Inferencia para la Matriz de Covarianza

Sea $\underline{\boldsymbol{x}_1}, \underline{\boldsymbol{x}_2}, \cdots, \underline{\boldsymbol{x}_n}$ una m.a de una población normal p-variada con vector de medias $\underline{\boldsymbol{\mu}}$-desconocida y matriz de varianzas-covarianzas $\boldsymbol{\Sigma}$-Desconocida, ie. $\underline{\boldsymbol{x}_i} \sim N_p(\underline{\boldsymbol{\mu}}\ ,\ \boldsymbol{\Sigma})$.

Se tiene interés en la siguiente PH:

$$\begin{cases} H_0 \ : \ \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \\ H_a \ : \ \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0 \end{cases}$$

donde, $\boldsymbol{\Sigma}_0$-es una matriz de valores fijos conocida para $\boldsymbol{\Sigma}$ (usualmente planteada por experiencia previa).

Dependiendo de la forma particular de $\boldsymbol{\Sigma}_0$, existen distintos nombres para la PH asociada:

1. $\boldsymbol{\Sigma}_0$-Cualquier matriz de valor fijo. Prueba general.
2. $\boldsymbol{\Sigma}_0 = \Delta$-Diagonal. Prueba de independencia de variables.
3. $\boldsymbol{\Sigma}_0 = \sigma^2 \boldsymbol{I}_p$. Prueba de homocedasticidad e independencia.
4. $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_p$. Prueba de Esfericidad, ie. variables con varianzas unitarias e incorreladas.
5. $\boldsymbol{\Sigma}_0 = \boldsymbol{B}_m + \sigma^2 \boldsymbol{I}_p$, con $\boldsymbol{B}$-de rango $m < p$. Prueba de homocedasticidad e independencia parcial. Con $m = 0$-se tiene la rueba de homocedasticidad e independencia.
6. $\boldsymbol{\Sigma}_0 = \boldsymbol{B}_m + \boldsymbol{I}_p$, con $\boldsymbol{B}$-de rango $m < p$. Prueba de Esfericidad parcial. Con $m = 0$-se tiene la Prueba de Esfericidad.

La estadística de Razón de Verosimilitud para la PH considerada está dado por:

$$\lambda := \frac{\underset{\underline{\mu}}{\text{Máx}} \; L(\underline{\mu}\,,\; \boldsymbol{\Sigma}_0)}{\underset{\underline{\mu}\,,\,\boldsymbol{\Sigma}}{\text{Máx}} \; L(\underline{\mu}\,,\; \boldsymbol{\Sigma})} = \frac{\text{Máximo de L-Restingida}}{\text{Máximo de L-No Restingida}}$$

$$= \frac{\text{Máximo de L-Bajo } H_0\text{-Cierta}}{\text{Máximo de L-General}}$$

Luego de realizar los cálculos y simplificaciones necesarias se obtiene que:

$$\lambda = \left[\left(\frac{n-1}{n}\right)^{p} \frac{|\boldsymbol{S}|}{|\boldsymbol{\Sigma}_0|}\right]^{\frac{n}{2}} \text{Exp}\left\{-\frac{1}{2}\left[(n-1)\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - np\right]\right\}$$

haciendo: $\frac{n-1}{n} \approx 1$, ie. $n-1 \approx n = v$, ie. $v = n-1 = n$, se tiene:

$$\lambda = \frac{|\boldsymbol{S}|^{\frac{v}{2}}}{|\boldsymbol{\Sigma}_0|^{\frac{v}{2}}} \text{Exp}\left\{-\frac{1}{2}\left[v\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - vp\right]\right\}$$

y haciendo $\lambda^\star = -2\log\lambda$, se tiene que:

$$\lambda^\star = v\left[\text{Log}|\boldsymbol{\Sigma}_0| - \text{Log}|\boldsymbol{S}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - p\right]$$

Bajo H_0-cierta, se tiene que:

$$\lambda^\star \sim \chi^2_k\ , \quad \textit{para} \quad n-1 \ \text{ grande}$$

con $k = p + \frac{p(p+1)}{2} - p = \frac{p(p+1)}{2}$, ie.

k=Parámetros en Θ menos parámetros en Θ_0 (ie. menos p).

Rechazamos H_0 si.

$$\lambda^\star > \chi^2_{\alpha\ ;\ k}$$

Ahora, otra forma alterna de $\lambda^\star$ es como sigue:

Si $\lambda_1, \lambda_2, \ldots, \lambda_p$ son los valores propios de $\boldsymbol{S\Sigma_0^{-1}}$, entonces se sabe que:

$$\text{tr}\left(\boldsymbol{S\Sigma_0^{-1}}\right) = \sum_{i=1}^{p} \lambda_i$$

y usando propiedades de determinantes se tiene que:

$$\text{Log}|\boldsymbol{\Sigma}_0| - \text{Log}|\boldsymbol{S}| = -\text{Log}|\boldsymbol{S\Sigma_0^{-1}}| = -\text{Log}\left(\prod_{i=1}^{p} \lambda_i\right),$$

luego,

$$\lambda^\star = v\left[\text{Log}|\boldsymbol{\Sigma}_0| - \text{Log}|\boldsymbol{S}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - p\right]$$

$$= v\left[-\text{Log}\left(\prod_{i=1}^{p}\lambda_i\right) + \sum_{i=1}^{p}\lambda_i - p\right]$$

$$\lambda^\star = v\left[\sum_{i=1}^{p}[\lambda_i - \text{Log}\lambda_i] - p\right] \sim \chi^2_k$$

y rechazamos H_0 si $\lambda^\star > \chi^2_{\alpha\ ;\ k}$

con $k = \frac{p(p+1)}{2}$.

Una Modificación para $\lambda^\star$-fue propuesta por Bartlet, (**para el caso de muestras pequeñas**) la cual es:

$$\lambda_1^\star = \left\{1 - \frac{1}{6(n-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}\lambda^\star \sim \chi_k^2$$

es decir,

$$\lambda_1^\star = c\lambda^\star \sim \chi_k^2$$

con

$$c = 1 - \frac{1}{6(n-1)}\left[2p+1-\frac{2}{p+1}\right]$$

que puede usarse para tamaños de muestras moderadamente pequeños.

EJEMPLO: Se tomaron 20 sujetos y se les midió los tiempos de reacción ante un estímulo en centésimas de segundo. A cada individuo se le midieron estos tiempos en 3 intervalos de tiempos distintos. Se asume que estas mediciones tienen una distribución $N_3(\underline{\boldsymbol{\mu}}\ ,\ \boldsymbol{\Sigma})$. Pruebe la hipótesis:

$$H_0\ :\ \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 10 \end{bmatrix} \quad v.s \quad H_a\ :\ \boldsymbol{\Sigma} \neq \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 10 \end{bmatrix}$$

Solución: Como

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 10 \end{bmatrix}\ ,\ \text{luego}\quad \boldsymbol{\Sigma}_0^{-1} = \begin{bmatrix} 0.41 & -0.23 & 0.03 \\ -0.23 & 0.42 & -0.16 \\ 0.03 & -0.16 & 0.17 \end{bmatrix}$$

$$\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1} = \begin{bmatrix} 0.85 & -0.01 & 0.03 \\ -0.58 & 1.68 & -0.08 \\ -0.41 & 0.72 & 0.68 \end{bmatrix}.$$

Los valores propios de $\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}$ son: $\lambda_1 = 1.61$, $\lambda_2 = 0.87$ y $\lambda_3 = 0.73$, luego: $\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) = \sum \lambda_i = 3.2216$, $|\boldsymbol{\Sigma}_0| = 86$, $|\boldsymbol{S}| = 88.6355$, de donde:

$$\lambda^\star = v\left[\text{Log}|\boldsymbol{\Sigma}_0| - \text{Log}|\boldsymbol{S}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - p\right]$$

$$= 20\left[\text{Log}(86) - \text{Log}(88.6355) + 3.2216 - 3\right]$$

$$\lambda^\star = 3.63$$

$$\lambda_1^\star = \left\{1 - \frac{1}{6(n-1)}\left[2p + 1 - \frac{2}{p+1}\right]\right\}\lambda^\star$$

$$= \left\{1 - \frac{1}{6(19)}\left[2(3) + 1 - \frac{2}{3+1}\right]\right\}3.63$$

$$\lambda_1^\star = 3.42$$

Para $\alpha = 0.05$, se tiene que: $\chi_{\alpha;v} = \chi_{0.05;6} = 12.592$.

En este caso, $v = p(p+1)/2 = 3(4)/2 = 6$, y como:

$$\lambda_1^\star = 3.42 < 12.592 = \chi^2_{\alpha;v},$$

,

entonces, no se rechaza H_0 y se concluye que la evidencia muestral apoya la hipótesis:

$$H_0 \ : \ \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 10 \end{bmatrix},$$

a un nivel de significancia del 5%.

Ejemplo Usando R. Para n-grande

Se desea contrstar las hipótesis:

$$H_0 \ : \ \boldsymbol{\Sigma} = \begin{bmatrix} 6.5 & -0.05 & -2.5 \\ -0.05 & 5.5 & 0.5 \\ -2.5 & 0.5 & 7.5 \end{bmatrix}$$

a un nivel de significancia del 5%.

Resultados usando la función de usuario: `sigma_sigma0_ngrande`

```
Sigma_0<-matrix(c(6.5, -0.05, -2.5,-0.05, 5.5, 0.5, -2.5 ,0.5, 7.5),byrow=TRUE,
res_sigma0<-sigma_sigma0_ngrande(grupo1[,1:3],Sigma_0,0.05)
kable(res_sigma0)
```

Lamda_est	df	Chi_Tabla	Valor_P
21.7799	6	12.5916	0.00132724

Ejemplo Usando R. Para n-pequeña (Modificación de Bartlet)

Resultados usando la función de usuario: `sigma_sigma0_npqna`

```
Sigma_0<-matrix(c(6.5, -0.05, -2.5,-0.05, 5.5, 0.5, -2.5 ,0.5, 7.5),byrow=TRUE,
res_sigma0npqna<-sigma_sigma0_npqna(grupo1[,1:3],Sigma_0,0.05)
kable(res_sigma0npqna)
```

Lamda1_est	c	df	Chi_Tabla	Valor_P
20.872	0.95833	6	12.592	0.001934

Dos o más Matrices de Covarianzas

Recordar que uno de los supuestos cuando se comparan dos o más vectores de medias, es que las respectivas matrices de Var-Cov asociadas a cada población diferente, sean iguales. Un test muy común para probar la igualdad de matrices de Var-Cov es el M-Test de Box.

Suponga que se tienen g-poblaciones diferentes, con matrices de Var-Cov asociadas dadas por: $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_g$, respectivamente. Se desea contrastar las hipótesis:

$$\begin{cases} H_0 \ : \ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma} \\ H_a \ : \ \mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j \quad p.a \quad i, j \end{cases}$$

donde $\mathbf{\Sigma}$-es una matriz de Var-Cov común.

Si se tienen g-muestras aleatorias, una para cada población, de tamaños $n_1, n_2, \ldots, n_g$, respectivamente y además se asume que cada muestra proviene de una población con distribución normal p-variada, entonces el estadístico de prueba de razón de verosimilitud para esta PH es:

$$\lambda = \prod_{i=1}^{g} \left(\frac{|\boldsymbol{S}_i|}{|\boldsymbol{S}_p|} \right)^{\frac{n_i - 1}{2}},$$

donde, $\boldsymbol{S}_i$-es la matriz de Var-Cov muestral asosciada a la m.a de la i-ésima población, $i = 1, 2, \ldots, g$ y

$$\boldsymbol{S}_p = \frac{1}{\sum_{i=1}^{g}(n_i - 1)} \left[\sum_{i=1}^{g} (n_i - 1)\boldsymbol{S}_i \right]$$

haciendo, $\{v_i = n_i - 1\}$ y $\{v = \sum_{i=1}^{g} v_i = \sum_{i=1}^{g}(n_i - 1)\}$, se tiene que:

$$\boldsymbol{S}_p = \frac{1}{v}\left[\sum_{i=1}^{g} v_i \boldsymbol{S}_i\right]$$

La **Estadística M de Box** se define como:
$M = -2\text{Log}\lambda \sim \chi^2$, (n-grande).

$$M = \left[\sum_{i=1}^{g}(n_i - 1)\right]\text{Log}|\boldsymbol{S}_p| - \sum_{i=1}^{g}(n_i - 1)\text{Log}|\boldsymbol{S}_i|$$

$$= v\text{Log}|\boldsymbol{S}_p| - \sum_{i=1}^{g} v_i \text{Log}|\boldsymbol{S}_i|$$

Bajo H_0-cierto, se espera que las matrices de Var-Cov muestrales no sean muy diferentes, en cuyo caso, el valor de λ estaría cerca a uno y por lo tanto M-sería pequeño.

Ahora, sea

EJEMPLO: En el departamento de salud y servicios sociales de Wisconsin se realizó un estudio para investigar el efecto de la propiedad o la certificación (o ambas) sobre los costos. Cuatro costos fueron seleccionados para el análisis; estos fueron calculados diariamente por paciente y fueron medidos en horas por paciente diario. Las variables fueron: X_1-Costo de la enfermería, X_2-Costo de alimentación, X_4-Costo de operación y mantenimiento y X_4-Costo de administración y lavandería. Se registraron un total de $n = 516$-observaciones de cada una de las cuatro variables, separadas previamente en tres grupos de interés: Privados ($n_1 = 271$), Públicos ($n_2 = 138$) y Gubernamentales ($n_3 = 107$). Se asume que el vector $\underline{\boldsymbol{x}} = (X_1, X_2, X_3, X_4)^t$ tiene una distribución $N_4(\underline{\boldsymbol{\mu}_i}\ ,\ \boldsymbol{\Sigma}_i)$, para $i = 1, 2, 3$. Se desea probar la hipótesis:

$$\begin{cases} H_0 \ : \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma} \\ H_a \ : \ \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j \quad p.a \quad i \neq j = 1, 2, 3 \end{cases}$$

$i = 1$ Privados

$$n_1 = 271\ , \quad \bar{\mathbf{x}}_1 = \begin{pmatrix} 2.066 \\ 0.480 \\ 0.082 \\ 0.36 \end{pmatrix} \quad , \quad S_1 = \begin{pmatrix} 0.291 & & & \\ -0.01 & 0.011 & & \\ 0.02 & 0.000 & 0.001 & \\ 0.010 & 0.003 & 0.000 & 0.10 \end{pmatrix}$$

$i = 2$ Públicos

$$n_2 = 138\ , \quad \bar{\mathbf{x}}_1 = \begin{pmatrix} 2.167 \\ 0.596 \\ 0.124 \\ 0.418 \end{pmatrix} \quad , \quad S_2 = \begin{pmatrix} 0.561 & & & \\ 0.011 & 0.025 & & \\ 0.001 & 0.004 & 0.005 & \\ 0.037 & 0.007 & 0.002 & 0.019 \end{pmatrix}$$

$i = 3$ Gubernamentales

$$n_3 = 107\ , \quad \bar{\mathbf{x}}_1 = \begin{pmatrix} 2.273 \\ 0.521 \\ 0.125 \\ 0.383 \end{pmatrix} \quad , \quad S_3 = \begin{pmatrix} 0.261 & & & \\ 0.030 & 0.017 & & \\ 0.003 & 0.000 & 0.004 & \\ 0.018 & 0.006 & 0.001 & 0.013 \end{pmatrix}$$

Con la información anterior se obtiene:

$$|\boldsymbol{S}_1| = 2.783 \times 10^{-8}\ , \qquad Log|\boldsymbol{S}_1| = -17.397$$

$$|\boldsymbol{S}_2| = 89.539 \times 10^{-8}\ , \qquad Log|\boldsymbol{S}_2| = -13.926$$

$$|\boldsymbol{S}_3| = 14.579 \times 10^{-8}\ , \qquad Log|\boldsymbol{S}_3| = -15.741$$

$$|\boldsymbol{S}_p| = 17.398 \times 10^{-8}\ , \qquad Log|\boldsymbol{S}_1| = -15.564$$

$$u = \left[\frac{1}{270} + \frac{1}{137} + \frac{1}{106} - \frac{1}{270 + 137 + 106}\right]\left[\frac{2(4)^2 + 3(4) - 1}{6(4+1)(3-1)}\right] = 0.0133$$

$$= (270 + 137 + 106)\,(-15.564) - [270(-17.397) + 137(-13.926) + 106(-15.741)] = 289.3$$

$$C = (1 - u)M = (1 - 0.0133)289.3 = 285.5$$

Ahora, para $\alpha = 0.05$, se tiene que:

$$\chi^2_{\alpha;k} = \chi^2_{0.05\ ;\ \frac{p(p+1)}{2}(g-1)} = \chi^2_{0.05\ ;\ 20} = 31.34,$$

Como $C > \chi^2_{\alpha;k}$, entonces se rechaza H_0 y se concluye que las matrices de Var-Cov asociadas a los vectores de costos para las las tres poblaciones consideradas son diferentes a un nivel de significancia del 5%.

Ejemplo Usando R. Prueba M de Box (n-grande)

Resultados de esta PH usando la función de usuario:

```
prueba_M_Box2()
```

```
x<-grupo1[,1:3]
y<-grupo2[,1:3]
res_box<-prueba_M_Box2(x,y,0.05)
kable(res_box)
```

M	U	C	df	Chi_Tabla	Valor_p
4.16033	0.0613499	3.9051	6	12.5916	0.689518

Resultados de esta PH utilizando la función `boxM` del paquete `biotools` del R.

```
mbox<-boxM(datos[1:55,1:3],datos$Grupos[1:55])
mbox
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos[1:55, 1:3]
## Chi-Sq (approx.) = 3.9051, df = 6, p-value = 0.6895
```

Estadístico$= C \sim \chi^2$	gl	p-Valor
3.9051	6	0.68952