Maestría en Computo Estadístico
Estadística Multivariada
Tarea 3
18 de marzo de 2021
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 3, EM.

Instrucciones

Subirla a la plataforma en un zip que contenga el código y el archivo pdf con los resultados.

1. La matriz de datos para un muestra aleatoria de tamaño $n = 3$ de una población normal bivariada está dada por:

$$\mathbf{X} = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix}.$$

a) Verifica que T^2 permanece sin cambios si cada observación $\mathbf{x}_j,\ j = 1, 2, 3$ es reemplazada por $\mathbf{C}\mathbf{x}_j$, donde

$$\mathbf{C} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

$$\mathbf{C}\mathbf{x}_j = \begin{pmatrix} x_{j1} - x_{j2} \\ x_{j1} + x_{j2} \end{pmatrix}.$$

Producen la matriz

$$\begin{pmatrix} (6-9) & (10-6) & (8-3) \\ (6+9) & (10+6) & (8+3) \end{pmatrix}$$

RESPUESTA
Calculemos primero T^2 de la matriz de datos original, recordemos que se define como

$$T^2 = n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu).$$

Entonces con ayuda de R, calculemos el vector $\bar{x}$ y la matriz de covarianzas S.

```
X = matrix(c(6,10,8,9,6,3), 3) # datos del problema
mu = colMeans(X) # vector de medias
mu
```

```
## [1] 8 6
```

```
P = diag(c(1,1,1))-matrix(1, nrow=3, ncol=3)/3 # matriz de centrado
S = t(X)%*%P%*%X/2 # matriz de covarianzas muestral
S
```

```
##      [,1] [,2]
## [1,]    4   -3
## [2,]   -3    9
```

Entonces tenemos que $\bar{x} = (8, 6)$, y la matriz de covarianzas es $S = \frac{1}{n} X'PX = \begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix}$. Lo que implica que la inversa de la matriz de covarianzas es $S^{-1} = \frac{1}{27} \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix}$. **Por lo que,**

$$
\begin{aligned}
T^2 &= 3 \begin{pmatrix} 8-\mu_1 & 6-\mu_2 \end{pmatrix} \begin{pmatrix} 1/3 & 1/9 \\ 1/9 & 4/27 \end{pmatrix} \begin{pmatrix} 8-\mu_1 \\ 6-\mu_2 \end{pmatrix} \\
&= \begin{pmatrix} 8-\mu_1 & 6-\mu_2 \end{pmatrix} \begin{pmatrix} 8-\mu_1 + (6-\mu_2)/3 \\ (8-\mu_1)/3 + 4(6-\mu_2)/9 \end{pmatrix} \\
&= \begin{pmatrix} 8-\mu_1 & 6-\mu_2 \end{pmatrix} \begin{pmatrix} 8-\mu_1 + 2 - \mu_2/3 \\ -\mu_1/3 + 16/3 - 4\mu_2/9 \end{pmatrix} \\
&= (8-\mu_1)(8-\mu_1+2-\mu_2/3) + (6-\mu_2)(-\mu_1/3 + 16/3 - 4\mu_2/9) \\
&= 64 - 8\mu_1 + 16 - 8\mu_2/3 - 8\mu_1 + \mu_1^2 - 2\mu_1 + \mu_1\mu_2/3 - 2\mu_1 + 32 - 8\mu_2/3 + \mu_1\mu_2/3 - 16\mu_2/3 + 4\mu_2^2/9 \\
&= 112 - 20\mu_1 - 32\mu_2/3 + 2\mu_1\mu_2/3 + \mu_1^2 + 4\mu_2^2/9. \quad \blacksquare
\end{aligned}
$$

Ahora calculemos T^2 utilizando la transformación con la matriz $\mathbf{C}$, para ello calculemos la matriz nueva, el vector de medias y la matriz de covarianzas de esta nueva matriz.

```
C = matrix(c(1, 1, -1, 1), 2) # matriz C
X_new = t(C%*%t(X)) # transformación de los datos
X_new
```

```
##      [,1] [,2]
## [1,]   -3   15
## [2,]    4   16
## [3,]    5   11
```

```
mu_new = colMeans(X_new) # vector de medias
mu_new
```

```
## [1]  2 14
```

```
S_new = t(X_new)%*%P%*%X_new/2 # matriz de covarianza muestral para los datos transf.
S_new
```

```
##      [,1] [,2]
## [1,]   19   -5
## [2,]   -5    7
```

Entonces tenemos que $\bar{x}_{new} = \bar{x} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 + \mu_2 \end{pmatrix} = (2, 14)$, y la matriz de covarianzas es $S_{new} = \frac{1}{n} X'PX = \begin{pmatrix} 19 & -5 \\ -5 & 7 \end{pmatrix}$. Lo que implica que la inversa de la matriz de covarianzas es $S^{-1} = \frac{1}{108} \begin{pmatrix} 7 & 5 \\ 5 & 19 \end{pmatrix}$.

Por lo que,

$$
\begin{aligned}
T^2 &= \frac{3}{108}\begin{pmatrix}2-(\mu_1-\mu_2) & 14-(\mu_1+\mu_2)\end{pmatrix}\begin{pmatrix}7 & 5\\ 5 & 19\end{pmatrix}\begin{pmatrix}2-(\mu_1-\mu_2)\\ 14-(\mu_1+\mu_2)\end{pmatrix}\\
&= \frac{1}{36}\begin{pmatrix}2-(\mu_1-\mu_2) & 14-(\mu_1+\mu_2)\end{pmatrix}\begin{pmatrix}14-7\mu_1+7\mu_2+70-5\mu_1-5\mu_2\\ 10-5\mu_1+5\mu_2+266-19\mu_1-19\mu_2\end{pmatrix}\\
&= \frac{1}{36}\begin{pmatrix}2-(\mu_1-\mu_2) & 14-(\mu_1+\mu_2)\end{pmatrix}\begin{pmatrix}84-12\mu_1+2\mu_2\\ 276-24\mu_1-14\mu_2\end{pmatrix}\\
&= \frac{1}{36}\left((2-(\mu_1-\mu_2))(84-12\mu_1+2\mu_2)+(14-(\mu_1+\mu_2))(276-24\mu_1-14\mu_2)\right)\\
&= \frac{1}{36}(168-108\mu_1+88\mu_2-14\mu_1\mu_2+12\mu_1^2+2\mu_2^2+3864-612\mu_1-472\mu_2+38\mu_1\mu_2+24\mu_1^2+14\mu_2^2)\\
&= \frac{1}{36}(4032-720\mu_1-384\mu_2+24\mu_1\mu_2+36\mu_1^2+16\mu_2^2)\\
&= 112-20\mu_1-32\mu_2/3+2\mu_1\mu_2/3+\mu_1^2+4\mu_2^2/9. \ \blacksquare
\end{aligned}
$$

2. Dadas la siguiente muestra de observaciones bivariadas:

$$\mathbf{X} = \begin{pmatrix}2 & 12\\ 8 & 9\\ 6 & 9\\ 8 & 10\end{pmatrix}$$

a) Evalua T^2, para probar $H_0 : \mu = [7, 11]$, usando los datos.

RESPUESTA

Teorema: 1 *(Visto en clase, pag. 20) Una generalización natural de la distancia cuadrada univariada t es su análago multivariado T^2 de Hotelling*

$$T^2 = n(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0)$$

donde

$$\bar{x} = \frac{1}{n}\mathbb{X}\mathbf{1}, \mu_0 = \begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{pmatrix}.$$

Y $n^{-1}S$ es la matrix de covariazas estimada de $\bar{x}$. Esto nos da un marco de referencia para probar hipótesis sobre el vector de medias, donde las hipótesis nula y alternativa son

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0.$$

Y con el estadístico T^2. Con decisión de rechazo: no rechazar H_0 si $T^2 \leq \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$, de otra manera se rechaza H_0.

Con ayuda de R, calculamos el estadístico T^2

```r
X = matrix(c(2,8,6,8,12,9,9,10), nrow=4) # datos del problema

mu = colMeans(X) # vector de medias
mu
```

```
## [1]  6 10
```

```r
mu_0 = matrix(c(7,11), nrow=1) # media a probar

P = diag(c(1,1,1,1))-matrix(1, nrow=4, ncol=4)/4
S = t(X)%*%P%*%X/3 # matriz de covarianzas muestras.
S
```

```
##           [,1]      [,2]
## [1,]  8.000000 -3.333333
## [2,] -3.333333  2.000000
```

```r
T2 = 4*(mu-mu_0)%*%solve(S)%*%t(mu-mu_0) # estadístico T2
T2
```

```
##          [,1]
## [1,] 13.63636
```

Es decir, **el estadístico T^2 =13.6363636.**

b) Especifica la distribución de T^2 (verificando la normalidad de los datos).

RESPUESTA

Ocupemos la proba de Shapiro−Wilks para probar normalidad univariada.

```r
# var x_1
shapiro.test(X[,1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X[, 1]
## W = 0.82743, p-value = 0.1612
```

```r
# var x_2
shapiro.test(X[,2])
```

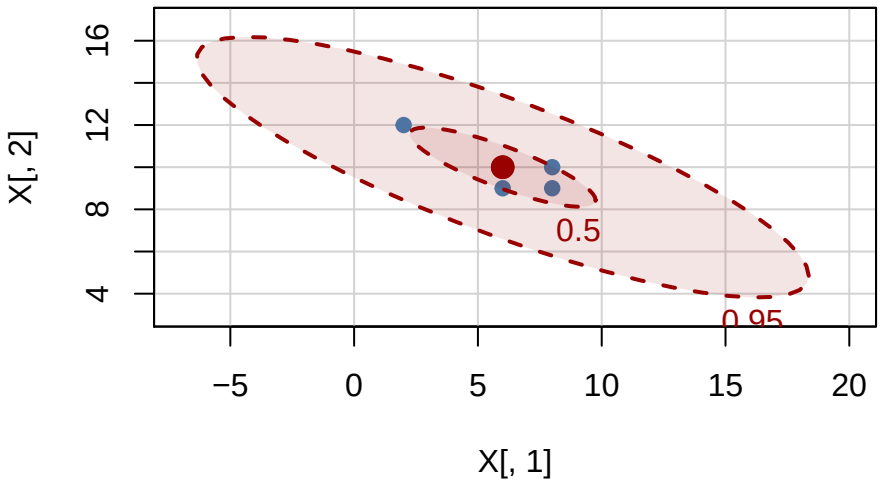
```
##
##  Shapiro-Wilk normality test
##
## data:  X[, 2]
## W = 0.82743, p-value = 0.1612
```

Observemos que en ambas pruebas el p−value es mayor que 0.05, por lo que implica que no hay evidencia significativa de no rechazar la hipótesis nula. Por lo que podemos concluir que que ambas variables se distribuyen normal. Ahora probemos que la tienen una distribución bivariada, para ello ocuparemos los diagramas de dispersión.

```r
library(car) # dataEllipse()
```

```
## Loading required package: carData
```

```r
dataEllipse(X[,1], X[,2], xlim=c(-7,20), ylim=c(3,17), pch = 19, col = c("steelblue", "#990000"),
lty = 2, ellipse.label = c(0.5, 0.95), levels = c(0.5,
0.95), fill = TRUE, fill.alpha = 0.1) # gráficamos la elipse
```



Observando el diagrama de dispersión, podríamos concluir que si se distribuye como una normal bivariada. Entonces como $\mathbf{x} \sim N_p(\mu, \sigma)$, esto implica que $\hat{\mathbb{X}}$ se distribuya $N_p(\mu, \Sigma/n)$. **Ocupando la diapositiva 13 (semana 3), tenemos que si $\hat{\mathbb{X}} \sim N_p(\mu, \Sigma/n)$ entonces el estadístico T^2 se distribuye $T^2(p, n-1)$ (distribución T^2 de Hotelling).** Si tuvieramos una muestra aleatoria grande, ocupando el teorema del límite central T^2 converge a la distribución χ^2_p, pero no es el caso de este ejercicio.

c) Usando (a) y (b) prueba H_0 en $\alpha = ,05$ ¿Que conclusión se tiene?

RESPUESTA
Calculemos el valor critico.

```r
n = 4 # datos del probleema
p = 2
F_valor = (n-1)*p*(qf(0.95, p, n-2))/(n-p) # valor crítico
F_valor
```

```
## [1] 57
```

Entonces como $T^2 = 13.6363636 \leq 57 = 3F_{2,2}(0{,}5)$,con 95 % de confianza no podemos rechazar la hipótesis nula, y **por lo tanto podemos concluir que la media poblacional es** $\mu = [7, 11]$.

d) Evalue T^2 utilizando la relación que tiene con la lambda de Wilks.

RESPUESTA

Teorema: 2 *(Visto en clase, pag. 48-semana 4) Suponga $\mathbf{x}_1, \cdots, \mathbf{x}_n$ es una muestra aleatoria de una población $N_p(\mu, \Sigma)$. Entonces la prueba basada en T^2 es equivalente a la prueba de la razón de verosimilitud para probar $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$ debido a que*

$$\Lambda^{n/2} = \left(\frac{|\hat{\Sigma}|}{\hat{\Sigma}_0}\right) = \left(1 + \frac{T^2}{n-1}\right)^{-1} \tag{1}$$

. De la ecuación anterior, T^2 se puede calcular como el cociente de dos determinantes. Resolviendo

para T^2 se obtiene

$$T^2 = \frac{(n-1)|\hat{\Sigma}_0|}{|\hat{\Sigma}|} - (n-1)$$

Con ayuda de R calculamos T^2 con la ecuación 1,

```
mu_0 <- as.vector(mu_0) # media a constractar

hat_Sigma_0 <- (t(X)-mu_0)%*%t(t(X)-mu_0) # matriz de covarianzas bajo H_0
hat_Sigma_0
```

```
##      [,1] [,2]
## [1,]   28   -6
## [2,]   -6   10
```

```
hat_Sigma <- (t(X)-mu)%*%t(t(X)-mu) # matriz de covarianzas
hat_Sigma
```

```
##      [,1] [,2]
## [1,]   24  -10
## [2,]  -10    6
```

```
T2_new <- (n-1)*det(hat_Sigma_0)/det(hat_Sigma)-3 # estadístico
T2_new
```

```
## [1] 13.63636
```

Por lo que podemos concluir nuevamente que T^2 =13.6363636.

e) Evalue Λ y la lambda de Wilks.

RESPUESTA

Teorema: 3 *(Visto en clase, pag. 46-semana 4) El estadístico de la Razón de Verosimilitud:*

$$\Lambda = \left(\frac{|\hat{\Sigma}|}{\hat{\Sigma}_0}\right)^{n/2}.$$

O de manera equivalente

$$\Lambda^{n/2} = \frac{|\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})'|}{|\sum_{j=1}^{n}(x_j - \mu_0)(x_j - \mu_0)'|}$$

Ocupando los resultados del inciso anterior tenemos,

```
Lambda = (det(hat_Sigma)/det(hat_Sigma_0))**2 # estadístico de razon de verosi.
Lambda
```

```
## [1] 0.03251814
```

Es decir, $\Lambda = 0.0325181$. Ahora calculemos el valor crítico de la razón de velorsimilitud

```
Lambda_valor_critico <- (1+F_valor/(n-1))**(-1)
Lambda_valor_critico**2
```

```
## [1] 0.0025
```

La decisión de rechazo es: no rechazar $H_0 si \Lambda \geq 0{,}0025$. Vemos que $\Lambda = 0{,}0325 \geq 0{,}0025$, por lo que no hay evidencia significativa para rechazar H_0 con una confianza del 95 %, es decir, **podemos concluir nuevamente que la media poblacional es** $\mu = [7, 11]$. ■.

3. El departamento de control de calidad de un fabricante de hornos de mi roondas es requerido por el gobierno federal para monitorear la antidad de radia ión emitida por los hornos que fabrican. Se realizaron mediciones de la radiación emitida por 42 hornos sele ionados al azar on las puertas erradas y abiertas. Los datos están en el archivo **datosradiacion.**

a) Construye un elipse de confianza del 95 % para μ, considerando la transformación de las variables:

$$x_1 = \sqrt[4]{\text{mediciones de la radiación con puerta cerrada}}$$
$$x_2 = \sqrt[4]{\text{mediciones de la radiación con puerta abierta}}$$

RESPUESTA

Teorema: 4 *(Visto en clase, pag. 62-semana 4) Recordando que el estadístico para probar* $H_0 : \mu = \mu_0$ *está dado por*

$$T^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0)$$

No se rechaza H_0 *si* $T^2 \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$*, es decir,*

$$n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Por tanto, la región de confianza para μ *de una población normal* $p-$*variada está dado por*

$$\mathbb{P}\left(n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) \right) = 1 - \alpha.$$

Más formalmente, una región de confianza $R(\mathbb{X})$ *del* $100(1-\alpha)\,\%$ *para el vector de medias* μ *de una distribución normal* $p-$*dimensional es el elipsoide determinado por todos los puntos posibles de* μ *que satisfacen*

$$n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Sabemos por el teorema del límite central que en una muestra aleatoria grande con distribución T^2 esta converge en probabilidad a una distribución χ^2_p, como en nuestro conjunto de datos tenemos que $n - p = 40$ es grande podemos traducir la región de confianza $R(\mathbb{X})$ del $100(1-\alpha)\,\%$ para el vector de medias μ de una distribución normal p−dimensional es el elipsoide determinado por todos los puntos posibles de μ que satisfacen

$$n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \leq \frac{n-1}{n-p} \chi^2_p(\alpha)$$

Con ayuda de calculamos el vector de medias $\bar{x}$, la matriz de covarianzas muestral S y el valor crítico $\frac{(n-1)}{(n-p)} \chi^2_p(\alpha)$.

```r
library(dplyr) # manipular data.frame
library(readxl) # leer xls
datos_radiacion <- read_excel("../datos_radiacion.xlsx", col_names = # cargamos los datos
                              c("puerta_cerrada", "puerta_abierta"), skip = 1)

datos_radiacion <- datos_radiacion %>% # transformamos los datos
  mutate(x_1 = as.numeric(puerta_cerrada)**(1/4),
         x_2 = as.numeric(puerta_abierta)**(1/4))

mu <- datos_radiacion %>% select(x_1, x_2) %>% colMeans() %>% as.vector()
mu #vecor de medias
```

```
## [1] 0.5642575 0.6029812
```

```r
n <- nrow(datos_radiacion) # datos del problema
p <- ncol(datos_radiacion)-2

S <- datos_radiacion %>% select(x_1, x_2) %>%  cov() %>% as.matrix()
solve(S) # inversa de la matriz de covarianzas muestral
```

```
##           x_1       x_2
## x_1  203.4981 -163.9069
## x_2 -163.9069  200.7691
```

```r
chi_valor = (n-1)*qchisq(0.95, p)/(n-p) # valor crítico.
chi_valor
```

```
## [1] 6.141251
```

Por lo tanto, nuestra región de confianza $R(\mathbb{X})$ son todos los puntos (μ_1, μ_2) que satisfacen

$$42\begin{pmatrix}0{,}5642575-\mu_1 & 0{,}6029812-\mu_2\end{pmatrix}\begin{pmatrix}203{,}4981 & -163{,}9069\\ -163{,}9069 & 200{,}7691\end{pmatrix}\begin{pmatrix}0{,}5642575-\mu_1\\ 0{,}6029812-\mu_2\end{pmatrix}\leq 6{,}141251$$

$$\begin{pmatrix}0{,}5642575-\mu_1 & 0{,}6029812-\mu_2\end{pmatrix}\begin{pmatrix}114{,}8253-203{,}4981\mu_1-98{,}83279+163{,}9069\mu_2\\ -92{,}4857+163{,}9069\mu_1+121{,}06-200{,}7691\mu_2\end{pmatrix}\leq 0{,}1462203$$

$$\begin{pmatrix}0{,}5642575-\mu_1 & 0{,}6029812-\mu_2\end{pmatrix}\begin{pmatrix}15{,}99251-203{,}4981\mu_1+163{,}9069\mu_2\\ 28{,}5743+163{,}9069\mu_1-200{,}7691\mu_2\end{pmatrix}\leq 0{,}1462203$$

$$\begin{aligned}9{,}023894-114{,}82536\mu_1+92{,}4857\mu_2-15{,}99251\mu_1+203{,}4981\mu_1^2-163{,}9069\mu_1\mu_2+\\ 17{,}22977+98{,}83279\mu_1-121{,}06\mu_2-28{,}5743\mu_2-163{,}9069\mu_2\mu_1+200{,}7691\mu_2^2\leq 0{,}1462203\\ \mathbf{26{,}25367-31{,}98508\mu_1-57{,}14911\mu_2-327{,}8138\mu_1\mu_2+203{,}4981\mu_1^2+200{,}7691\mu_2^2}\leq 0{,}1462203.\end{aligned}$$

Y la elipsoide de confianza al $95\,\%$ de confianza esta dada por la forma cuádratica

$$\mathbf{26{,}10745-31{,}98508\mu_1-57{,}14911\mu_2-327{,}8138\mu_1\mu_2+203{,}4981\mu_1^2+200{,}7691\mu_2^2=0}.$$

b) Prueba si $\mu = (,562, ,589)$ está en la región de confianza.

RESPUESTA

Para probar si $\mu = (,562, ,589)$ evaluemos la desigualdad para ver si se cumple

```
mu_1 = 0.562 # media a comprobar
mu_2 = 0.589
26.25367-31.98508*mu_1-57.14911*mu_2-327.8138*mu_1*mu_2+203.4981*(mu_1**2)+200.7691*(mu_2**2)
```

```
## [1] 0.02963164
```

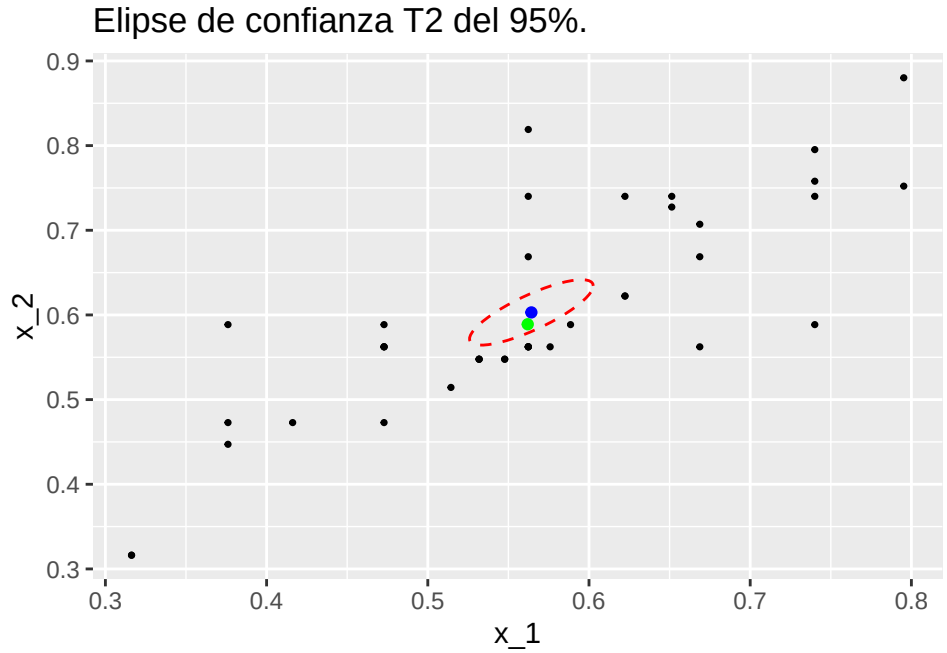
Entonces como la desigualdad 0,02963164 $\leq$ 0,1462203 si se cumple, **entonces** $\mu = (,562, ,589)$ **si se encuentra en la región de confianza**.

c) Calcula los valores y vectores propios de **S** y obten la gráfica del elipsoide de confianza.

RESPUESTA
Con ayuda de R calculamos los vectores y valores propios.

```
mu_new <- c(.562, .589) # media a constractar
eigVal  <- eigen(S)$values # valores y vectores propios
eigVec  <- eigen(S)$vectors
```

```
library(ggplot2) # graficar
ggplot(datos_radiacion, aes(x_1, x_2))+ # grficamos la elipsoide y los datos
  geom_point(size=0.6)+
  stat_ellipse(data=datos_radiacion,aes(x_1, x_2), type="norm", col="red", size=.5,
               linetype=2, level=.05) +# región de confianza al 95%
  geom_point(data=data.frame(x_1=mu[1], x_2=mu[2]), aes(x_1,x_2), col="blue")+
  geom_point(data=data.frame(x_1=mu_new[1], x_2=mu_new[2]), aes(x_1,x_2), col="green")+
  labs(title = "Elipse de confianza T2 del 95%.")
```



Observamos que la media si cae dentro del elipsoide de confianza.

d) Realiza una prueba para la hipótesis $H_0 : \mu = (,55, ,60)$ en un nivel de significacia de $\alpha = 0{,}05$. Es consistente el resultado con la gráfica de la elipse de confianza del 95 % para μ obtenida en el inciso anterior? Explica.

RESPUESTA
Calculemos el estadístico T^2,

```r
X <- as.matrix(datos_radiacion %>% select(x_1, x_2))
mu_0 <- c(0.55, 0.60)

T2 <- n*t(mu-mu_0)%*%solve(S)%*%(mu-mu_0)
T2
```

```
##          [,1]
## [1,] 1.227116
```

```r
chi_valor
```

```
## [1] 6.141251
```

Entonces, como $T^2 = 1.2271163 \leq 6.1412512 = 1{,}025\chi^2_p(\alpha)$ con 95 % de confianza no podemos rechazar la hipótesis nula, y **por lo tanto podemos concluir que la media poblacional es** $\mu = [0{,}55, 0{,}60]$. Y notemos que es consistente al resultado con la gráfica de la elipse de confianza del 95 % para μ, esto se debe a que la correlación de estas variables es muy pequeña (cercana a 0) por lo que no afecta tanto en los dos enfoques. ■.

4. Sabemos que T^2 es igual al $t - valor$ cuadrado univariado más grande construido a partir de la combinación lineal $\mathbf{a}'\mathbf{x}_j$, con $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu_0)$ (ver notas de la semana 5).

a) Usando los resultados del ejercicio anterior y la misma hipótesis nula H_0 del inciso d), evalua $\mathbf{a}$ para los datos transformados de radiaciones de los hornos.

RESPUESTA
Calculamos $\mathbf{a}$ con la ayuda de R,

```r
a <- solve(S)%*%(mu-mu_0)
rownames(a) <- c("Puerta cerrada", "Puerta abierta")
a
```

```
##                      [,1]
## Puerta cerrada  2.412731
## Puerta abierta -1.738364
```

b) Verifica que el valor t^2 calculado con esta $\mathbf{a}$ es igual a la T^2 del ejercicio anterior.

RESPUESTA

Teorema: 5 *(Visto en clase, pag. 7-semana 5) Tenemos que el estadístico* t^2 *se define como*

$$t^2 = \frac{n(\mathbf{a}'\bar{x} - \mathbf{a}'\mu)}{\sqrt{\mathbf{a}'S\mathbf{a}}}.$$

Calculemos t^2 con ayuda de R,

```r
t2 <- n*(t(a)%*%mu-t(a)%*%mu_0)**2/(t(a)%*%S%*%a)
t2
```

```
##          [,1]
## [1,] 1.227116
```

Por lo tanto, **podemos observar que** $t2 =$ **1.2271163** $=$ **1.2271163** $= T^2$.

5. Los datos en el archivo **datososos** representan las longitudes en centímetros de siete osos hembras a los 2, 3, 4 y 5 años de edad.

a) Obtener los intervalos de confianza simultaneos T^2 del 95 % para las cuatro medias poblacionales de la longitud por año.

RESPUESTA

Teorema: 6 *(Visto en clase, pag. 10-semana 5) Sea* $\mathbf{x}_1, \cdots, \mathbf{x}_n$ *una muestra aleatoria obtenida de una población* $N_p(\mu, \Sigma)$ *positiva definida. Entonces, simultáneamente para toda* $\mathbf{a}$*, el intervalo*

$$\left(\mathbf{a}'\bar{\mathbb{X}} - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}, \mathbf{a}'\bar{\mathbb{X}} + \sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}\right)$$

Contendrá $\mathbf{a}'\mu$ *con probabilidad* $1-\alpha$*. Estos intervalos simultáneos se denomina intervalos* T^2*, ya que la probabilidad de cobertura es determinada por la distribución* T^2*. Notese que las elecciones sucesivas de* $\mathbf{a}' = [1000\cdots0], \mathbf{a}' = [0100\cdots0], \cdots, \mathbf{a}' = [000\cdots1]$ *para los intervalos* T^2*, nos permiten obtener los intervalos de confianza para las medias de los componentes,* $\mu_1, \mu_2, \cdots, \mu_p$*, esto es*

$$\begin{aligned}
\bar{x}_1 - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} \leq &\mu_1 \leq \bar{x}_1 + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} \\
\bar{x}_2 - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \leq &\mu_2 \leq \bar{x}_2 + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \\
&\vdots \\
\bar{x}_p - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}} \leq &\mu_p \leq \bar{x}_p + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}}.
\end{aligned}$$

Con lo anterior procedemos a calcularlo en R y suponemos que los datos se distribuyen como una normal multivariada, es decir, cumple los supuestos del Teorema **??**

```
datos_osos <- read_excel("../datos_osos.xlsx") # cargamos los datos
n <- nrow(datos_osos) # datos del problema
p <- ncol(datos_osos)

valor_critico <- sqrt((n-1)*p*(qf(0.95, p, n-p))/(n-p)) # valor critico
mu <- colMeans(datos_osos) # vecotr de medais
mu
```

```
## Longitud2 Longitud3 Longitud4 Longitud5
##  143.2857  159.2857  173.1429  177.1429
```

```
S_mof <- matrix(sqrt(diag(cov(datos_osos))/n),1) # a'Sa

lim_inf <- mu-valor_critico*S_mof # limites de los intervalso de confianza
lim_inf
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 130.6851 127.0216 160.3082 155.3749
```

```
lim_sup <- mu+valor_critico*S_mof
lim_sup
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 155.8863 191.5498 185.9776 198.9108
```

Entonces, los intervalos de confianza para las cuatro medias poblacionales de la longitud por año μ_1, μ_2, μ_3 y μ_4 son

variable	límite inferior	media	límite superior
Longitud2	130.6851024	μ_1 =143.2857143	155.8863261
Longitud3	127.0216268	μ_2 =159.2857143	191.5498017
Longitud4	160.3081582	μ_3 =173.1428571	185.9775561
Longitud5	155.3748671	μ_4 =177.1428571	198.9108472

Cuadro 1: Intervalos de confianza simultaneos T^2 al 95 % de confianza.

b) Respecto al inciso a), obtener los intervalos de confianza simultaneos T^2 del 95 % para los tres aumentos anuales sucesivos en la longitud media.

RESPUESTA

En este problem, queremos contrastar los tres aumentos anuales sucesivos en la longitud media, es decir, $\mu_2 - \mu_1$, $\mu_3 - \mu_2$ y $\mu_4 - \mu_3$. Por lo que consideraremos, $\mathbf{a}' = \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix}, \mathbf{a}' = \begin{pmatrix} 0 & -1 & 1 & 0 \end{pmatrix}$ y $\mathbf{a}' = \begin{pmatrix} 0 & 0 & -1 & 1 \end{pmatrix}$. Entonces utilizando el teorema (6) podemos encontrar los intervalos múltiples para el cambio en la longitud media,

```
S <- cov(datos_osos) # matriz de covarianzas muestral
A <- matrix(c(-1,1,0,0,0,-1,1,0,0,0,-1,1),4) # a'
var_x <- c()
for (i in 1:3){ #calculamos la 'varianza'
  var_x[i] <- sqrt((t(A[,i])%*%S%*%(A[,i]))/n)
}

dif_mu <- diff(mu) # vector de diferencias de medias
dif_mu
```

```
## Longitud3 Longitud4 Longitud5
##  16.00000  13.85714   4.00000
```

```
lim_inf_d <- dif_mu-valor_critico*var_x# limites de los intervalso de confianza
lim_inf_d
```

```
## Longitud3 Longitud4 Longitud5
## -21.22649 -22.73077 -20.65385
```

```
lim_sup_d <- dif_mu+valor_critico*var_x
lim_sup_d
```

```
## Longitud3 Longitud4 Longitud5
##  53.22649  50.44505  28.65385
```

Entonces,los intervalos múltiples para el cambio en la longitud media poblacional para los tres aumentos anuales sucesivos en la longitud media $\mu_2 - \mu_1, \mu_3 - \mu_2$ y $\mu_4 - \mu_3$ son

variable	límite inferior	aumento anual	límite superior
Diferencia 2 a 3 años	-21.2264919	$\mu_2 - \mu_1 = 16$	53.2264919
Diferencia 3 a 4 años	-22.7307684	$\mu_3 - \mu_2 = 13.8571429$	50.4450541
Diferencia 4 a 5 años	-20.6538466	$\mu_4 - \mu_3 = 4$	28.6538466

Cuadro 2: Intervalos de confianza simultaneos T^2 al 95 % de confianza.

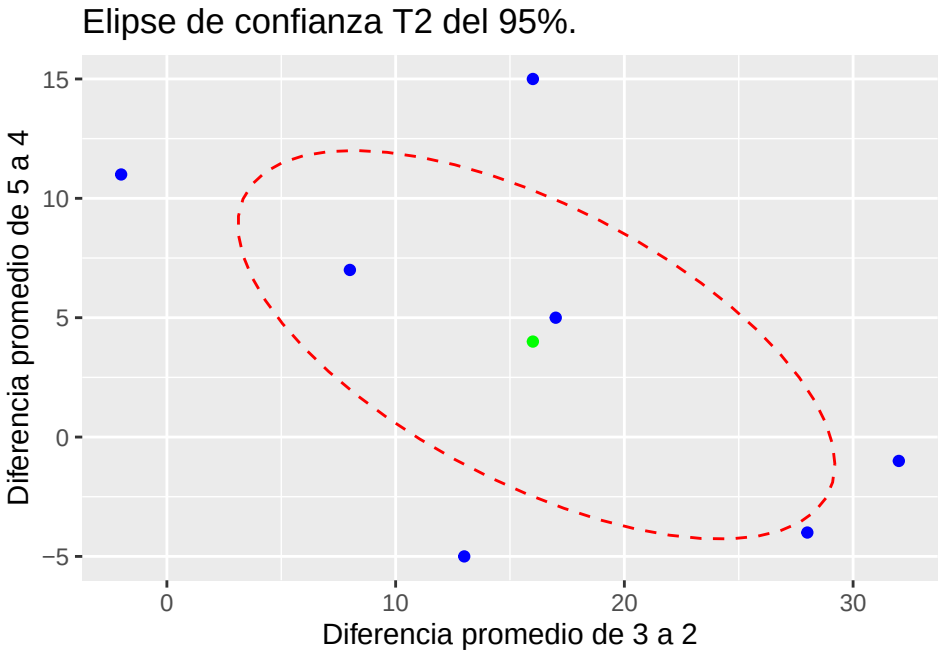
c) Obtener la elipse de confianza T^2 del 95 % para el aumento medio de la longitud de 2 a 3 años y el aumento medio de la longitud de 4 a 5 años.

RESPUESTA

Con ayuda de R procedemos a calcular la elipse de confianza T^2,

```
X <- datos_osos %>%
  mutate(x_1=Longitud3-Longitud2,
         x_2=Longitud5-Longitud4) %>%
  select(x_1, x_2) %>% as.data.frame()

ggplot(X, aes(x_1, x_2))+
  geom_point(col="blue")+
  stat_ellipse(data=X,aes(x_1, x_2), type="t", col="red", size=.5, linetype=2, level=.50)+
  geom_point(data=data.frame(x_1=colMeans(X)[1], x_2=colMeans(X)[2]), aes(x_1,x_2), col="green")+
  labs(x="Diferencia promedio de 3 a 2", y="Diferencia promedio de 5 a 4",
       title = "Elipse de confianza T2 del 95%.")
```



d) Construir los intervalos de confianza de 95 % de Bonferroni para el conjunto formado por las cuatro longitudes medias y los tres aumentos anuales sucesivos en la longitud media, compara los resultados con los obtenidos en a) y b).

RESPUESTA

Teorema: 7 *Con un nivel de confiazna global más grande o igual a* $1-\alpha$*, podemos construir los* $m = p$ *intervalos de confianza de Bonferroni:*

$$\begin{aligned}
\bar{x}_1 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \leq & \mu_1 \leq \bar{x}_1 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \\
\bar{x}_2 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \leq & \mu_2 \leq \bar{x}_2 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \\
& \vdots \\
\bar{x}_p - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}} \leq & \mu_p \leq \bar{x}_p + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}}.
\end{aligned}$$

Utilizando el teorema (7) procedemos a calcular los nuevos intervalos de confianza,

```
mu <- colMeans(datos_osos) # vecotr de medias

S_mof <- matrix(sqrt(diag(cov(datos_osos))/n),1)

valor_critico <- abs(qt(0.05/(2*p),n-1)) # cambiamos el valro criico

lim_inf_bon <- mu-valor_critico*S_mof # limites de los intervalso de confianza
lim_inf_bon
```

```
##           [,1]     [,2]     [,3]     [,4]
## [1,] 138.0904 145.9831 167.8511 168.1678
```

```
lim_sup_bon <- mu+valor_critico*S_mof
lim_sup_bon
```

```
##         [,1]     [,2]     [,3]     [,4]
## [1,] 148.481 172.5883 178.4347 186.1179
```

Entonces, los intervalos de confianza para las cuatro medias poblacionales de la longitud por año μ_1, μ_2, μ_3 y μ_4 son

variable	lím. inf.	lim. inf. bonf.	media	lim. sup. bonf.	lím. sup.
Longitud2	130.6851024	138.090424	μ_1 =143.2857143	148.4810045	155.8863261
Longitud3	127.0216268	145.9830825	μ_2 =159.2857143	172.588346	191.5498017
Longitud4	160.3081582	167.8510517	μ_3 =173.1428571	178.4346626	185.9775561
Longitud5	155.3748671	168.1678147	μ_4 =177.1428571	186.1178996	198.9108472

Cuadro 3: Comparación de los intervalos de confianza simultaneos T^2 al 95 % de confianza.

Comparando los intervalos notamos que los invertalos considerando el método de Bonferroni es más cortos que los intervalos de confianza simultáneos. Ahora calculemos los intervalos de los aumentos en medias,

```
S <- cov(datos_osos)
A <- matrix(c(-1,1,0,0,0,-1,1,0,0,0,-1,1),4)
var_x <- c()
for (i in 1:3){
```

```
  var_x[i] <- sqrt((t(A[,i])%*%S%*%(A[,i]))/n)
}

dif_mu <- diff(mu)

valor_critico <- abs(qt(0.05/(2*3),n-1)) # cambiamos el vector critico

lim_inf_d_b <- dif_mu-valor_critico*var_x # limites de los intervalso de confianza
lim_inf_d_b
```

```
##  Longitud3  Longitud4  Longitud5
##  1.6703152 -0.2267315 -5.4900656
```

```
lim_sup_d_b <- dif_mu+valor_critico*var_x
lim_sup_d_b
```

```
## Longitud3 Longitud4 Longitud5
##  30.32968  27.94102  13.49007
```

variable	lím. inf.	lim. inf. bon.	aumento anual	lim. sup. bon.	lím. sup.
Diferencia 2 a 3 años	-21.2264919	1.6703152	$\mu_2 - \mu_1 = 16$	30.3296848	53.2264919
Diferencia 3 a 4 años	-22.7307684	-0.2267315	$\mu_3 - \mu_2 = 13.8571429$	27.9410172	50.4450541
Diferencia 4 a 5 años	-20.6538466	-5.4900656	$\mu_4 - \mu_3 = 4$	13.4900656	28.6538466

Cuadro 4: Comparación de los intervalos de confianza simultaneos T^2 al 95 % de confianza.

Comparando el cuadro , podemos observar que los intervalos de confianza con el método de Bonferroni son más estrechos. Además, los intervalos de confianza simultaneos todos contienen al 0 por lo que se podría concluir que los incrementos anuales son iguales, pero los intervalos con el método de Bonferroni la diferencia en los años 2 y 3 no contiene al 0 lo que podríamos concluir que existe evidencia significativa para decir que son diferentes. Estas diferencias tienen sentido, ya que con el método de Bonferroni se esta controlando la tasa de error global intependiente de la estructura de correlación entre las variables. ■.