Tarea2 Análisis multivariado

José Antonio García Ramirez

February 15, 2018

1. Eejercicio 1:

Sea $x \sim N_3(\mu, \Sigma)$, donde

$$\mu = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} \quad y \quad \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

1.1 Encuentra la distribución de $3x_1 - 2x_2 + x_3$

Es un resultado visto en clase que si $x \sim N_n(\mu, \Sigma)$ entonces para un vector fijo a se tiene que $a^t x$ genera una distribución normal con distribución $N_1(a^t\mu, a^t\Sigma a)$ Por lo que si tomamos $a = (3, -2, 1)^t)$\$, entonces $a^t x = 3x_1 - 2x_2 + x_3$ que se distribuye $N_1(a^t\mu, a^t\Sigma a) = N_1((3, -2, 1)^t(2, -3, 1), (3, -2, 1)^t(2, -3, 1) = N_1(13, (3, -2, 1)^t(2, 1, 1)) = N_1(13, 9)$

1.2 Reetiqueta las variables si es necesario, y encuentra un vector 2×1, tal que x_2 y $x_2 - a^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ son independientes.

Fijemos $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, al considerar el vector $x_2 - a^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1 - a_2)x_2 - a_1 x_1$, y encontrar condiciones sobre los a_i para que los vectores sean ortogonales se tendrá la independencia.

Entonces aplicamos el método de Gram-Schmidt para ortogonalizar vectores, como buscamos una dirección podemos dividir $(1 - a_2)x_2 - a_1 x_1$ entre $-a_1$ y comenzamos la ortogonalización: $a_1 =< x_2, x_2 >$ y $a_2 = 1 - < x_1, x_2 >$ Entonces al hacer el producto tenemos que:

$$\left\langle x_2, x_2 - a^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\rangle =< x_2, x_2 - a_1 x_1 - a_2 x_2 >==< x_2, x_2 > - < x_2, x_2 >< x_2, x_1 > -(1 - < x_1, x_2 >) < x_2, x_2 >= 0$$

$\Rightarrow x_2$ y $x_2 - a^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ son ortogonales e independientes.

2. Ejercicio 2:

Encuentra los estimadores de máxima verosimilitud del vector 2×1 de medias μ y de la matriz 2×2 de covarianzas Σ a partir de la m. a.

$$X = \begin{pmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{pmatrix}$$

Obtenida de una población normal bivariada.

En clase, vimos un teaser de la prueba acerca de que la media muestral y la varianza muestral son estimadores maximoverosimiles, por lo que:

$$\bar{X} = \hat{\mu} \begin{pmatrix} 4 \\ 6 \end{pmatrix}$$

y

$$\hat{\Sigma} = \frac{n-1}{n} S = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{2} \end{pmatrix}$$

3. Ejercicio 3:

Sea $x \sim N_3(\mu, \Sigma)$, donde

$$\mu = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

y

$$\Sigma = \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

. Determina la distribución condicional de X_1 dado que $X_2 = x_2$ y $X_3 = x_3$.

Por definición

$$f(X_1|X_2 = x_2, X_3 = x_3) = \frac{f_{X_1X_2X_3}(x_1, x_2, x_3)}{f_{X_2X_3}(x_2, x_3)}$$

El caso de dimensión tres puede verse como el caso anterior particionando la matrix de manera correcta, como vimos en clase que las marginales y las condicionales de una normal bivariada son normales, entonces

$$\mu_{1|23} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = 1 + (0 \ \ -1)\begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x_2 + 1 \\ x_3 - 2 \end{pmatrix} = 2 - \frac{x_3}{2} \quad y$$

$$\Sigma_{1|23} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 4 - (0 \ \ -1)\begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 0 \\ -1 \end{pmatrix} = 4 - \frac{1}{2} = \frac{7}{2}$$

Así $f_{X_1|X_2X_3} \sim N_1\left(2 - \frac{x_3}{2}, \frac{7}{2}\right)$

4. Ejercicio 4:

Realiza las siguientes operaciones para una normal bivariada:

4.1 Escribe una función que de como resultado la probabilidad asociada a un par de valores x,y considerando que estos se comportan como una normal bivariada (utiliza como parámetros adicionales, las medias μ_1,μ_2,σ_1,σ_2,ρ). La función que implemente es la siguiente:

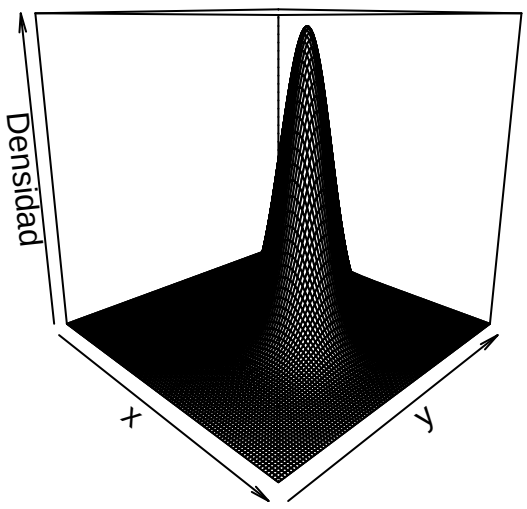
```
bi <- function(x, y, mu.1 = 0, mu.2 = 1, sigma.1 = 1 ,sigma.2 = 1, rho)
{
  #fijo los parametros para casos default y para hacer el siguiente ejercicio más facil
  sin.x <- 2*pi*sqrt(sigma.1*sigma.2*(1-rho**2))
  exp <- ( -1/ (2*(1-rho^2)) )* ( ( (x-mu.1)**2/ sigma.1) + ((y-mu.2)^2/sigma.2) -
          2*rho*( ( (x-mu.1)/sqrt(sigma.1))*( (y-mu.2)/sqrt(sigma.2))) )
  p <- (1/sin.x)*exp(exp)
  return(p)
}
```

Y testeo la función

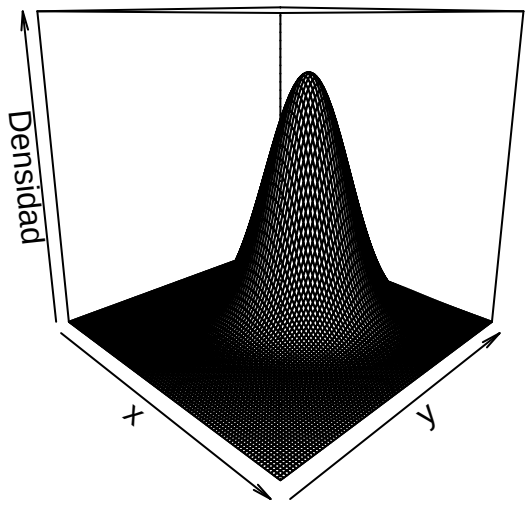
```
mu.1 = 0
mu.2 = 0
sigma.1 <- 1
sigma.2 <- 1
rho <- 0
x <- rnorm( 1, mu.1, sigma.1 )
y <- rnorm(1, mu.1, sigma.2)
bi(x,y, mu.1 , mu.2, sigma.1, sigma.2, rho ) #caso maximo
```

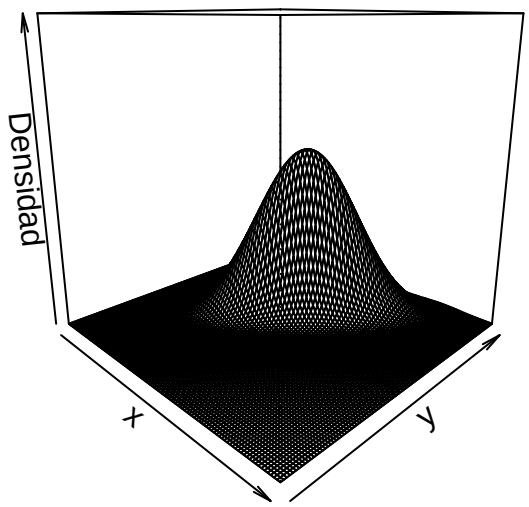
```
## [1] 0.1189183
```

4.2 Realiza gráficas de la densidad sobre una región de valores $x \in [-4, 4]$, $y \in [-4, 4]$ (utiliza la función pers()), y verifica los efectos que tiene variar los valores del parámetro ρ


rho=−0.5637471312657
Densidad
x
y

rho=0.229097998701036





Como lo discutimos en clase, cuando la correlación aumenta en valor absoluto se nota que se forman elipses y el punto medio de la distribución, su valor esperado, tienda a tener una probabilidad mayor si ρ se aproxima a cero, además cuando la correlación cambia de negativo a positivo la elipse refleja sus ejes, "invirtiéndola"

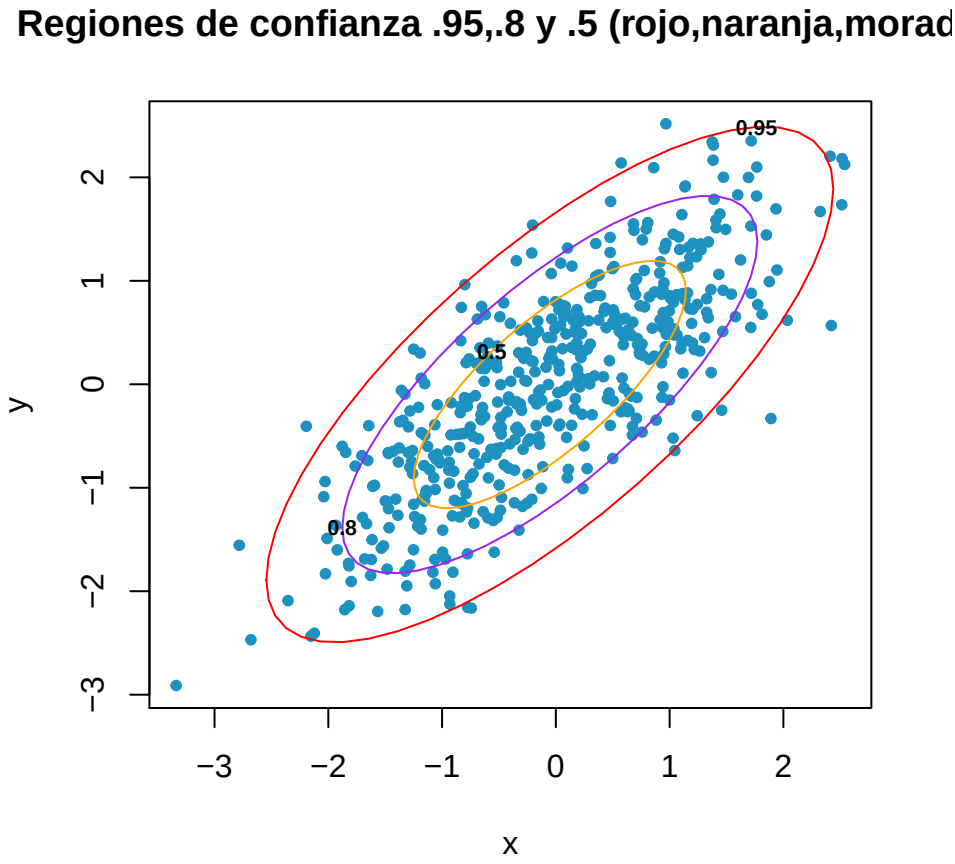
5. Ejercicio 5:

Obten lo siguiente a partir de una muestra de una normal bivariada. Para el muestreo genera 500 observaciones utilizando la función *mrvnorm* de la libreria 'MASS', con $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ y $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$.

5.1 Calcula los elipsoides de confianza del $95, 80$ y 50 con la función que se presenta a continuación y gráficalos sobre los datos

```
confelli<-function(b, C , df, level=0.95, xlab="",ylab="",add=T,prec=51, color)
{
  d <- sqrt(diag(C))
  dfvec <- c(2,df)
  phase <- acos( C[1,2]/(d[1]*d[2]) )
  angles <- seq(-pi, pi, len=prec)
  mult <- sqrt( dfvec[1]*qf(level,dfvec[1],dfvec[2]) )
  xpts <- b[1] + d[1]*mult*cos(angles)
  ypts <- b[2] + d[2]*mult*cos(angles+phase)
  if(add) lines(xpts,ypts, col = color)
    else plot(xpts,ypts,type="l",xlab=xlab,ylab=ylab, col = color)
  a<-round(runif(1,1,51))
  text(xpts[a],ypts[a],paste(level),adj=c(0.5,0.5), font=2,cex=0.7)
}
```

```r
library(MASS)
S <- matrix(c(1,0.8,0.8,1), ncol=2, nrow = 2) #matriz de varianza
mu <- c(0,0)
m.a <- mvrnorm(n=500, mu = mu, Sigma = S)
        #realizamos el inciso b
media <- apply(m.a, 2, mean )
s.1 <- apply(m.a, 2, var) #estimamos la varianza
s.2 <- var( m.a[,1], m.a[,2])
s <- matrix(c(s.1[1], s.2, s.2, s.1[1]), byrow = TRUE, ncol = 2)
    #graficamos
plot(m.a, col="#1D91C0", alpha = 0.1, pch=20, main="Regiones de confianza .95,.8 y .5 (rojo,naranja,mora
     xlab="x", ylab="y")
confelli( b= media, C=s, df=498, col ='red') #porque solo estimamos 2 parametros
confelli(b=media, C=s, df=498, level=0.80, col = 'purple')
confelli(b=media, C=s, df=498, level=0.50, col= 'orange')
```



5.2 Obten la media la matriz de covarianzas de los datos generados en el muestreo.

Esto se calculo en el inciso anterior a continuación muestro explícitamente los valores de la media y de la matriz de covarianzas.

```r
media #media de la uestra generada
```

```
## [1] -0.052851210 -0.002253739
```

```r
s # varianza de la muestra generada
```

```
##           [,1]      [,2]
## [1,] 1.0301981 0.7812592
## [2,] 0.7812592 1.0301981
```

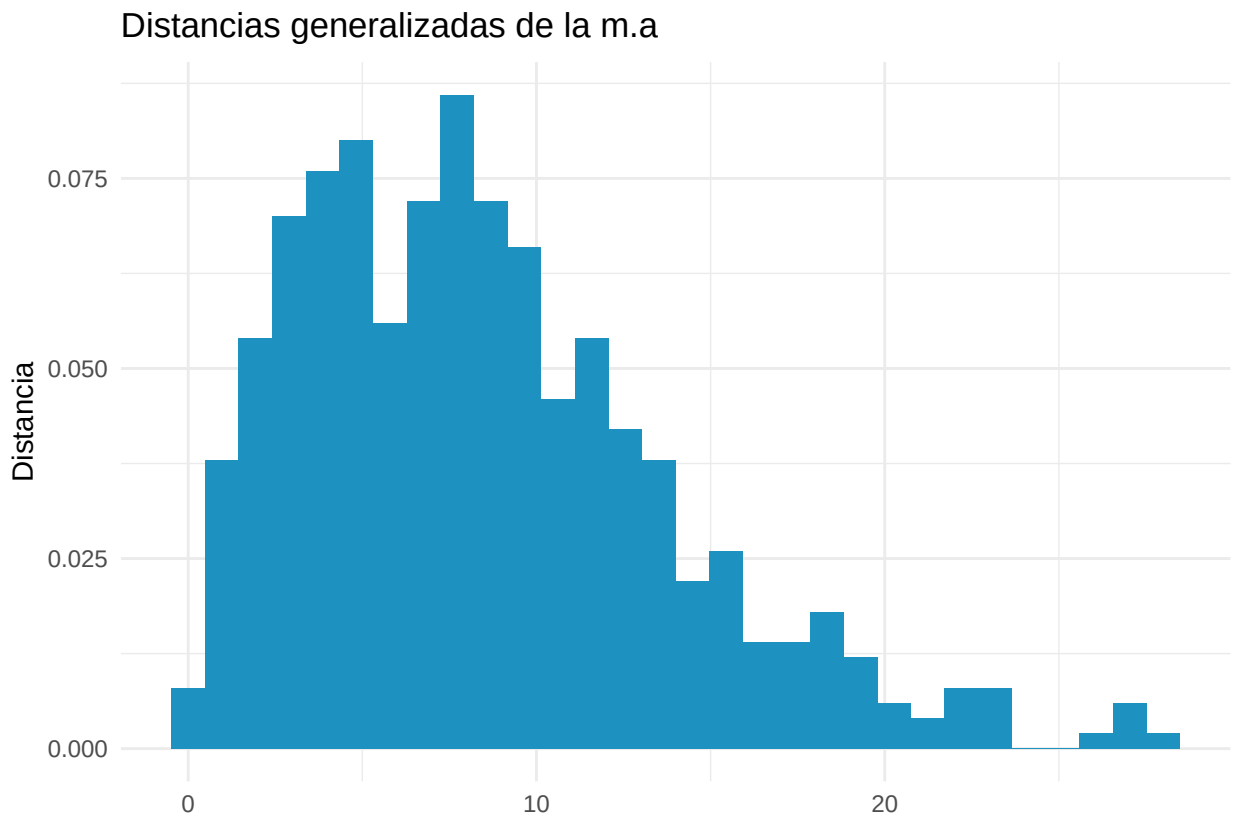
5.3 Calcula la distancia generalizada.

Este es el código que utilice para calcular las distancias generalizadas, de los puntos de la muestra.

```r
dist.gen <- function(x)
{
  # x: dataframe con la muestra
  media <- apply(x, 1, mean) #medias
  s_1 <- solve(var(x))   #varianzas
  centrados <- data.frame(x1= x[, 1]-media[1],x2= x[,2]- media[2]) #centro la muestra
  d.gen <- vector(mode='numeric' , length= dim(x)[1])
  for( i in 1:dim(x)[1])
  {
    d.gen[i] <- as.matrix(centrados[i, ])%*%s_1%*%t(centrados[i,])
  }
  return(d.gen)
}
distancias <- dist.gen(m.a)
```

5.4 Construye un histograma con las distancias

```r
distancias <- data.frame(dist = distancias)
library(ggplot2)
ggplot(distancias, aes(dist, fill =I('#1D91C0'))) +
  geom_histogram(aes(y=..count../sum(..count..))) +
  theme_minimal() + xlab('') +ylab('Distancia') + ggtitle('Distancias generalizadas de la m.a')
```



5.5 Prueba si las distancias son normales mediante la prueba de Shapiro-Wilks

De la salida siguiente vemos que la prueba arroja un $p-value$ cercano a cero, en particular la documentación de la prueba en R no marca el nivel de confianza de la prueba, pero en general podemos decir que la probabilidad de cometer un error de tipo 1 al rechazar la hipótesis nula es bajo dado que es cierta, es bajo por lo que las distancias no son normales, de hecho, en clase vimos que se distribuyen como una ji-cuadrada.

```
shapiro.test(distancias$dist)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  distancias$dist
## W = 0.94438, p-value = 9.55e-13
```

6. Ejercicio 6:

Existen distribuciones cuyas marginales son normales, pero la distribución conjunta no es normal. Considera la siguiente función que genera n pares bivariados.

```
checker <- function(n)    # Genera n pares marginales normales
  # que no son normales bivariadas
{
  checker <- NULL         # comienza una lista
  for (i in 1:n)
  {
    x <- rnorm(2)       # par de normales independientes
    if(x[1]>0)  x[2] <-  abs(x[2])
    else     x[2] <- -abs(x[2])
```

```
    checker <- rbind(checker, x)
  }
  checker
}
```

6.1 Use la función anterior y genera un número grande de pares bivariados. Grafica las columnas individuales y usa algún método para probar si estas columnas están normalmente distribuidas.

```
m.a <- as.data.frame(checker(1000))
shapiro.test(m.a$V1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m.a$V1
## W = 0.99884, p-value = 0.7832
```

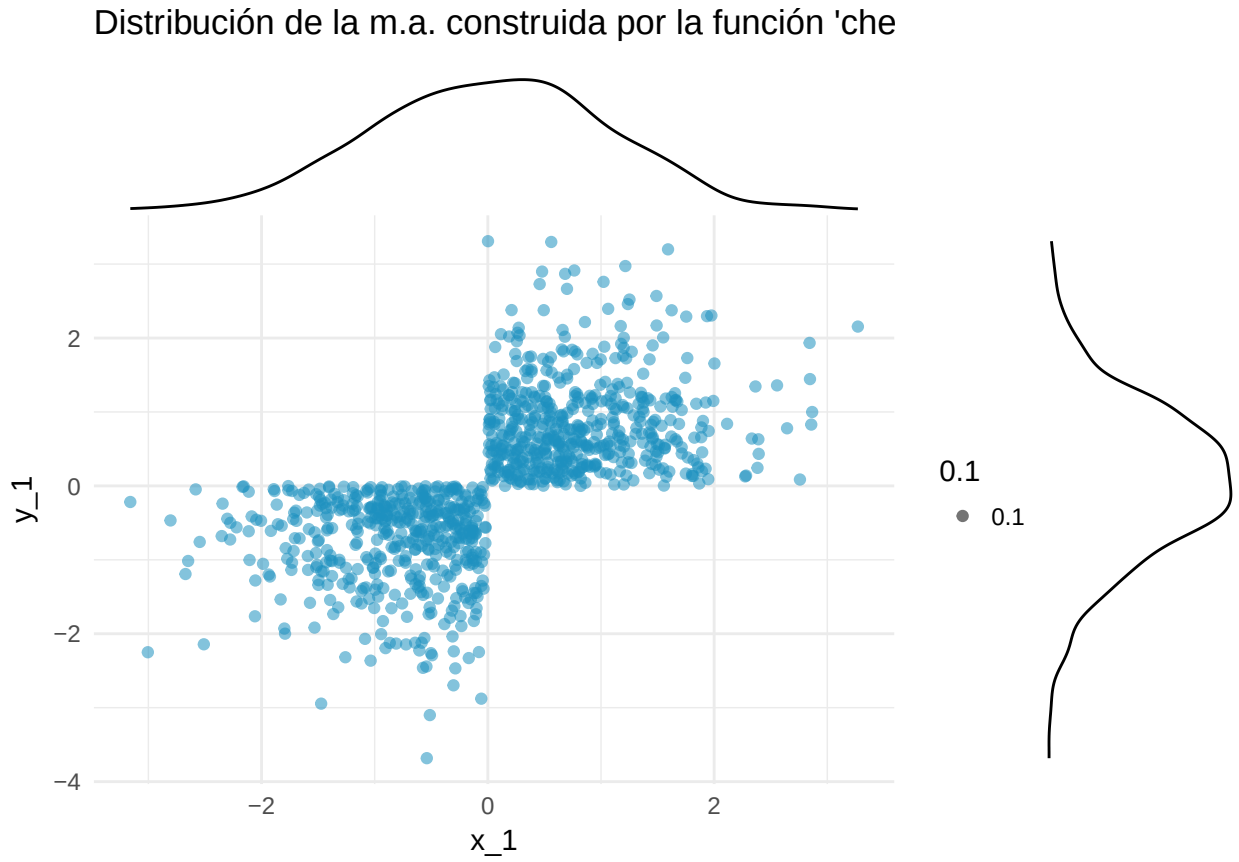
```
shapiro.test(m.a$V2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m.a$V2
## W = 0.99799, p-value = 0.2769
```

De manera análoga al inciso anterior utilice la prueba Shapiro Wilks, De la salida anterior vemos que la prueba arroja un $p - value$ cercano mayor a 0.05 para ambas variables por lo que hay evidencia para pensar que se distribuyen como normales, de hecho, en clase vimos que se distribuyen como una ji-cuadrada. Agregue la prueba de Kolmogorov para saber el nivel de significancia de la prueba

6.2 Grafica los pares bivariados para demostrar que los datos no siguen una distribución normal bivariada.

```
library(ggExtra)
p <-ggplot(m.a, aes(x=V1, y=V2, color = I('#1D91C0'), alpha = .1 )) + geom_point()+
  guides(fill=FALSE)+ xlab('x_1') + ylab('y_1')+
  ggtitle("Distribución de la m.a. construida por la función 'cheker'") +
  theme_minimal()
ggMarginal(p, type = "density", margins = "both", size = 4, marginCol = I('#1D91C0'))
```



Y sorprendentemente es fácil notar que la muestra no es normal, aunque sus marginales lo son, por que por ejemplo no muestra valores en los sectores 2 y 3 del plano, aunque es simétrica, y dista de mostrar gorma elíptica.