Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 2: 22/07/24 – 27/07/24

Introducción

Recuerde que:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \\ \vdots \\ X_p \end{bmatrix}_{p\times 1}$$

es un **vector aleatorio** $p-$variado con función de densidad de probabilidad conjunta (**fdpc**):

Introducción

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_p)$$

con parámetros:

- **Vector de Medias**

$$\boldsymbol{\mu}_{p\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \\ \vdots \\ \mu_p \end{bmatrix}_{p\times 1}$$

- **Matriz de Varianzas y Covarianzas**

$$\boldsymbol{\Sigma}_{p\times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{kk} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} & \cdots & \sigma_{kp} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pk} & \cdots & \sigma_{pp} \end{bmatrix}_{p\times p}$$

Introducción

$$\mathbf{X}_{N\times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nk} & \cdots & X_{Np} \end{bmatrix}_{N\times p} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_j^\top \\ \vdots \\ \mathbf{X}_N^\top \end{bmatrix}_{N\times p}$$

Introducción

representa a una población que tiene N unidades de investigación, digamos

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_N,$$

donde

$$\mathbf{X}_j^\top = \begin{bmatrix} X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \end{bmatrix}_{1\times p}$$

contiene la **información multivariada** de la $j-$ésima unidad de investigación, $j = 1, 2, \ldots, N$. Además asuma que

$$f_{\mathbf{X}_j}(\mathbf{x}_j) = f_{\mathbf{X}_j}(x_{j1}, x_{j2}, \ldots, x_{jk}, \ldots, x_{jp})$$

es la **función de densidad de probabilidad** (fdp) de $\mathbf{X}_j^\top$.

Introducción

Observación: Las mediciones de las $\underline{p \text{ variables en } \boldsymbol{X}_j^\top}$ generalmente estarán correlacionadas, mientras que las mediciones $\underline{\text{entre las } \boldsymbol{X}_j^\top}$ suelen ser independientes.

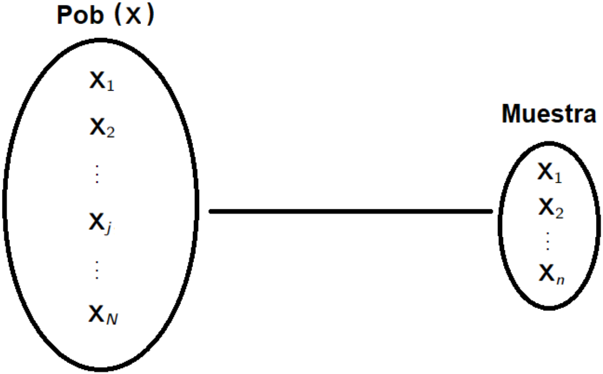
Si $\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_n^\top$ son observaciones **independientes** que se extraen de

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_p)$$

entonces $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ es una **muestra aleatoria** de $f_{\mathbf{X}}(\mathbf{x})$ y por lo tanto

$$f_{\mathbf{X}_1,\ldots,\mathbf{X}_n}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = f_{\mathbf{X}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}_n}(\mathbf{x}_n) = \prod_{j=1}^{n} f_{\mathbf{X}_j}(\mathbf{x}_j)$$

Introducción



$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_p)$$

Una muestra $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ de $f_{\mathbf{X}}(\mathbf{x})$ puede ser escrita por

$$\mathbf{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np} \end{bmatrix}_{n\times p} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_j^\top \\ \vdots \\ \mathbf{X}_n^\top \end{bmatrix}_{n\times p}$$

Introducción

En particular, si

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$$

es una muestra aleatoria de una población **Normal Multivariada**, denotada por $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; la cual:

- es generalización a varias dimensiones de la **normal univariada**.
- juega un papel funtamental en el **Análisis Multivariado**, puesto que muchos de los **estadísticos** usados en el análisis multivariado tiene una **distribución**[1] **aproximadamente normal multivariada** independiente de la naturaleza de los datos.

[1] Distribución de Muetreo

Introducción

Recuerde que si $X \sim \mathrm{N}(\mu, \sigma^2)$:

- $x \in \mathbb{R}$
- $\mu \in \mathbb{R}$
- $\sigma^2 > 0$
- $$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \mathbf{I}_{(-\infty,+\infty)}(x)$$
- el término

$$\frac{(x-\mu_X)^2}{\sigma_X^2} = \left[\frac{x-\mu_X}{\sigma_X}\right]^2 = [x-\mu_X]\left(\sigma_X^2\right)^{-1}[x-\mu_X]$$

es la **distancia cuadrática entre** x **y** μ, la cuál está medida en unidades de σ.

Distribución Normal Multivariada

Distribución Normal Multivariada

Distribución Normal Mutlivariada

El vector aleatorio $\mathbf{X}_{p\times 1} = [X_1, X_2, \ldots, X_p]^\top$ tiene Distribución Normal Multivariada ($p-$variada) con **vector de medias** $\boldsymbol{\mu}_{p\times 1}$ y **matriz de varianzas y covarianzas** $\boldsymbol{\Sigma}_{p\times p}$ (simétrica y definida positiva), si su función de densidad de probabilidad está dada por

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

con $\mathbf{x} \in \mathbb{R}^p$.

Observación: Usamos la notación $\boldsymbol{X}_{p\times 1} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *para referirnos a que e vector aleatorio* $\boldsymbol{X}_{p\times 1}$ *sigue una distribución normal* $p-$*variada con vector de media* $\boldsymbol{\mu}$ *y matriz de varianzas y covarianzas* $\boldsymbol{\Sigma}$.

Distribución Normal Multivariada

Si

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$$

es una muestra aleatoria de una población $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces

$$f_{\mathbf{X}_j}(\mathbf{x}_j) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})\right\}$$

y la fdpc de $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ queda dada por

$$f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{j=1}^{n} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})\right\}$$

$$= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})\right\}$$

Distribución Normal Multivariada

Una definición alternativa

El vector aleatorio $\mathbf{X}_{p\times 1} = [X_1, X_2, \ldots, X_p]^\top$ tiene distribución normal multivariada sí y sólamente sí para cualquier vector fijo $\mathbf{a}_{p\times 1}$, **toda combinación lineal**

$$\mathbf{a}_{1\times p}^\top \mathbf{X}_{p\times 1} = \sum_{j=1}^{p} a_j x_j$$

tiene distribución **normal univariada**.

Distribución Normal Multivariada

Observaciones:

- $\mathbf{a}^{\top}_{1\times p}\mathbf{X}_{p\times 1}$ es una **combinación lineal** de los elementos de $\mathbf{X}$ con los coeficientes del vector $\mathbf{a}$ (usaremos en algunas técnicas, por ejemplo en **Análisis de Componentes Principales**)
- $\mathbf{a}^{\top}_{1\times p}\mathbf{X}_{p\times 1}$ es visto como una **proyección** de $\mathbf{X}$ (espacio $p-$ dimensional) en un espacio unidimensional.

Distribución Normal Multivariada

Example

Para $p = 2$ (caso **bivariado**) se tiene que

$$\mathbf{X}_{p\times 1} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \right)$$

Esto es, el vector aleatorio $\mathbf{X}$ tiene **distribución normal bivariada** con vector de medias $\boldsymbol{\mu}_{2\times 1}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}_{2\times 2}$.

Distribución Normal Multivariada

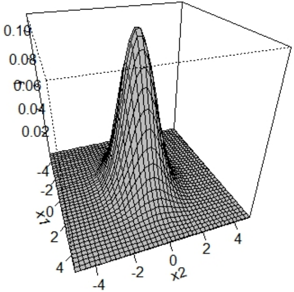


Figura: Ilustración de un vector normal bivariado con $\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 2, \sigma_{22} = 2, \sigma_{12} = \sigma_{21} = 0$

Observe que:

- la densidad es una superficie (de **campana**) sobre el plano (x_1, x_2) **centrada** en su **vector de medias** $\mu = \mathbf{0}$.
- la **altura** de la superficie es la **fdp**.

En la figura asumimos que las **variables** son **no correlacionadas** $(\sigma_{12} = \sigma_{21} = 0)$.

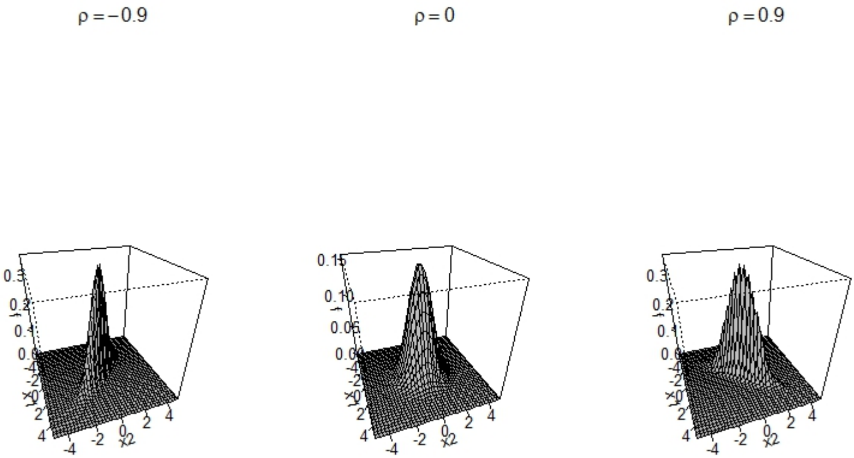
Distribución Normal Multivariada

Observe que:

- la densidad es una superficie (de **campana**) sobre el plano (x_1, x_2) **centrada** en su **vector de medias** $\mu = \mathbf{0}$.
- la **altura** de la superficie es la **fdp**.

En la figura asumimos que las **variables** son **no correlacionadas** ($\sigma_{12} = \sigma_{21} = 0$). Si consideraramos **correlación positiva**, la gráfica se "**aplastaría**" y la curva se alinearía de una esquina a otra. En el caso que la **correlación** entre las variables fuera **negativa**, la superficie también se "**aplastaría**" y se alinearía en las otras esquinas.

Distribución Normal Multivariada



Distribución Normal Multivariada

```
# --- Densidad de na Normal Bivariada --- #
require(mvtnorm)
mu<-c(0,0)                                  # mu=0
Sigma<-matrix(c(2,0,0,1),ncol = 2) # sigma12=0

x1<-seq(-5,5,length=40)
x2<-x1

# --- densidad de la f normal multiv.
f<-matrix(0,
          nrow = length(x1),
          ncol = length(x2))
```

Distribución Normal Multivariada

```
for(i in 1:length(x1))
for(j in 1:length(x2))
f[i,j]<-dmvnorm(c(x1[i],x2[j]),
                mean = mu,
                sigma = Sigma)

# --- Grafico
persp(x1, x2, f,
      theta = 70,  # ángulo de visualización
      phi = 30,   # ángulo de visualización
      col="gray",
      ticktype = "detailed"
      )
```

Distribución Normal Multivariada

Un resultado Importante:

Si

$$\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

y si

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$$

entonces

$$\mathbf{Z} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$$

$$(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$$

Observación: $\chi_p^2 \equiv Gamma(\alpha = p/2, \beta = 2)$

Distribución Normal Multivariada

Observación:

- *La cantidad* $(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ *es llamada* ***Distancia de Mahalanobis*** *entre* $\boldsymbol{X}$ *y* $\boldsymbol{\mu}$*. Note que ésta es una distancia (cuadrática) entre dos vectores, que es ponderada por* $\boldsymbol{\Sigma}^{-1}$*.*
- *Esa ponderación se hace necesaria cuando las variables en* $\boldsymbol{X}$ *tienen magnitudes muy diferentes y en consecuencias varianzas muy diferentes. Si en esos casos no se hace esa ponderación (estandarización) las variables con mayor varianza podrian dominar esa distancia.*
- Si distancias de Mahalanobis observadas o muestrales,

$$d_i^2 = (\mathbf{X}-\bar{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\bar{\mathbf{x}})$$

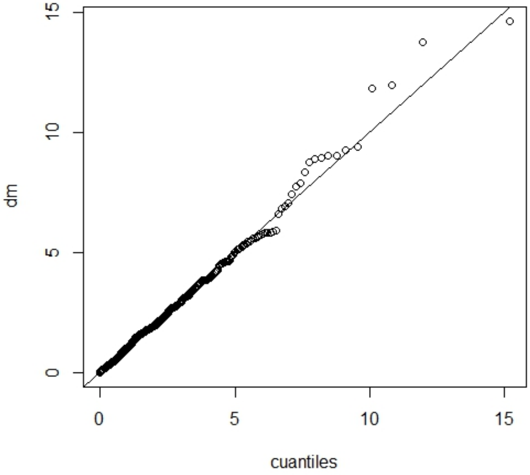
se ajustan a una distribución χ_p^2 tendremos indicios o sopechas sobre la normalidad multivariada de $\mathbf{X}$.

Distribución Normal Multivariada

Vamos a verificar la distribución de distancia de Mahalanobis en R:

```
# --- Distribucion de la Distancia de Mahalanobis --- #
require(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
x<-mvrnorm(n,mu,sigma)
xbarra<-colMeans(x)
s<-cov(x)
dm<-mahalanobis(x,xbarra,s)
cuantiles<-qchisq(ppoints(length(x)),df=2)
qqplot(cuantiles,dm)
abline(0,1)
```

Distribución Normal Multivariada



Distribución Normal Multivariada

La fdp de $\mathbf{X}_{2\times 1} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ está dada por

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right)^2 \right.\right.$$

$$\left.\left. + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right)^2 \right] - 2\rho_{12} \left[\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} \right] \right\}$$

Note que en el **exponente** se tiene la **forma geométrica** de una **elipse**. Entonces para cada para de X_1, X_2 que se hagan cortes de la fdp, se tendrá una elipse que es constante para todos los valores de X_1, X_2.

Pontificia Universidad JAVERIANA Bogotá

Distribución Normal Multivariada

Observaciones:

- si $\rho_{12} = 0$, esto es, $\underline{X_1}$ y $\underline{X_2}$ no están correlacionadas, entoces

$$f_{\mathbf{X}}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

  Nota: ***En el caso normal bivariado con correlación cero implica independencia***.

- si $\rho_{12} = \pm 1$, esto es si X_1 y X_2 tienen una correlación perfecta, esto es, una es exactamente función de la otra, entonces

$$f_{\mathbf{X}}(x_1, x_2) = 0$$

Pontificia Universidad JAVERIANA Bogotá

Sea

$$\{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2, \quad c > 0\}$$

el conjunto de todos los puntos $\mathbf{x}$ cuya distancia a $\boldsymbol{\mu}$ es constante. Entonces, la densidad de la normal multivariada es constante sobre superficies donde la distancia cuadrática $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ es constante. Estos conjuntos de puntos son llamados **contornos** [2].

Observación:

- *Los **ejes de los contornos están en la dirección de autovectores de $\boldsymbol{\Sigma}^{-1}$** y las longitudes de sus ejes son proporcionales a las raíces cuadradas de los recíprocos de sus vectores propios.*
- *Los **contornos** proporcionan la **densidad en cada curva**.*

[2] Un contorno corresponde a la superficie de una elipsoide centrada en μ

Recuerde que, lo **autovalores** de $\mathbf{\Sigma}$ se obtienen en la solución de

$$|\mathbf{\Sigma} - \lambda\, \mathbf{I}\,| = 0$$

Para $\mathbf{\Sigma}_{2\times 2}$, tenemos

$$\lambda_1 = \sigma_{11} + \sigma_{12} \qquad \lambda_2 = \sigma_{11} - \sigma_{22}$$

y los autovectores se obtienen apartir de la solución de

$$\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i \qquad \mathbf{e}_i = \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix}$$

luego

$$\mathbf{e}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Distribución Normal Multivariada

Si $\sigma_{21} = \sigma_{12} > 0$, entonces

- λ_1 sería el mayor valor propio y su vector propio asociado $(\mathbf{e}_1)$ caería sobre una línea recta de 45 grados através del punto $\boldsymbol{\mu}$.
- El eje mayor estaría determinado por $\pm c\sqrt{\lambda_1}\mathbf{e}_1$.
- λ_2 sería el menor valor propio y su vector propio asociado $(\mathbf{e}_2)$ caería sobre un recta perpendicular a la recta de 45 grados através del punto $\boldsymbol{\mu}$.
- El eje menor estaría determinado por $\pm c\sqrt{\lambda_2}\mathbf{e}_2$.

Si $\sigma_{21} = \sigma_{12} > 0$, entonces

Distribución Normal Multivariada

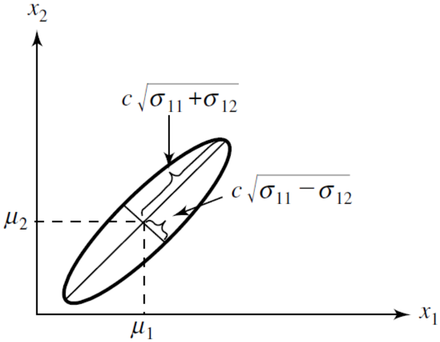
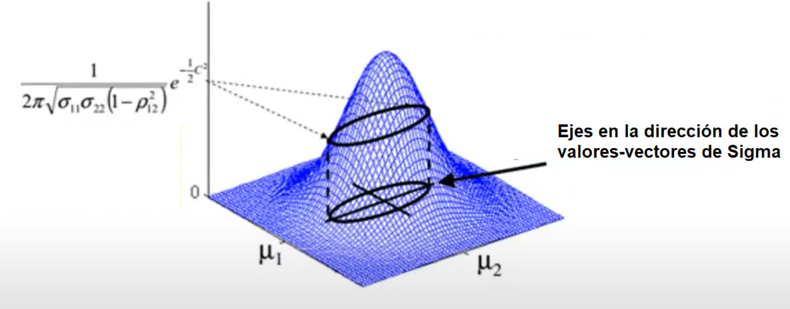


Figura: Figura 4.3 de Johnson and Wichern (2013) - Applied Multivariate Statistical Analysis, pp. 154

Distribución Normal Multivariada



Distribución Normal Multivariada

Gráfico de Contornos: Se intercepta a la densidad bivariada con un plano que tiene una elipse pequeña, la cual se hace más grande en la medida que se baja ese plano.

```
# --- Grafico de Contornos --- #
 contour(x1,
        x2,
        f,
        draw=T,
        nlevels = 20,
        labcex = 0.8,
        xlab=expression(x[1]),
        ylab=expression(x[2])
        )
```

Distribución Normal Multivariada

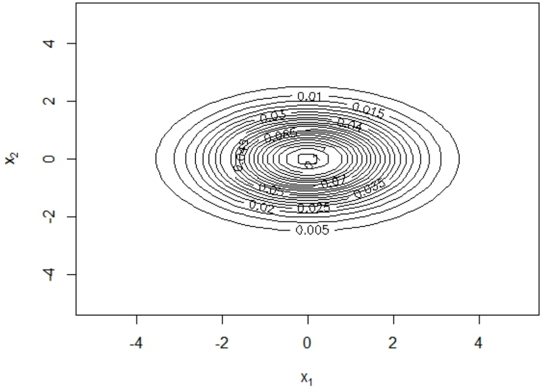


Figura: Contornos de una normal bivariada con $\mu_1 = \mu_2 = 0, \sigma_{11} = 2, \sigma_{22} = 1, \sigma_{12} = \sigma_{21} = 0$

Distribución Normal Multivariada

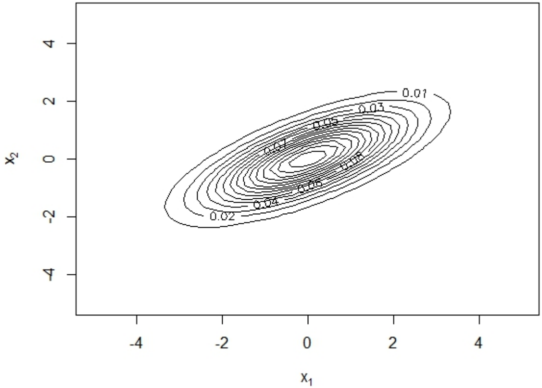


Figura: Contornos de una normal bivariada con $\mu_1 = \mu_2 = 0, \sigma_{11} = 2, \sigma_{22} = 1, \sigma_{12} = \sigma_{21} = 1$

Distribución Normal Multivariada

Si las observaciones fueran generadas por una distribución normal multivariada:

- todas las distribuciones bivariadas deberian ser normales y los contornos de densidad constante deberían ser eplipses.
- Para un α dado se esperaría que alrededor del $\alpha\,\%$ de las observaciones caigan dentro de la elipse dada por

$$\left\{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})^\top \mathsf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq \chi^2_2(\alpha)\right\}$$

Observación: Estas son verificaciones informales

¿Cómo verificamos normalidad multivariada formalmente?

Distribución Normal Multivariada

¿Cómo verificamos normalidad multivariada formalmente?

Rta: Probando las hipótesis:

$$\begin{aligned} \mathsf{H}_0 &: \text{Los datos provienen de una poblacion Normal Multivariada} \\ \mathsf{H}_1 &: \text{Los datos NO provienen de una poblacion Normal Multivariada} \end{aligned}$$

Matemáticamente

$$\begin{aligned} \mathsf{H}_0 &: \mathbf{X}_{p\times 1} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathsf{H}_1 &: \mathbf{X}_{p\times 1} \nsim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

Distribución Normal Multivariada

¿Cómo verificamos normalidad multivariada formalmente?

Rta: Probando las hipótesis:

$$
\begin{aligned}
\mathrm{H}_0 &: \text{Los datos provienen de una poblacion Normal Multivariada} \\
\mathrm{H}_1 &: \text{Los datos NO provienen de una poblacion Normal Multivariada}
\end{aligned}
$$

Matemáticamente

$$
\begin{aligned}
\mathrm{H}_0 &: \mathbf{X}_{p\times 1} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathrm{H}_1 &: \mathbf{X}_{p\times 1} \nsim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
$$

Existen varias estadísticas de pruebas y varios paquetes para probar estas hipótesis, dentro de los cuales tenemos:

Distribución Normal Multivariada

```
# ----- Prueba de Shapiro ----- #
require(mvShapiroTest)
mvShapiro.Test(X)

# ----- Otras Pruebas ----- #
require(MVN)
mvn(X, mvnTest="mardia")    # test de Mardia
mvn(X, mvnTest="hz")        # test de Henze-Zirkler
mvn(X, mvnTest="royston")   # test de Royston
mvn(X, mvnTest="dh")        # test de Doornik-Hansen
```

Veremos en detalle más adelante (Semana 3)

Distribución Normal Multivariada

Algunas propiedades de la distribución normal multivariada:

1. Las **combinaciones lineales** de $\mathbf{X}_{p\times 1} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tienen distribución normal.
   i. Si $Y = \mathbf{a}^\top \mathbf{X}$ con $\mathbf{a}_{p\times 1}$ fijo, entonces
   $$Y \sim \mathrm{N}_1\left(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma}\, \mathbf{a}\right)$$
   ii. Si $\mathbf{Y}_{q\times 1} = \mathbf{A}\,\mathbf{X}$, donde $\mathbf{A}_{q\times p}$ y $\mathbf{b}_{q\times 1}$ son fijos $(q < p)$, entonces
   $$\mathbf{Y} \sim \mathrm{N}_q\left(\mathbf{A}\,\boldsymbol{\mu}, \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^T\right)$$
   iii. Si $\mathbf{Y}_{q\times 1} = \mathbf{A}\,\mathbf{X} + \mathbf{b}$, donde $\mathbf{A}_{q\times p}$ y $\mathbf{b}_{q\times 1}$ son fijos $(q < p)$, entonces
   $$\mathbf{Y} \sim \mathrm{N}_q\left(\mathbf{A}\,\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^T\right)$$

Distribución Normal Multivariada

2. La distribución **marginal** de cada variable aleatoria de $\mathbf{X}$ tiene una **distribución normal univariada**, esto es

$$X_i \sim \mathrm{N}(\mu_i, \sigma_{ii}) \qquad \forall\, X_i \in \mathbf{X}$$

3. **Cualquier subconjunto** de los componentes de $\mathbf{X}$ tienen una **distribución normal multivariada**.
4. La **covarianza cero** implica que los componentes correspondientes se distribuyen de forma **independiente**.
5. Las **distribuciones condicionales** de los componentes son normales (**multivariadas**).

Distribución Normal Multivariada

Suponga que en $\mathbf{X}_{p\times 1}$ se consideran dos subgrupos de tamaño q y $p-q$:

$$\mathbf{X}_{p\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \\ \cdots \\ X_{q+1} \\ X_{q+2} \\ \vdots \\ X_p \end{bmatrix}_{p\times 1} = \begin{bmatrix} \mathbf{X}^{(1)}_{q\times 1} \\ \cdots \\ \mathbf{X}^{(2)}_{(p-q)\times 1} \end{bmatrix}_{p\times 1}$$

Dada esta partición sigue que

Distribución Normal Multivariada

$$\boldsymbol{\mu}_{p\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \\ \cdots \\ \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)}_{q\times 1} \\ \cdots \\ \boldsymbol{\mu}^{(2)}_{(p-q)\times 1} \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{p\times p} = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \vdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix}$$

Distribución Normal Multivariada

donde

- Σ_{11} es la **matriz** de $q \times q$ de **varianzas y covarianzas** de los elementos del subvector $\underline{\mathbf{X}}^{(1)}$.
- Σ_{22} es la **matriz** de $(p-q)\times(p-q)$ de **varianzas y covarianzas** de los elementos del subvector $\underline{\mathbf{X}}^{(2)}$.
- Σ_{12} es la **matriz** de $q\times(p-q)$ de **covarianzas** entre los elementos del subsector $\underline{\mathbf{X}}^{(1)}$ con los elementos del subvector $\underline{\mathbf{X}}^{(2)}$.
- Σ_{21} es la **matriz** de $(p-q)\times q$ de **covarianzas** entre los elementos del subvector $\underline{\mathbf{X}}^{(2)}$ con los elementos del subsector $\underline{\mathbf{X}}^{(1)}$.

Distribución Normal Multivariada

De ahí:

- (Propiedad 3):

$$\mathbf{X}^{(1)} \sim \mathrm{N}_q(\boldsymbol{\mu}^{(1)}, \Sigma_{11})$$
$$\mathbf{X}^{(2)} \sim \mathrm{N}_{p-q}(\boldsymbol{\mu}^{(2)}, \Sigma_{22})$$

- Si $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ son **independientes** entonces

$$\Sigma_{12} = \mathbf{0}$$

- $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ son **independientes** si y solo si

$$\Sigma_{12} = \mathbf{0}$$

Distribución Normal Multivariada

- Si

$$\begin{bmatrix} \mathbf{X}^{(1)}_{q_1\times 1} \\ \cdots \\ \mathbf{X}^{(2)}_{q_2\times 1} \end{bmatrix} \sim \mathrm{N}_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}^{(1)}_{q_1\times 1} \\ \cdots \\ \boldsymbol{\mu}^{(2)}_{q_2\times 1} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \vdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix} \right)$$

entonces

$$\mathbf{X}^{(1)} \Big| \mathbf{x}^{(2)}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}_{1.2}, \Sigma_{1.2}\right)$$

donde

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left[\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right]$$

$$\Sigma_{1.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Distribución Normal Multivariada

Example

Suponga que

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathrm{N}_3 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \right)$$

Son $\mathbf{X}^{(1)} = [X_1, X_2]^\top$ y $\mathbf{X}^{(2)} = X_3$ independientes?

Obs: Example 4.6 from Johnson and Wicher (2014), Applied Multivariate Statistical Analysis, 6th Edition, pp. 160.

Distribución Normal Multivariada

Solución:

Como

$$\mathbf{X}^{(1)} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \qquad \mathbf{X}^{(2)} = X_3$$

sigue que

- $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_3 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(2)} \end{bmatrix}$
- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \cdots \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$

evidentemente

$$\boldsymbol{\mu}^{(1)} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \text{y} \qquad \boldsymbol{\mu}^{(2)} = \mu_3$$

- $$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 1 & \vdots & 0 \\ 1 & 3 & \vdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & 2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \vdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix}$$

evidentemente

$$\Sigma_{11} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix} \qquad \Sigma_{12} = \mathbf{0}$$

Distribución Normal Multivariada

$$\Sigma_{12} = \text{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, X_3\right) = \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma_{22} = \sigma_{33} = 2$$

Así, tenemos que

$$\begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(2)} \end{bmatrix} \sim \text{N}_{2+1}\left(\begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \cdots \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \vdots & \mathbf{0} \\ \cdots & \vdots & \cdots \\ \mathbf{0} & \vdots & \Sigma_{22} \end{bmatrix}\right)$$

Como

$$\Sigma_{12} = \mathbf{0}$$

entonces $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ son independientes.

Distribución Normal Multivariada

Ejercicio: Genere muestras de tamaño $n = 100$ de

- $\mathbf{X} \sim \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$
- $X_3 \sim \text{Beta}(3, 2)$
- $X_4 \sim \text{Gamma}(2, 2)$

y para cada muestra realice

- boxplot
- histograma
- Q-Qplot de las distancias de Mahalanobis