Análisis de la Normalidad Multivariante. Identificación de Outliers.

1. Análisis de la Normalidad Multivariante

Dado un conjunto de datos **X** (una muestra de tamaño *n* de un vector aleatorio *p*-dimensional), se trata de investigar si la distribución poblacional es normal *p*-variante.

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \underline{X}_1^t \\ \vdots \\ \underline{X}_n^t \end{bmatrix}$$

No existen procedimientos de bondad de ajuste tan satisfactorios como en el caso univariante. Uno de los motivos es que las frecuencias esperadas de regiones multivariantes serán en general pequeñas a no ser que *n* sea muy elevado (curso de la dimensionalidad).
A cambio se han propuesto diversas ideas y procedimientos para analizar la normalidad multivariante.

En primer lugar, recordemos que si $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ entonces cada una de las componentes sigue una ley normal univariante. Por tanto se puede estudiar la normalidad de cada una de las variables mediante procedimientos gráficos (gráfico de cuantiles, histograma, etc) o bien formales, donde destaca el test de Shapiro-Wilk.

Además, cada par de variables (X_i, X_j) sigue una ley normal bivariante. En tal caso las nubes de puntos observados deben sugerir una forma elíptica o circular. El tamaño y orientación dependerá de la intensidad y signo del coeficiente de correlación lineal entre ambas variables.

Otra opción es utilizar las llamadas distancias de Mahalanobis. En general, dados dos vectores *p*-dimensionales $\underline{u}$ y $\underline{v}$, y una matriz *M* no singular $p_x p$, se define la distancia al cuadrado de Mahalanobis

$$D^2(\underline{u},\underline{v};M) = (\underline{u}-\underline{v})^t M^{-1}(\underline{u}-\underline{v})$$

Una distancia de este tipo aparece en la expresión de la función de densidad de la ley $N_p(\underline{\mu},\Sigma)$

$$f(\underline{x}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^t \Sigma^{-1}(\underline{x}-\underline{\mu})}$$

Y se sabe que para un vector aleatorio que siga una ley normal multivariante se verifica que

$$(\underline{X}-\underline{\mu})^t \Sigma^{-1}(X-\underline{\mu}) \sim \chi^2_p$$

por lo que diversos autores han recomendado estudiar si la versión muestral de la anterior expresión se distribuye al menos de forma aproximada por tal distribución,

$$\text{¿} \quad D_i^2 = (\underline{X}_i - \overline{\underline{X}})^t \hat{\Sigma}^{-1}(\underline{X}_i - \overline{\underline{X}}) \sim \chi^2_p \quad ?$$

Se pueden aplicar procedimientos como el gráfico de cuantiles para comparar la distribución de estos valores con la ley teórica. Se trataría de dibujar la nube de puntos a partir de

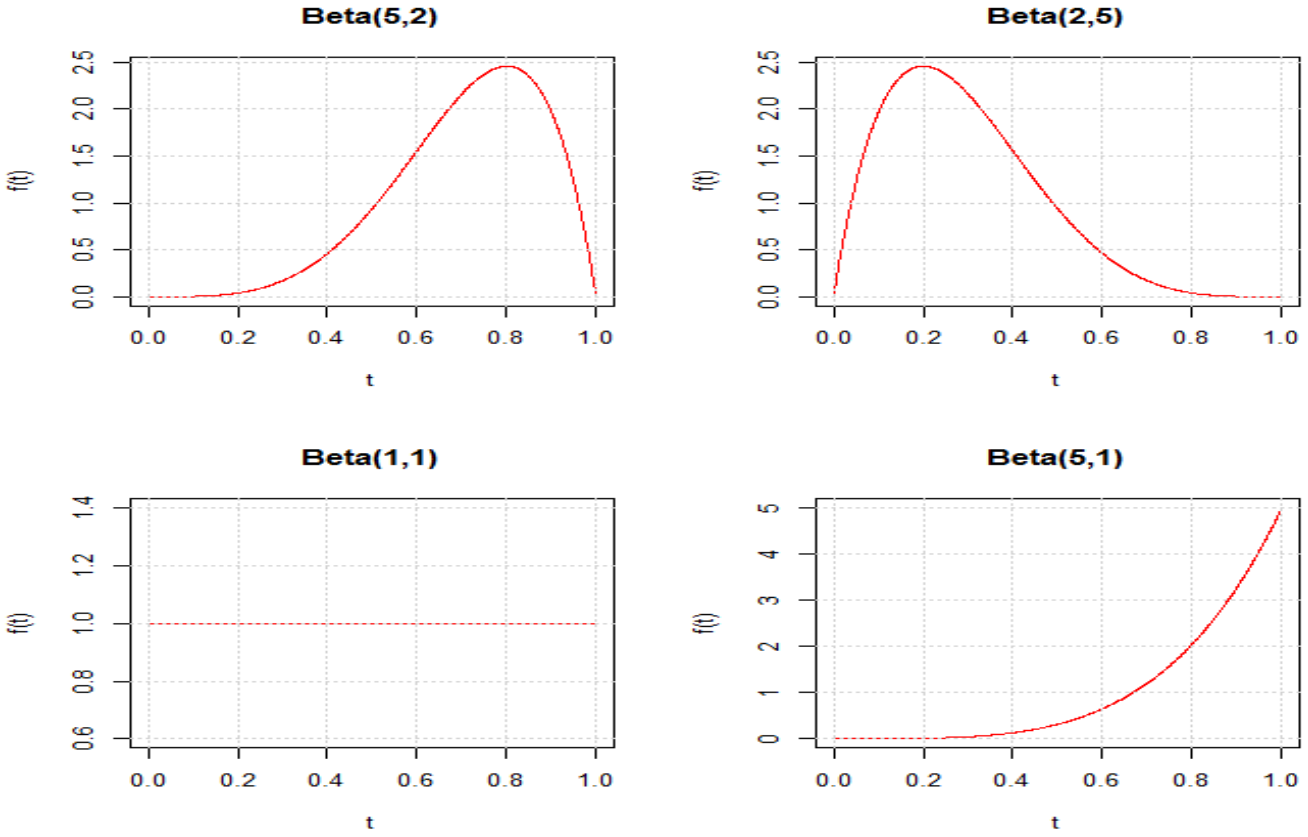
$$\{(D^2_{(i)}, \chi^2_{p,(i-0.5)/n}), i = 1,2,\ldots,n\}$$

Sin embargo se ha observado que la aproximación proporcionada por la ley χ^2_p puede ser muy pobre.

Otra opción es utilizar la siguiente propiedad. Suponiendo normalidad multivariante,

$$u_i = \frac{nD_i^2}{(n-1)^2} \sim Beta(p/2, (n-p-1)/2)$$

$$T \sim Beta(a,b) \Rightarrow f(t) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}, \quad a>0, b>0, 0 \le t \le 1$$



Sería preferible en tal caso construir gráficos de cuantiles comparando la distribución de los u_i con los de la ley Beta correspondiente, por lo que se dibuja la nube de puntos de los pares $(u_{(i)},v_i)$ donde los v_i son los cuantiles correspondientes a

$$P[T \leq v_i] = \frac{i-\alpha}{n-\alpha-\beta+1}, \quad T \sim Beta(a,b)$$

$$a = \frac{p}{2}, b = \frac{(n-p-1)}{2}, \quad \alpha = \frac{p-2}{2p}, \beta = \frac{n-p-3}{2(n-p-1)}$$

El gráfico resultante permite detectar desviaciones de la normalidad multivariante o identificar observaciones atípicas.

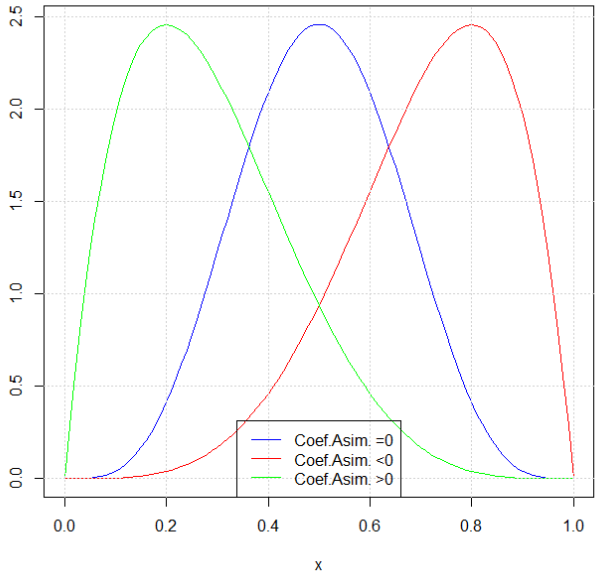
También se han propuesto algunos métodos inferenciales para realizar contrastes de bondad de ajuste. Uno de ellos es el Test de Mardia, que se basa en una generalización de los coeficientes de asimetría y curtosis.

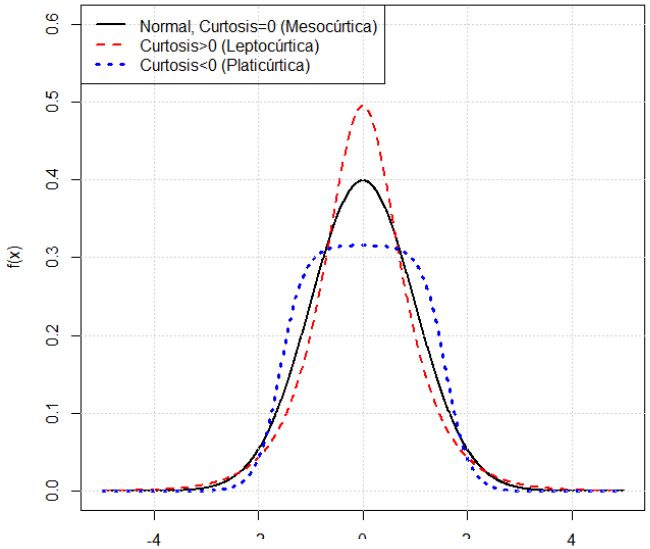
Para una variable aleatoria univariante se definen los coeficientes poblaciones de asimetría y curtosis:

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3}, \quad \delta = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

Para una muestra univariante se definen los coeficientes de asimetría y curtosis muestrales como

$$\hat{\gamma} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}, \quad \hat{\delta} = \frac{n\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{2}} - 3$$





lepto: fino

Ornitorrinco: “Platypus” (inglés)
Platy: amplio, ancho

Para poblaciones multivariantes, Mardia definió los siguientes coeficientes poblacionales de simetría y curtosis:

$$\beta_1 = E\left\{[(\underline{X} - \underline{\mu})^t \Sigma^{-1} (\underline{X} - \underline{\mu})]^3\right\}, \quad \beta_2 = E\left\{[(\underline{X} - \underline{\mu})^t \Sigma^{-1} (\underline{X} - \underline{\mu})]^2\right\}$$

Se puede comprobar que bajo normalidad multivariante

$$\beta_1 = 0, \quad \beta_2 = p(p+2)$$

Para la estimación de ambos coeficientes se definen los siguientes valores, donde se utiliza la estimación de máxima verosimilitud de la matriz de dispersión poblacional (div. por *n*):

$$g_{ij} = (\underline{X}_i - \overline{\underline{X}})^t \widetilde{\Sigma}^{-1} (\underline{X}_j - \overline{\underline{X}})$$

Se tiene entonces

$$b_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}^3, \quad b_2 = \frac{1}{n} \sum_{i=1}^{n} g_{ii}^2$$

Se pueden transformar en sendos estadísticos cuyas distribuciones *aproximadas* son chi-cuadrado (con p(p+1)(p+2)/6 g.l.) y N(0,1) respectivamente, lo que permite realizar contrastes de hipótesis sobre los coeficientes poblacionales. Para el coeficiente de asimetría se pueden realizar correcciones que mejoren la aproximación para muestras de tamaño reducido (como hace la función **mardiaTest** de la librería **MVN** de R).

2. Identificación de Outliers Multivariantes

Algunos de los métodos anteriores pueden ayudar a detectar observaciones atípicas, ya sea en las nubes de puntos o en los gráficos de cuantiles.

Suponiendo normalidad multivariante se pueden utilizar también algunos métodos inferenciales.

Para contrastar la existencia de una observación atípica, Wilks propuso el siguiente estadístico:

$$w = \max_i \frac{|(n-2)\hat{\Sigma}_{(-i)}|}{|(n-1)\hat{\Sigma}|} = 1 - \frac{nD^2_{(n)}}{(n-1)^2}$$

Se puede utilizar un procedimiento basado en la ley F de Snedecor teniendo en cuenta que

$$F_i = \frac{n-p-1}{p}\left[\frac{1}{1-nD_i^2/(n-1)^2} - 1\right] \sim F_{p,n-p-1}$$

Se puede trabajar con el máximo de estos coeficientes

$$F_{(n)} = \max_i F_i = \frac{n-p-1}{p}\left[\frac{1}{w} - 1\right]$$

De modo que el p-valor se obtiene con

$$P[F_{(n)} > f] = 1 - P[\max F_i \le f] = 1 - P[F_1 \le f \cap \ldots \cap F_n \le f] = 1 - P[F \le f]^n$$