1 Análisis Multivariado I - Práctica 1 - Parte 2 (en R)

1.1 Outliers

1. Implemente en R una función que calcule el *medcouple* para una muestra $x = \{x_1, \ldots, x_n\}$ de una distribución continua univariada $\mathcal{F}$. Brys *et. al.* (2003, *A Comparison of Some New Measures of Skewness*) definen el *medcouple* como sigue.

   Si $x_1 \leq x_2 \leq \ldots \leq x_n$ y m_n es la mediana, entonces

   $$\mathrm{MC}_n = \underset{x_i \leq m_n \leq x_j}{\text{mediana}}\, h\left(x_i, x_j\right)$$

   donde la función h se define para los valores $x_i \neq x_j$ como

   $$h\left(x_i, x_j\right) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}$$

   y como

   $$h\left(x_{s_i}, x_{s_j}\right) = \begin{cases} -1 & i + j - 1 < k \\ 0 & i + j - 1 = k \\ +1 & i + j - 1 > k \end{cases}$$

   para los casos en que $x_i = x_j = m_n$, en donde $s_1 < \cdots < s_k$ son los índices de las observaciones iguales a la mediana, es decir, $x_{s_j} = m_n$ para todo $j = 1, \ldots, k$.

2. Con la función del ejercicio anterior:

   (a) Calcule el MC de los datos correspondientes a la velocidad del viento, V, en el dataset `airquality` y compararlo con el obtenido con `mc()` de la librería `robustbase`.

   (b) Construya el boxplot clásico y el boxplot ajustado usando `adjbox`. En ambos casos, identifique cuáles son los datos detectados como outliers y compare.

   (c) Repita (a) y (b) para el dataset `los` de la librería `robustbase` y para los datos de ozono, O, del dataset `airquality`. En este último caso, elimine las observaciones *missing*.

   (d) Qué observa en los casos anteriores? En todos los casos realice gráfico del estimador de la densidad. Qué reflexiones puede aportar?

   (e) Realice el bagplot de (V, O) para la submuestra sin observaciones faltantes con distintos factores de expansión: 1.5, 2 y 3. Qué observa?

   Para facilitar las comparaciones nombre a cada observación con su índice utilizando ya sea la instrucción `row.names()` *o la instrucción* `names()`

3. Genere una muestra $X_1, \ldots, X_n$ donde $n = 50, 100$ y 200, con distribuión G, donde G es una de las siguientes distribuciones

   - $G = N(2, 25)$
   - $G \sim \chi^2_2$
   - $G = 0.9N(0, 25) + 0.1N(\mu, 0.01)$ donde μ toma los valores $5, 10, 15, 20$

- $G = \Gamma(3, 1)$
- $X_i = \log(V_i)$ donde $V_i \sim \Gamma(3, 1)$

En todos los casos, fije la semilla para poder reproducir los resultados.

(a) Calcule el MC.

(b) Utilizando el boxplot clásico y el ajustado identifique cuáles son los datos detectados como outliers y compare.

(c) Qué puede decir en el caso $G = 0.9N(0, 25) + 0.1N(\mu, 0.01)$? Coinciden con el 10% de datos generados como $N(\mu, 0.01)$.

Para el caso (c), guarde en un vector llamado `outs` *el índice de los datos generados como* $N(\mu, 0.01)$ *y comparelos con las observaciones detectadas por los boxplots.*

4. Simule tres muestras de una normal asimétrica $Z \sim \mathcal{SN}(\lambda)$ para diferentes valores de $\lambda \in \mathbb{R}$ que considere de interés. Tener en cuenta que si Y_0 e Y_1 son variables $N(0, 1)$ independientes y $\delta \in (-1, 1)$ entonces

$$Z = \delta |Y_0| + \sqrt{(1 - \delta^2)} Y_1 \sim \mathcal{SN}(\lambda(\delta))$$

con $\lambda = \delta/\sqrt{1 - \delta^2}$.

En todos los casos, fije la semilla para poder reproducir los resultados. Realice un gráfico adecuado para verificar si se observa la asimetría.

(a) Calcule el MC.

(b) Indique si el boxplot clásico detecta alguna observación como outlier.

(c) Repita (b) con el boxplot ajustado. Qué observa?

5. Genere una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ para $n = 50, 100$ y 200, $\mathbf{x}_i \in \mathbb{R}^2$ tales que $\mathbf{x}_i \sim \mathbf{x}$ donde $\mathbf{x} = (X_1, X_2)^{\mathrm{T}}$, X_1 es independiente de X_2, $X_1 \sim N(2, 1)$, $X_2 \sim \mathcal{SN}(\lambda(\delta))$ con $\delta = 0.98$. Fije la semilla para poder reproducir los resultados.

Ayuda: Utilice la librería `sn` *y la función* `rsn`

(a) Para cada coordenada de la muestra, identifique los datos que son detectados como outliers por el boxplot clásico y el boxplot ajustado, respectivamente. Qué observa?

(b) Qué observaciones son identificadas como outliers mediante el bagplot?

(c) Repita (a) y (b) cuando $X_1 \sim 0.9N(2, 1) + 0.1N(5, 0.01)$ y $X_2 \sim \mathcal{SN}(\lambda(\delta))$ con $\delta = 0.98$.

Para facilitar la comparación nombre a cada observación con su índice utilizando la instrucción `row.names()`

6. Sea $\mathbf{x} = (X_1, X_2)^{\mathrm{T}} \sim N_2(\mathbf{0}, \mathbf{\Psi})$ con $\mathbf{\Psi}$ la matriz de correlación asociada a $\mathbf{\Sigma} = \begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$ y $X_0 \sim N(0, 1)$ independiente de $\mathbf{x}$ tal que con $\delta_j \in (-1, 1), j = 1, 2$. Definamos $Z_j = \delta_j |X_0| + \sqrt{\left(1 - \delta_j^2\right)} X_j$, entonces $Z_j \sim \mathcal{SN}(\lambda(\delta_j))$.

El vector $\mathbf{z} = (Z_1, Z_2)^{\mathrm{T}}$ es un vector normal multivariado asimétrico, $\mathbf{z} \sim \mathcal{SN}_2(\boldsymbol{\alpha}, \mathbf{\Omega})$, con densidad conjunta $f_2(\mathbf{z}) = 2\phi_2(\mathbf{z}; \mathbf{\Omega})\Phi\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}\right)$ donde:

- $\phi_2(\mathbf{z};\boldsymbol{\Omega})$ es la densidad de $N_2(\mathbf{0},\boldsymbol{\Omega})$,
- $\boldsymbol{\alpha}^{\mathrm{T}} = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Delta}^{-1}}{\sqrt{1+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}}}$, $\boldsymbol{\lambda} = (\lambda(\delta_1), \lambda(\delta_2))^{\mathrm{T}}$,
- $\boldsymbol{\Delta} = \mathrm{DIAG}\left(\sqrt{1-\delta_1^2}, \sqrt{1-\delta_2^2}\right)$ y
- $\boldsymbol{\Omega} = \boldsymbol{\Delta}\left(\boldsymbol{\Psi} + \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}\right)\boldsymbol{\Delta}$.

(a) Genere muestras normales asimétricas univariadas con $\delta_1 = 0.99$ y $\delta_2 = -0.5$. Realice gráficos adecuados para verificar la asimetría de $Z_1 \sim \mathcal{SN}(\lambda(\delta_1))$, $Z_2 \sim \mathcal{SN}(\lambda(\delta_2))$ y $\mathbf{z} \sim \mathcal{SN}_2(\boldsymbol{\alpha},\boldsymbol{\Omega})$.

Ayuda: Utilice la librería `sn` *y la función* `rmsn`

(b) Realice un gráfico de la densidad y un mapa de contorno para $\mathbf{z} \sim \mathcal{SN}(\boldsymbol{\alpha},\boldsymbol{\Omega})$ y compare con el de $\mathbf{z}^\star \sim N_2(\mathbf{0},\mathbf{I}_2)$.

(c) Realice un bagplot para las muestras de $\mathbf{z}$ y $\mathbf{x}$ y verifique si las observaciones atípicas detectadas con el bagplot se corresponden con observaciones atípicas detectadas con los boxplots unidimensionales de X_1, X_2, Z_1, Z_2.

Teniendo en cuenta la distribución de los datos, Ud. diría que esas observaciones son realmente outliers?

(d) Para las observaciones en la muestra de $\mathbf{x}$, calcule el cuadrado de la distancia de Mahalanobis:

i. usando los parámetros poblacionales y determinar cuáles se encuentran fuera de la región determinada por el punto de corte $\chi^2_{2,0.05}$.

ii. repetir (i) usando los estimadores clásicos de los parámetros poblacionales.

iii. repetir (i) usando como estimador robusto de los parámetros poblacionales

- el estimador de Donoho-Stahel.

  Utilice la librería `rrcov` *y la función* `CovSde`*. Puede ser que necesite recuperar ciertos resultados usando* `@` *en lugar de* `$`*.*

- el $S-$estimador.

  Utilice la librería `rrcov` *y la función* `CovSest`*.*

(e) Genere una muestra con la misma distribución que el vector aleatorio $\mathbf{y} \sim 0.9F_2 + 0.1H_2$ donde $F_2 = \mathcal{SN}_2(\boldsymbol{\alpha},\boldsymbol{\Omega})$ con $\boldsymbol{\alpha} = (10,4)^{\mathrm{T}}$, $\boldsymbol{\Omega} = \mathbf{I}_2$ y H_2 es la distribución $N(\boldsymbol{\mu}, 0.25\mathbf{I}_2)$ con $\boldsymbol{\mu} = (-1,-1)^{\mathrm{T}}$.

i. Grafique su densidad y el mapa de contorno.

ii. Realice los contornos de profundidad usando las medidas de atipicidad

- de Donoho-Stahel (SDO),
- ajustada (AO) y
- direccional (DO).

Utilice la librería `mrfDepth` *para su cálculo.*
Qué medida elegiría para detectar posibles datos atípicos para una muestra de este tipo?