área.

# Ejercicio 2: Calculando los intervalos de confianza simultaneos para la diferencia en los componentes de la media.

Dos muestras de dimensión $n_1 = 45$, y $n_2 = 55$ se tomaron a partir de propietarios de viviendas con y sin AC. Se utilizaron dos medidas de consumo de electricidad ( kwh). $X_1 =$ consumo *on-peak* durante julio. $X_2 =$consumo total *off-peak* durante julio. El resumen estadístico está dado por:

$$\begin{aligned} \bar{\mathbf{x}}_1 &= \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} \quad n_1 = 45 \\ \bar{\mathbf{x}}_2 &= \begin{bmatrix} 130 \\ 355 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 8632 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix} \quad n_2 = 55 \end{aligned} \tag{8}$$

- Encuentre el intervalo de confianza simultaneo para las diferencias en los componentes de las medias al 95 %.
- Encuentre y dibuje la elipse de confianza para $\mu_1 - \mu_2$ al 95 %.
- Encuentre las intervalos de confianza simultaneos de Bonforreni para la diferencia en la media poblacional.

# Comparación de dos muestras cuando $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

- Cuando $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ no es posible encontrar un medida de distancia del tipo $T^2$, cuya distribución no dependa de $\boldsymbol{\Sigma}_1$ y $\boldsymbol{\Sigma}_2$
- La magnitud de las discrepancias (respecto a la suposición de normalidad) que son críticas en la situación multivariada dependen probablemente, en gran medidad, sobre el número de variables $p$.
- No obstante, para $n_1$ y $n_2$ grandes, se evita la complejidad debido a matrices de covarianza distintas.

# Comparación de dos muestras cuando $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

## Resultado

*Sean las dimensiones de las muestras, de tal manera que $n_1 - p$, y $n_2 - p$ son grandes. Entonces, un elipsoide de confianza del $100(1-\alpha)\,\%$ para $\boldsymbol{\mu_1} - \boldsymbol{\mu_2}$ está dado por todas las $\boldsymbol{\mu_1} - \boldsymbol{\mu_2}$ que satisfacen*

$$[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leq \chi^2_p(\alpha), \quad (9)$$

*donde $\chi^2_p(\alpha)$ es el percentil superior $(100\alpha)$-ésismo de una distribución chi-cuadrada con p grados de libertad. También, el intervalo de confianza simultaneo $100(1-\alpha)\,\%$ para todas las combinaciones lineales $\mathbf{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ se obtienen mediante:*

$$\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \textit{pertenecen a} \quad \mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm \sqrt{\mathbf{a}'\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)} \quad (10)$$

## Observación

*Si,* $n_1 = n_2 = n$, *entonces* $(n-1)/(n+n-2) = 1/2$, *así*
$\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 = \frac{1}{n}(\mathbf{S}_1 + \mathbf{S}_2) = \frac{(n-1)\mathbf{S}_1 + (n-1)\mathbf{S}_2}{n+n-2}\left(\frac{1}{n} + \frac{1}{n}\right) =$
$\mathbf{S}_{pooled}\left(\frac{1}{n} + \frac{1}{n}\right)$

con muestras iguales, el procedimiento para muestras grandes es esencialmente el mismo al procedimiento basado en la matriz de varianza agregada (*pooled*).

# Ejemplo 3: Procedimiento para muestras grandes para inferencias acerca de diferencia en medias

Nos interesa analizar los datos discutidos en el ejemplo 2 utilizando la aproximación para muestras grandes (sobreponga los resultados).

- Encontrar la combinación lineal más crítica que nos lleva al rechazo de $H_0 : \mu_1 - \mu_2 = 0$. Interprete y grafique sus resultados.

## Multiples poblaciones

- Frecuentemente se necesita comparar más de dos poblaciones. En este caso las muestras aleatorias recolectadas para cada una de las $g$ poblaciones se pueden escribir de la siguiente manera:

  Población 1: $\mathbf{X_{11}}, \mathbf{X_{12}}, \ldots, \mathbf{X_{1n_1}}$
  Población 2: $\mathbf{X_{21}}, \mathbf{X_{22}}, \ldots, \mathbf{X_{2n_2}}$
  $\vdots$
  Población g: $\mathbf{X_{g1}}, \mathbf{X_{g2}}, \ldots, \mathbf{X_{gn_g}}$

- MANOVA se utiliza en primera instancia para investigar si los vectores de las medias poblacionales son similares, y en dado caso, cuales son los componentes que difieren.

# Suposiciones

1. $\mathbf{X_{l1}}, \mathbf{X_{l2}}, \ldots, \mathbf{X_{ln_l}}$ es una muestra aleatoria de dimensión $n_l$ de una población con media $\mu_l$, $l = 1, 2, \ldots, g$. Las muestras aleatorias para diferentes poblaciones son independientes.
2. Todas las poblaciones tienen como covarianza común $\mathbf{\Sigma}$.
3. Cada población sigue una normal multivariada.

La condición 3 se puede relajar considerando el teorema del límite central cuando $n_l$ es grande.

Se recomienda repasar la versión univariada (ANOVA).

# Modelo MANOVA para comparar los vectores de medias en $g$ poblaciones.

Un modelo MANOVA se especifica de la siguiente manera:

$$\mathbf{X}_{lj} = \mu + \tau_l + \mathbf{e}_{lj}, \quad j = 1, 2, \ldots, n_l \quad \text{y} \quad l = 1, 2, \ldots, g \tag{11}$$

donde $\mathbf{e}_{lj}$ son variables independientes $N_p(\mathbf{0}, \mathbf{\Sigma})$. El vector $\mu$ representa la media global de las $g$ poblaciones, y $\tau_l$ representa la desviación de la media de la población $l$ respecto a la media global (también llamado tratamiento específico), donde $\sum_{l=1}^{g} n_l \tau_l = 0$.

En analogía al modelo univariado (ANOVA), el vector de observaciones se puede descomponer de la siguiente manera:

$$\underset{\text{(observación)}}{\mathbf{x}_{lj}} = \underset{\text{(media global } \hat{\mu})}{\bar{\mathbf{x}}} + \underset{\text{(efecto de tratamiento } \hat{\tau}_l)}{(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})} + \underset{\text{(residuos } \hat{\mathbf{e}}_{lj})}{(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)}$$

# Residuos

Con un poco de álgebra se obtiene de los residuos la expresión:

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{x})(\mathbf{x}_{lj}-\bar{\mathbf{x}})' = \sum_{l=1}^{g}(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})'+\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)', \quad (12)$$

donde, el primer termino se refiera a la *suma total corregida de los cuadrados y productos cruzados*, el segundo término a la *suma de los cuadrados y productos cruzados entre tratamientos*, y el tercer término a la *suma de cuadrados y productos cruzados entre residuos.*

La *suma de cuadrados y productos cruzados entre residuos* se puede expresar como

$$\mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}} - l)(\mathbf{x}_{lj} - \bar{\mathbf{x}} - l)' = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g, \tag{13}$$

donde $\mathbf{S}_l$ es la matriz de covarianza muestral para la muestra $l$.

Esta matriz es una generalización de $(n_1 + n_2 - 2)\mathbf{S}_{pooled}$

De manera analoga al caso univariado, la hipótesis de no efectos de tratamiento

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = \mathbf{0}. \tag{14}$$

se prueba al considerar la razones relativas entre la suma de tratamientas y la suma de residuos.

De manera equivalente se puede considerar la razón entre la suma de residuos y la suma total.

# Tabla resumen de MANOVA

En la siguiente tabla se resumen los calculos que se necesitan en el test estadístico de MANOVA

| Fuente de variación | Matriz suma de cuadrados y productos cruzados (SSP) | (d.f.) |
|---|---|---|
| Tratamiento | $\mathbf{B} = \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$ | g-1 |
| Resiudos (error) | $\mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)'$ | $\sum_{l=1}^{g} n_l - g$ |
| Total | $\mathbf{B} + \mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}})'$ | $\sum_{l=1}^{g} n_l - 1$ |

donde los grados de libertad (d.f.) se obtienen de la geometría univariada, así como de la teoría multivariada relacionada con la densidad de Wishart.

# Lambda de Wilks

Un test para $H_0 : \tau_1 = dots = \tau_g = \mathbf{0}$ toma en cuenta varianzas generalizadas.

Se rechaza $H_0$ si la razón de varianzas generalizas es pequeña:

$$\lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj}-\bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_{lj}-\bar{\mathbf{x}}_l)'|}{|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj}-\bar{\mathbf{x}})(\bar{\mathbf{x}}_{lj}-\bar{\mathbf{x}})'|}, \tag{15}$$

Esta cantidad se le conoce como lambda de Wiks, la cual estar relacionada con el criterio de la razón de verosimilitud. Además, se puede expresar en términos de los eigenvalores $\hat{\lambda}_1, \ldots, \hat{\lambda}_s$ de $\mathbf{W}^{-1}\mathbf{B}$ como:

$$\lambda^* = \prod_{i=1}^{s}\left(\frac{1}{1+\hat{\lambda}_i}\right), \tag{16}$$

, donde $s = min(p, g-1)$, el rango de $B$.

Otros estadísticos para testear la igualdad de multiples medias multivariadas, como son el *estadístico de Pillai*, el estadístico de *Lawley-Hotelling*, y el *estadístico de Roy*, también pueden ser escritos como funciones particulares de los eigenvalores de $\mathbf{W}^{-1}\mathbf{B}$.

# Distribución de $\lambda^*$

La distribución exacta de $\lambda^*$ se puede derivar para casos especiales:

| p | g | distribución muestral para datos normales multivariados |
|---|---|---|
| $p = 1$ | $g \geq 2$ | $\left(\frac{\sum n_l - g}{g-1}\right)\left(\frac{1-\lambda^*}{\lambda^*}\right) \sim F_{g-1, \sum n_l - g}$ |
| $p = 2$ | $g \geq 2$ | $\left(\frac{\sum n_l - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\lambda^*}}{\sqrt{\lambda^*}}\right) \sim F_{2(g-1), 2(\sum n_l - g - 1)}$ |
| $p \geq 1$ | $g = 2$ | $\left(\frac{\sum n_l - p - 1}{p}\right)\left(\frac{1-\lambda^*}{\lambda^*}\right) \sim F_{p, \sum n_l - p - 1}$ |
| $p \geq 1$ | $g = 3$ | $\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\lambda^*}}{\sqrt{\lambda^*}}\right) \sim F_{2p, 2(\sum n_l - p - 2)}$ |

# Distribución de $\lambda^*$ para muestras grandes.

Para muestras grandes Bartlett (1938) propuso una modificación de $\lambda^*$ para testear $H_0$. Especificamente, ha desmostrado que si $H_0$ es verdad y $\sum n_l = n$ es grande,

$$-\left(n-1-\frac{(p+g)}{2}\right) ln\lambda^* = -\left(n-1-\frac{(p+g)}{2}\right) ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right) \tag{17}$$

se aproxima a una distribución chi-cuadrada con $p(g-1)$ grados de libertad.

De esta manera, para $\sum n_l = n$ grande, rechazamos $H_0$ a nivel de significancia $\alpha$ si

$$-\left(n-1-\frac{(p+g)}{2}\right) ln\left(\frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|}\right) > \chi^2_{p(g-1)}(\alpha), \tag{18}$$

donde $chi^2_{p(g-1)}(\alpha)$ es percentil superior $(100\alpha)$ de la distribución chi-cuadrada con $p(g-1)$ d.f.

# Ejercicio 4: MANOVA y lambda de Wilks para testear igualdad de 3 vectores de medias.

Considere las siguientes muestras independientes:

Población 1: $\begin{bmatrix} 9 \\ 3 \end{bmatrix}, \begin{bmatrix} 6 \\ 2 \end{bmatrix}, \begin{bmatrix} 9 \\ 7 \end{bmatrix}$

Población 2: $\begin{bmatrix} 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

Población 3: $\begin{bmatrix} 3 \\ 8 \end{bmatrix}, \begin{bmatrix} 1 \\ 9 \end{bmatrix}, \begin{bmatrix} 2 \\ 7 \end{bmatrix}$

Pruebe la hipótesis nula $H_0 : \tau_1 = \tau_2 = \tau_3 = \mathbf{0}$ versus $H_1 :$ al menos un $\tau_0 \neq \mathbf{0}$ aplicando la distibución exacta de la lambda de Wilks, así como la aproximación de Bartlett.