TP Análisis Multivariado I: Test para normal multivariada

Fecha de Entrega: 4 de Octubre de 2021

Existen tres librerías de `R` para testear si un conjunto de observaciones $\mathbf{x}_1, \ldots, \mathbf{x}_n$, que se suponen independientes e idénticamente distribuídas provienen de una distribución $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a saber las librerías

- `mvShapiroTest` a través de la función `mvShapiro.Test` que se aplica a la matriz de observaciones $\mathbf{X}$ y que fue propuesto por Villaseñor-Alva & Gonzalez-Estrada (2009)
- `mvnormtest` desarrollada por Slawomir Jarek, a través de la función `mshapiro.test` que se aplica a $\mathbf{U} = \mathbf{X}^{\mathrm{T}}$
- `mvnormalTest` que posee varias funciones para evaluar alejamiento de la normalidad entre los que podemos mencionar las funciones
  - `mardia` que calcula las medidas de asimetría y curtosis de Mardia (1970) y se aplica a la matriz $\mathbf{X}$ y devuelve además el test de Shapiro-Wilks para cada coordenada de las observaciones.
  - `mvnTest` que se aplica a la matriz $\mathbf{X}$ y se basa en el procedimiento dado en Zhou and Shao (2014) y aproxima la distribución del estadístico por simulación, se recomienda tomar $B = 1000$. Además de devolver el estadístico multivariado y su pvalor en \$mv.test, devuelve los estadísticos de Shapiro-Wilks univariados.

donde $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix}$

El objetivo de este ejercicio es comparar para distintos tamaños de muestra el nivel y potencia de los distintos procedimientos.

Para ello, en cada caso deberá realizar $NR = 1000$ replicaciones generando muestras de tamaño $n = 20, 30, 50, 100, 200$ en dimensión $p = 2, 5, 10, 20$ según combinaciones indicadas con $\star$ en la Tabla 1

	n				
p	20	30	50	100	200
2	$\star$	$\star$	$\star$	$\star$	$\star$
5	$\star$	$\star$	$\star$	$\star$	$\star$
10			$\star$	$\star$	$\star$
20				$\star$	$\star$

Table 1: Combinaciones de (n, p) a evaluar.

Para que los resultados sean reproducibles se deberá fijar la semilla en cada replicación. Por otra parte, en cada replicación deberá armar un vector para guardar los resultados en un archivo que identifique las condiciones y que despues se levantarán para procesar. Por ejemplo,

```
n <- 100
p <- 20
#############################################################################################
# puede darle nombres a la hipotesis nula y alternativas para su identificacin, por ejemplo
# 'H0' = Hipotesis nula
# 'Ha' = la alternativa dada en a)
# paste('Hb',epsilon,sep="-") las alternativas dada en b) de acuerdo al valor de epsilon
# 'Hc' = la alternativa dada en c)
#############################################################################################

alternativa <- 'H0'

nombre.archivo <- paste('pvalores-n-',n,'-p-',p,'-',alternativa,'.txt', sep="")

pvalor.Shapiro.alva <- pvalor.Shapiro.Jarek <- pvalor.mardia<- pvalor.Zhou<- rep(NA,length=NR)

for (irep in 1:NR){
set.seed(1234+irep)
.....

pvalor.Shapiro.alva[irep] <- ***$p.value
pvalor.Shapiro.Jarek[irep] <- ***$p.value
pvalor.mardia[irep] <- ***$p.value
pvalor.Zhou[irep] <- ***$p.value

to.write <- c(irep,n,p,pvalor.Shapiro.alva[irep],pvalor.Shapiro.Jarek[irep],
            pvalor.mardia[irep],pvalor.Zhou[irep])

 write(t(to.write), file=nombre.archivo, ncolumns= length(to.write), append = TRUE)

 }
```

De esta forma para cada alternativa y para la hipótesis nula y para cada combinación de (n, p), se guardarán los $p-$valores obtenidos para cada procedimiento y test en un archivo que corresponde a una matriz de 7×NR. Luego, se calculará la proporción de rechazos para un nivel α para cada procedimiento que se reportará en una tabla o en un gráfico, identificando por ejemplo el procedimiento en diferentes colores o símbolos, .

1. En primer lugar, para evaluar el comportamiento en nivel considere observaciones generadas de una distribución $N(\mathbf{0}_p, \mathbf{I}_p)$. Considere nivel $\alpha = 0.05$. Estudie si hay diferencias.

2. Para evaluar potencia, fije el nivel en $\alpha = 0.05$ y estudie el comportamiento para las siguientes distribuciones

   a) Genere las observaciones $\mathbf{x}_i$ de modo que tengan la misma distribución que el vector $\mathbf{x}$, donde $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ se define como

   $$\begin{aligned} x_j &= y_j v_{j+1} \qquad 1 \leq j \leq p-1 \\ x_p &= y_p v_1 \end{aligned}$$

   donde $y_1, \ldots, y_p$ son i.i.d. $N(0, 1)$ y $v_j = \text{signo}(y_j)$, es decir, $v_j = 1$ si $y_j > 0$, $v_j = -1$ si $y_j < 0$ y $v_j = 0$ si $y_j = 0$.

   Esta distribución fue introducida por Dutta & Genton (2014) y tiene la propiedad que para cualquier conjunto de índices $(j_1, \ldots, j_{p-1}$ el vector de dimensión $p-1$ $(x_{j_1}, \ldots, x_{j_{p-1}})^{\mathrm{T}}$ tiene distribución $N(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})$ pero $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ no es $N(\mathbf{0}, \mathbf{I}_p)$ pues

   $$\mathbb{P}\left(\prod_{j=1}^{p} x_j > 0\right) = \mathbb{P}\left(\prod_{j=1}^{p-1} y_j v_{j+1} \times y_p v_1 > 0\right) = \mathbb{P}\left(\prod_{j=1}^{p} y_j v_j > 0\right) = 1\,,$$

   es decir, $\mathbf{x}$ tiene soporte en el cuadrante $\{\mathbf{u} \in \mathbb{R}^p : u_1 \times u_2 \times \ldots u_p > 0\}$.

   b) Genere las observaciones como $\mathbf{x}_i = (1 - B_i)\mathbf{z}_i + B_i \mathbf{y}_i$ donde $B_i, \mathbf{y}_i$ y $\mathbf{z}_i$ son independientes $B_i \sim Bi(1, \epsilon)$, $\mathbf{z}_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$ y $\mathbf{y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Considere los siguientes valores para ϵ, $\epsilon = 0.10, 0.25, 0.5$ y tome $\boldsymbol{\mu} = 2\mathbf{1}_p$ y $\boldsymbol{\Sigma} = 0.5\mathbf{I}_p$ y $\boldsymbol{\Sigma} = 0.7\mathbf{I}_p + 0.3\mathbf{1}_p\,\mathbf{1}_p^{\mathrm{T}}$, donde $\mathbf{1}_p$ el vactor de $\mathbb{R}^p$ con todas sus componentes iguales a 1.

   c) Genere las observaciones tales que $\mathbf{x}_i \sim \mathcal{T}_{p,3}$, es decir,

   $$\mathbf{x}_i = \mathbf{z}_i \sqrt{\frac{3}{V_i}}$$

   donde $\mathbf{z}_1, \ldots, \mathbf{z}_n, V_1, \ldots, V_n$ son independientes, $\mathbf{z}_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$ y $V_i \sim \chi^2_k$. Puede usar si prefiere la función `rmvt` de la librería `mvtnorm`

Referencias

Dutta, S. & Genton, M. (2014). A non-Gaussian multivariate distribution with all lower-dimensional Gaussians and related families, *Journal of Multivariate Analysis*, **132**, 82-93.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519-530.

Villaseñor-Alva, J.A. & Gonzalez-Estrada, E. (2009). A generalization of Shapiro-Wilk's test for multivariate normality. *Communications in Statistics: Theory and Methods*,**38**,1870-1883.

Zhou, M., & Shao, Y. (2014). A powerful test for multivariate normality. *Journal of Applied Statistics*, **41**, 351-363.