Actividad 1.4 La Normal Multivariada

Raúl Correa Ocañas

2024-08-14

Ejercicio 1.

Hallar el procedimiento para el cálculo de probabilidad de que $P(X_1 \leq 3, X_2 \leq 2)$ con X_1, X_2 se distribuyen normalmente con $\mu = \begin{bmatrix} 2.5 \\ 4 \end{bmatrix}$ y $\Sigma = \begin{bmatrix} 1.2 & 0 \\ 0 & 2.3 \end{bmatrix}$.

```
upper_bound = c(3,2)
mu = c(2.5, 4)
n_dim = length(mu)

sigma = matrix(c(1.2, 0, 0, 2.3), nrow = n_dim)

pmnorm(x = upper_bound, mean = mu, varcov = sigma)
```

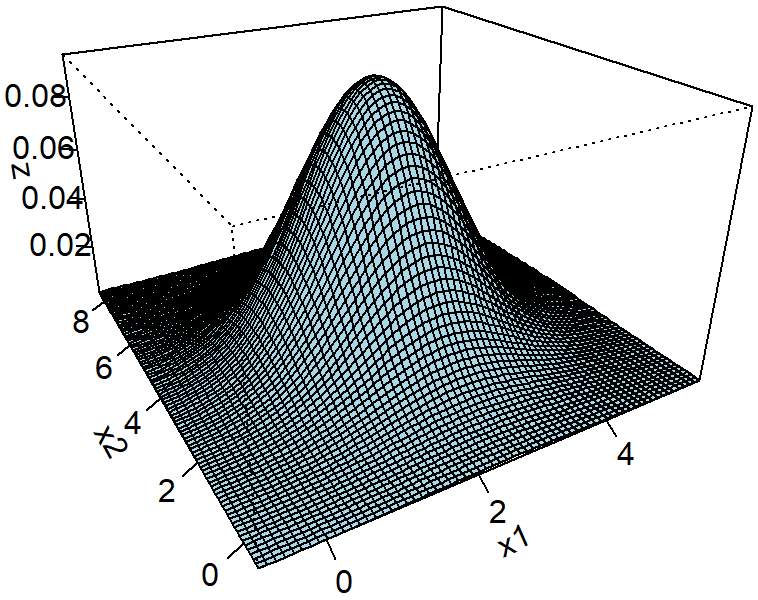
```
## [1] 0.06328658
```

Ejercicio 2.

```
library(mnormt)
x1 = seq(mu[1]-3*sigma[1,1]**(1/2), mu[1]+3*sigma[1,1]**(1/2), 0.1)
x2 = seq(mu[2]-3*sigma[2,2]**(1/2), mu[2]+3*sigma[2,2]**(1/2), 0.1)

f = function(x1, x2) dmnorm(cbind(x1, x2), mu, sigma)
z = outer(x1, x2, f)

persp(x1, x2, z, theta=-30, phi=25, expand=0.6, ticktype='detailed', col = "lightblue")
```

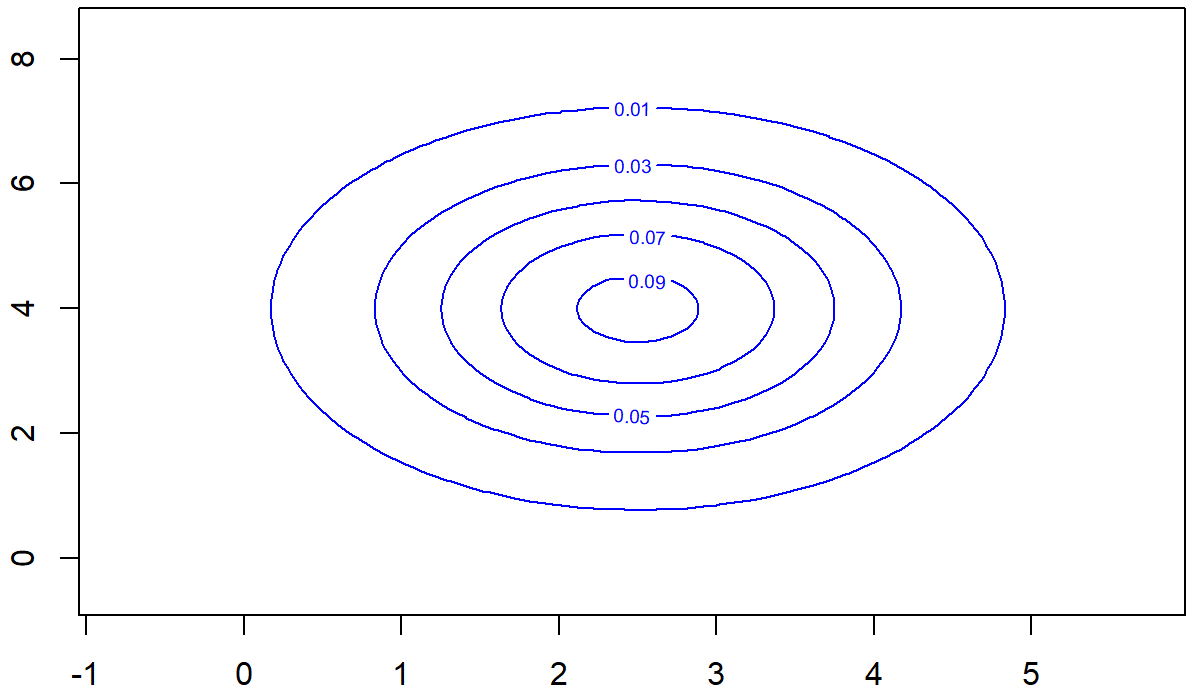


Ejercicio 3.

¿A qué altura (0.01, 0.03, 0.05, 0.07, 0.09, 0.1) crees que esté el contorno con el diámetro más pequeño? ¿y el más grande?

Los datos más cercanos a la media tienen mayor altura debido a la probabilidad de estar en un intervalo cerca a la media incrementa mientras más cerca esté a la media. Esto implica que su contorno debe ser chico relativamente a otras alturas, porque para poder incluir datos con menor altura tendrian que estar mas alejados de la media, lo cual se traduce en un contorno mayor al observar el plano X_1 y X_2. Por lo tanto, el contorno de altura 0.1 tiene el contorno más pequeño, y el contorno con altura 0.01 es el más grande.

```
contour(x1, x2, z, col = "blue", levels = c(0.01, 0.03, 0.05, 0.07, 0.09))
```



¿Cómo se relaciona el resultado del primer problema con el segundo?

El primer problema se enfoca en calcular la probabilidad de que dos variables aleatorias normalmente distribuidas caigan dentro de un rectángulo definido por $X_1 \leq 3, X_2 \leq 2$ utilizando una distribución normal multivariada. En el segundo problema, se visualiza la función de densidad conjunta de estas dos variables. El área bajo la superficie tridimensional en el gráfico del segundo problema hasta el punto $(3, 2)$ es la probabilidad que se calculó en el primer problema.

¿Cómo se relacionan los gráficos de los incisos 2 y 3?

El gráfico del problemaa 2 es una representación tridimensional de la función de densidad conjunta de las dos variables, donde la altura del gráfico en un punto representa el valor de la densidad en ese punto. Por otro lado, el gráfico del problema 3 muestra las líneas de contorno de esta misma función de densidad para diferentes niveles de altura. Las líneas de contorno en el gráfico son proyecciones del gráfico tridimensional sobre el plano X_1 y X_2. Por lo tanto, ambos gráficos son diferentes formas de representar la misma función de densidad.

¿Cómo se relaciona la altura del contorno (pregunta3) con el porcentaje de datos que abarcaría el contorno?

La altura del contorno en una distribución normal multivariada está inversamente relacionada con el porcentaje de datos que abarca. Contornos de mayor altura representan regiones de mayor densidad de probabilidad, y estas regiones tienden a estar cerca de la media de la distribución. Estas regiones más densas abarcan un porcentaje menor de los datos totales. En contraste, contornos de menor altura representan regiones de menor densidad de probabilidad, que están más alejadas de la media y abarcan un porcentaje mayor de los datos.

Ejercicio 4.

Las variables X_1 y X_2 son conjuntamente normales e independientes con distribución univariada $N(0,1)$. Dadas las combinaciones lineales:

$Y_1 = 3 + 2X_1 - X_2$

$Y_2 = 5 + X_1 + X_2$

1. Encuentra la función de densidad bivariante de Y_1 y Y_2.

$Y \sim N(\mu, \Sigma)$

$\mu = \begin{bmatrix} E(Y_1) \\ E(Y_2) \end{bmatrix}$

$E(Y_1) = E(3) + E(2X_1) - E(X_2)$ y $E(Y_2) = E(5) + E(X_1) + E(X_2)$, por lo que $\mu = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$

$\Sigma = \begin{bmatrix} Cov(Y_1, Y_1) & Cov(Y_1, Y_2) \\ Cov(Y_2, Y_1) & Cov(Y_2, Y_2) \end{bmatrix}$

$Cov(Y_1, Y_1) = Var(Y_1) = Var(3) + Var(2X_1) + Var(-X_2)$

$\therefore Cov(Y_1, Y_1) = 5$

$Cov(Y_2, Y_2) = Var(Y_2) = Var(5) + Var(X_1) + Var(X_2)$

$\therefore Cov(Y_2, Y_2) = 2$

$Cov(Y_1,Y_2) = Cov(Y_2,Y_1) = Cov(2X_1 - X_2, X_1 + X_2) $

$Cov(Y_1,Y_2) = Cov(2X_1,X_1) + Cov(2X_1,X_2) - Cov(X_1,X_2) - Cov(X_2,X_2) $

$Cov(Y_1, Y_2) = 2 + 0 + 0 - 1 = 1$

$\therefore \Sigma = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$

Finalmente,

$(Y_1, Y_2) \sim N_2(\begin{bmatrix} 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix})$

□

2. ¿Cómo se distribuye la variable $Z = 3Y_1 + 4Y_2 - 1$?

$Z = 3(3 + 2X_1 - X_2) + 4(5 + X_1 + X_2) - 1$

$Z = 9 + 6X_1 - 3X_2 + 20 + 4X_1 + 4X_2 - 1$

$Z = 28 + 10X_1 + X_2$

Debido a que Z es una combinación lineal de la forma:

$c_1X_1 + c_2X_2 + \cdots + c_{n-1}X_{n-1} + c_nX_n$

La distribución de Z también será una distribución normal con:

$\mu = E(3Y_1) + E(4Y_2) - E(1) = 3 \times 3 + 4 \times 5 - 1 = 28$

$\Sigma = Var(28) + Var(10X_1) + Var(X_2) = 101$

$\therefore Z \sim N(28, 101)$

□

Ejercicio 5.

```
data_path = '../data/datos-A2.csv'
df = read.csv(data_path)
df = as.data.frame(df)
df
```

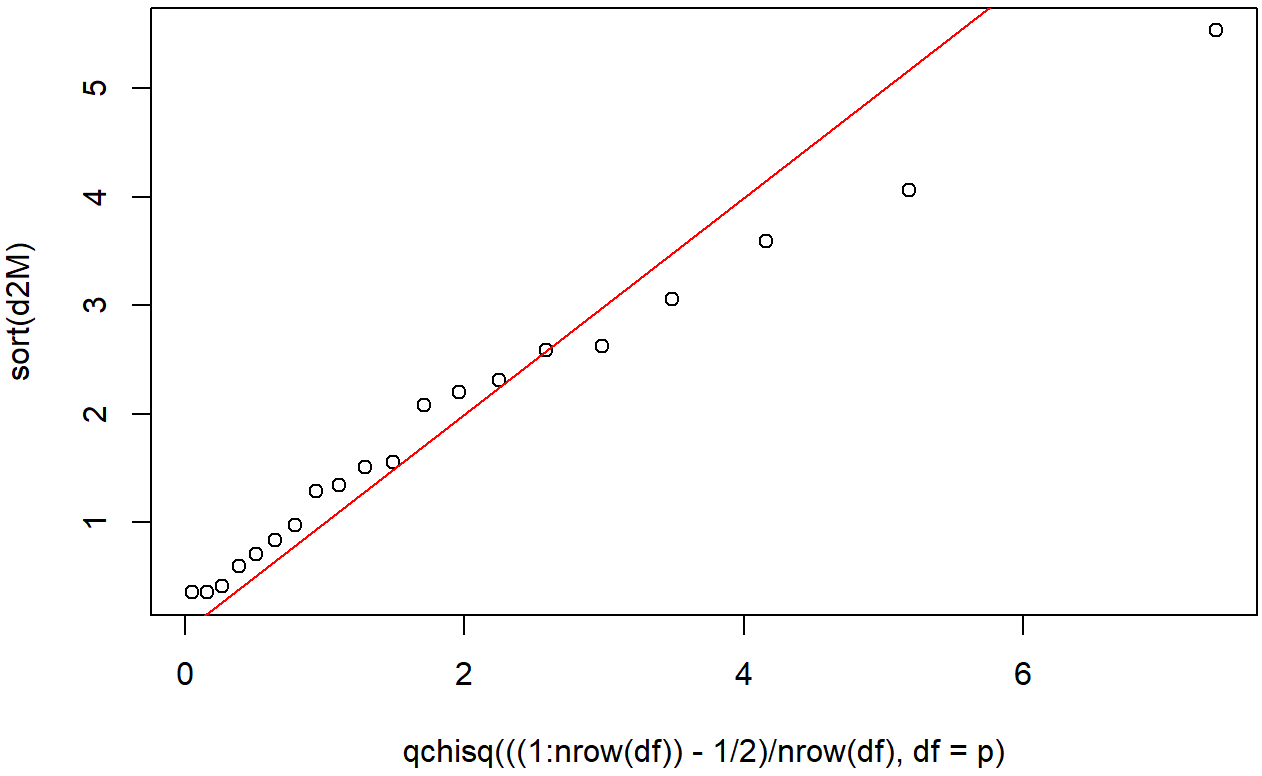
```
##      x   y
## 1  0.5 4.7
## 2  0.2 6.5
## 3  0.3 5.2
## 4  0.2 6.0
## 5  0.3 5.8
## 6  0.4 4.2
## 7  0.2 4.4
## 8  0.1 5.0
## 9  0.2 3.9
## 10 0.1 5.6
## 11 0.1 3.7
## 12 0.1 6.6
## 13 0.1 4.5
## 14 0.0 6.7
## 15 0.0 4.6
## 16 0.1 4.0
## 17 0.4 5.0
## 18 0.1 5.0
## 19 0.1 3.3
## 20 0.1 6.1
```

```
p = 2
X = colMeans(df)

S = cov(df)

d2M =  mahalanobis(df,X,S)

plot(qchisq(((1:nrow(df)) - 1/2)/nrow(df),df=p),sort( d2M ) )
abline(a=0, b=1,col="red")
```




```
mvn(df,subset = NULL,mvn = "mardia", covariance = FALSE,showOutliers = FALSE)
```

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  3.59823747819632  0.46309914697164    YES
## 2 Mardia Kurtosis -1.43530997731026 0.151198785877334    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        1.2355    0.0024    NO
## 2 Anderson-Darling     y        0.2451    0.7257    YES
##
## $Descriptives
##    n Mean   Std.Dev Median Min Max 25th  75th      Skew   Kurtosis
## x 20 0.18 0.1361114    0.1 0.0 0.5 0.10 0.225 0.8185140 -0.3698838
## y 20 5.04 1.0054588    5.0 3.3 6.7 4.35 5.850 0.1357527 -1.2067384
```

A. Interpreta el gráfico de QQ-Plot. Indica: ¿qué semejanzas hay entre este gráfico y el caso univariado? ¿qué diferencias?

En el caso univariado, se tiene una gráfica que compara los quartiles de la data con los cuartiles teóricos. Mientras más alineados esten a la linea, más confianza se tiene de que los datos formen parte de una distribución normal. En este caso podemos observar paralelos en como se ve la gráfica de QQ-Plot multivariada con el caso

univariado, con la diferencia de que ahora se compara la distancia de Mahalanobis con un valor de Chi cuadrada, la linea siendo los valores esperados en el caso de ser una distribución normal.

B. Interpreta los valores p de los resultados correspondientes a Mardia Skewness y Mardia Kurtosis. Recuerde que para el caso de Normalidad Ho: Los datos se distribuyen normalmente, H1: Los datos no se distribuyen normalmente. Observa las significancias univariadas y multivariadas.

Segun los resultados obtenidos, el sesgo de los datos tiene un valor de p de 0.463, por lo que no se tiene evidencia estadística para rechazar la hipótesis nula. Es decir, no se puede negar que el sesgo presente en los datos corresponde a lo esperado de una distribución normal. En el caso de la curtosis, se tiene un valor de p de 0.151, nuevamente sin poder rechazar la hipótesis nula y sin la posibilidad de negar que la curtosis es lo esperado de una distribución normal. Estos resultados indican que la distribución conjunta de los datos parece ser normal.

Por otro lado, haciendo un análisis de univariado para cada variable muestra que el valor de p para la variable x es de 0.0024. Con este valor, se rechaza la hipótesis nula y se concluye que la se tiene evidencia estadística para afirmar que la distribución univariada de x no es univariada. En el caso de la variable y, su valor de p es de 0.7257, indicando que no se tiene evidencia estadística para rechazar la hipótesis nula. Por lo tanto, se puede argumentar que la distribución univariada de y es normal.

C. Concluye sobre la normalidad multivariada de la matriz X y la normalidad univariada de cada una de las variables que componen la matriz X.

La matriz X muestra evidencia de seguir una distribución normal multivariada, ya que las pruebas de Mardia para asimetría y curtosis no indican desviaciones significativas de la normalidad. Sin embargo, cuando se examina cada variable individualmente, se observa que la variable x no sigue una distribución normal, mientras que la variable y sí lo hace. Por lo tanto, aunque los datos parecen ser normales en un contexto multivariado, hay una falta de normalidad univariada en la variable x. Bien si las pruebas indican normalidad multivariada, un modelo normal multivariado tiene como supuestos que sus variables sean normales, por lo cual sería más confiable decir que el conjunto como distribución multivariada no es normal.