Entrega 3 Problemas y Talleres MATII Estadística grado informática 2019-2020

Ricardo Alberich

13-05-2020

Contenidos

1 Entregas 3 Problemas: Estadística Inferencial 1

Contestad cada GRUPO de 3 a los siguientes problemas y cuestiones en un fichero Rmd y su salida en html o pdf.

Cambien podéis incluir capturas de problemas hechos en papel. Cada pregunta vale lo mismo y se reparte la nota entre sus apartados.

1.1 Problema 1: Contraste de parámetros de dos muestras.

Queremos comparar los tiempos de realización de un test entre estudiantes de dos grados G1 y G2, y determinar si es verdad que los estudiantes de G1 emplean menos tiempo que los de G2. No conocemos σ_1 y σ_2. Disponemos de dos muestras independientes de cuestionarios realizados por estudiantes de cada grado, $n_1 = n_2 = 50$.

Los datos están en http://bioinfo.uib.es/~recerca/MATIIIGMAT/NotasTestGrado/, en dos ficheros `grado1.txt` y `grado2.txt`.

```
G1=read.table("http://bioinfo.uib.es/~recerca/MATIIIGMAT/NotasTestGrado/grado1.txt",
              header=TRUE)$x
G2=read.table("http://bioinfo.uib.es/~recerca/MATIIIGMAT/NotasTestGrado/grado2.txt",
              header=TRUE)$x
n1=length(na.omit(G1))
```

```
n2=length(na.omit(G2))
media.muestra1=mean(G1,na.rm=TRUE)
media.muestra2=mean(G2,na.rm=TRUE)
desv.tip.muestra1=sd(G1,na.rm=TRUE)
desv.tip.muestra2=sd(G2,na.rm=TRUE)
```

Calculamos las medias y las desviaciones típicas muestrales de los tiempos empleados para cada muestra. Los datos obtenidos se resumen en la siguiente tabla:

$$\begin{array}{llll} n_1 & = 50, & n_2 & = 50 \\ \overline{x}_1 & = 9.7592926, & \overline{x}_2 & = 11.4660825 \\ \tilde{s}_1 & = 1.1501225, & \tilde{s}_1 & = 1.5642932 \end{array}$$

Se pide:

1. Comentad brevemente el código de R explicando que hace cada instrucción.
2. Contrastad si hay evidencia de que las notas medias son distintas entre los dos grupos. En dos casos considerando las varianzas desconocidas pero iguales o desconocidas pero distintas. Tenéis que hacer el contraste de forma manual y con funciones de `R` y resolver el contrate con el p-valor.
3. Calculad e interpretar los intervalos de confianza para la diferencia de medias asociados a los dos test anteriores.
4. Comprobad con el test de Fisher y el de Levene si las varianza de las dos muestras son iguales contra que son distintas. Tenéis que resolver el test de Fisher con `R` y de forma manual y el test de Levene con `R` y decidir utilizando el p-valor.

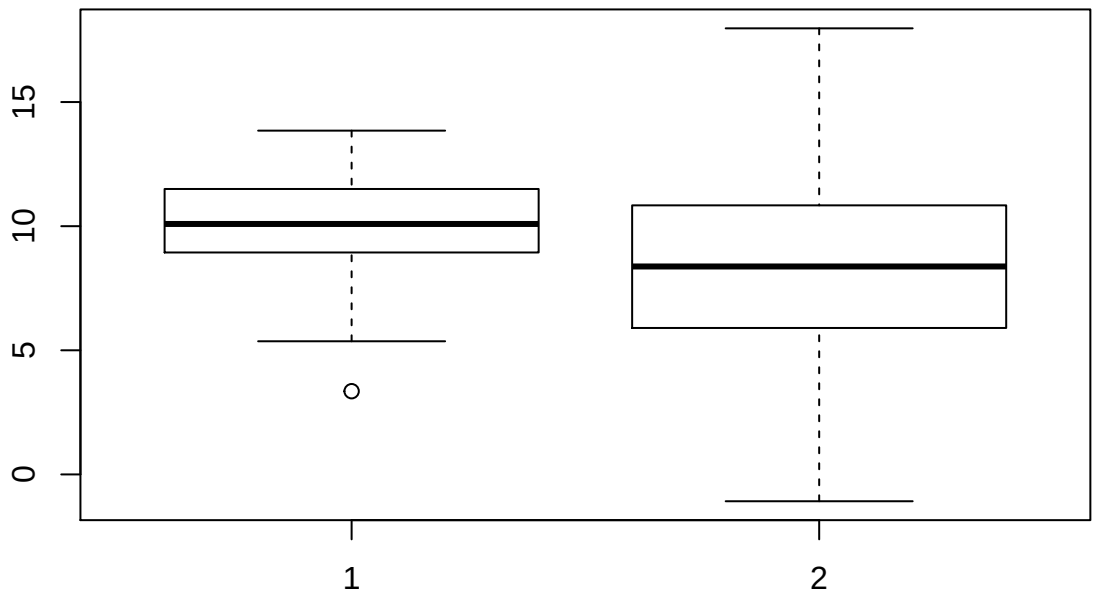
1.2 Problema 2 : Contraste dos muestras

Simulamos dos muestras con las funciones siguientes

```
x1=rnorm(100,mean = 10,sd=2)
x2=rnorm(100,mean = 8,sd=4)
```

Dibujamos estos gráficos

```
boxplot(x1,x2)
```




```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

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

```
##
## Attaching package: 'car'
```

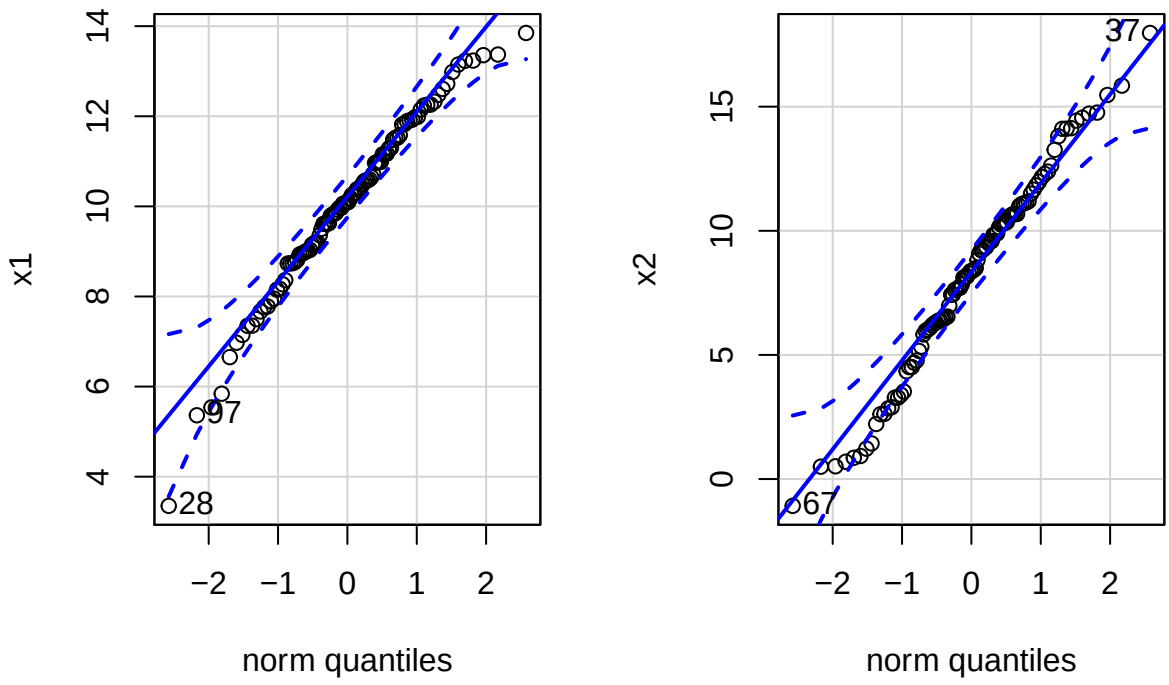
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
par(mfrow=c(1,2))
qqPlot(x1)
```

```
## [1] 28 97
```

```r
qqPlot(x2)
```




```
## [1] 37 67
```

```r
par(mfrow=c(1,1))
```

Realizamos algunos contrastes de hipótesis de igual de medias entre ambas muestras

```r
t.test(x1,x2,var.equal = TRUE,alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 4.1038, df = 198, p-value = 0.00002968
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.099638      Inf
## sample estimates:
## mean of x mean of y
## 10.041892  8.200902
```

```r
t.test(x1,x2,var.equal = FALSE,alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = 4.1038, df = 142.94, p-value = 0.00006799
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9542407 2.7277402
## sample estimates:
## mean of x mean of y
## 10.041892  8.200902
```

```r
t.test(x1,x2,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 4.1038, df = 198, p-value = 0.00005936
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9563407 2.7256402
## sample estimates:
## mean of x mean of y
## 10.041892  8.200902
```

Se pide

1. ¿Cuál es la distribución y los parámetros de las muestras generadas?
2. ¿Qué muestran y cuál es la interpretación de los gráficos?
3. ¿Qué test contrasta si hay evidencia a favor de que las medias poblacionales de las notas en cada grupo sean distintas? Di qué código de los anteriores resuelve este test.
4. Para el test del apartado anterior dad las hipótesis nula y alternativa y redactar la conclusión del contraste.

1.3 Problema 3 : Bondad de ajuste

Queremos analiza los resultados de aprendizaje con tres tecnologías. Para ello se seleccionan 3 muestras de 50 estudiantes y se les somete a evaluación después de un curso.

```r
nota=factor(sample(c(1,2,3,4),p=c(0.1,0.4,0.3,0.2),replace=TRUE,size=150),
            labels=c("S","A","N","E"))
tecnologia=rep(c("Mathematica","R","Python"),each=50)
frec=table(nota,tecnologia)
frec
```

```
##     tecnologia
## nota Mathematica Python  R
```

```
##    S           4      4  6
##    A          20     18 26
##    N          17     19 11
##    E           9      9  7
```

```
col_frec=colSums(frec)
col_frec
```

```
## Mathematica      Python           R
##          50          50          50
```

```
row_frec=rowSums(frec)
row_frec
```

```
##  S  A  N  E
## 14 64 47 25
```

```
N=sum(frec)
teoricas=row_frec%*%t(col_frec)/N
teoricas
```

```
##      Mathematica    Python         R
## [1,]    4.666667  4.666667  4.666667
## [2,]   21.333333 21.333333 21.333333
## [3,]   15.666667 15.666667 15.666667
## [4,]    8.333333  8.333333  8.333333
```

```
dim(frec)
```

```
## [1] 4 3
```

```
dim(teoricas)
```

```
## [1] 4 3
```

```
sum((frec-teoricas)^2/teoricas)
```

```
## [1] 4.729195
```

```
chisq.test(table(nota,tecnologia))
```

```
## Warning in chisq.test(table(nota, tecnologia)): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(nota, tecnologia)
## X-squared = 4.7292, df = 6, p-value = 0.579
```

Se pide

1. Discutid si hacemos un contraste de independencia o de homogeneidad de las distribuciones de las notas por tecnología. Escribid las hipótesis del contraste.
2. Interpretad la función `chisq.test` y resolved el contraste.
3. Interpretad `teoricas=row_frec%*%t(col_frec)/N` reproducid manualmente el segundo resultado de la primera fila.

1.4 Problema 4 : Bondad de ajuste. La ley de Benford

La ley de Benford es una distribución discreta que siguen las frecuencias de los primero dígitos significativos (de 1 a 9) de algunas series de datos curiosas.

Sea una v.a. X con dominio $D_X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ diremos que sigue una ley de Benford si

$$P(X = x) = \log_{10}\left(1 + \frac{1}{x}\right) \text{ para } x \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}\,.$$

Concretamente lo podemos hacer así

```
prob=log10(1+1/c(1:9))
prob
```

```
## [1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195
## [8] 0.05115252 0.04575749
```

```
MM=rbind(c(1:9),prob)
df=data.frame(rbind(prob))
# Y hacemos una bonita tabla
colnames(df)=paste("Díg.",c(1:9),sep =" ")
knitr::kable(df,format ='markdown')
```

	Díg. 1	Díg. 2	Díg. 3	Díg. 4	Díg. 5	Díg. 6	Díg. 7	Díg. 8	Díg. 9
prob	0.30103	0.1760913	0.1249387	0.09691	0.0791812	0.0669468	0.0579919	0.0511525	0.0457575

En general esta distribución se suele encontrar en tablas de datos de resultados de observaciones de funciones científicas, contabilidades, cocientes de algunas distribuciones ...

Por ejemplo se dice que las potencias de números enteros siguen esa distribución. Probemos con las potencias de 2. El siguiente código calcula las potencias de 2 de 1 a 1000 y extrae los tres primeros dígitos.

```
# R pasa los enteros  muy grande a reales. Para nuestros propósitos
# es suficiente para extraer los tres primeros dígitos.
muestra_pot_2_3digitos=str_sub(as.character(2^c(1:1000)),1,3)
head(muestra_pot_2_3digitos)
```

```
## [1] "2"  "4"  "8"  "16" "32" "64"
```

```r
tail(muestra_pot_2_3digitos)
```

```
## [1] "334" "669" "133" "267" "535" "107"
```

```r
#Construimos un data frame con tres columnas que nos dan el primer,
#segundo y tercer dígito respectivamente.
df_digitos=data.frame(muestra_pot_2_3digitos,
                      primer_digito=as.integer(
                        substring(muestra_pot_2_3digitos, 1, 1)),
                      segundo_digito=as.integer(
                        substring(muestra_pot_2_3digitos, 2, 2)),
                      tercer_digito=as.integer(
                        substring(muestra_pot_2_3digitos, 3, 3)))
head(df_digitos)
```

```
##   muestra_pot_2_3digitos primer_digito segundo_digito tercer_digito
## 1                      2             2             NA            NA
## 2                      4             4             NA            NA
## 3                      8             8             NA            NA
## 4                     16             1              6            NA
## 5                     32             3              2            NA
## 6                     64             6              4            NA
```

Notad que los NA en el segundo y el tercer dígito corresponden a número con uno o dos dígitos.

Se pide:

1. Contrastad con un test χ^2 que el primer dígito sigue una ley de Benford. Notad que el primer dígito no puede ser 0. Resolved manualmente y con una función de `R`.

2. Contrastad con un test χ^2 que el segundo dígito sigue una ley de uniforme discreta. Notad que ahora si puede ser 0. Resolved con funciones de `R`.

3. Contrastad con un test χ^2 que el tercer dígito sigue una ley de uniforme discreta. Notad que ahora si puede ser 0. Resolved con manualmente calculado las frecuencias esperadas y observadas, el estadístico de contraste y el p-valor utilizando `R`. Comprobad que vuestros resultados coinciden con los de la función de `R` que calcula este contraste.

4. Dibujad con `R` para los apartados 1 y 2 los diagramas de frecuencias esperados y observados. Comentad estos gráficos

1.5 Problema 5 : ANOVA

El siguiente código nos da las notas numéricas (variable `nota`) de los mismos ejercicios para tres tecnologías en tres muestra independientes de estudiantes de estas tres tecnologías diferentes

```r
head(nota)
```

```
## [1]  59.42540  31.25305  68.38927  76.77203 113.38209  63.38222
```

```r
library(nortest)
lillie.test(nota[tecnologia=="Mathematica"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nota[tecnologia == "Mathematica"]
## D = 0.091608, p-value = 0.3683
```

```r
lillie.test(nota[tecnologia=="R"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nota[tecnologia == "R"]
## D = 0.077947, p-value = 0.6288
```

```r
lillie.test(nota[tecnologia=="Python"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nota[tecnologia == "Python"]
## D = 0.091377, p-value = 0.3722
```

```r
lillie.test(nota)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nota
## D = 0.042685, p-value = 0.7235
```

```r
bartlett.test(nota~tecnologia)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  nota by tecnologia
## Bartlett's K-squared = 0.29005, df = 2, p-value = 0.865
```

```r
library(car)
leveneTest(nota~tecnologia)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.3749  0.688
##       147
```

```
sol_aov=aov(nota~tecnologia)
```

Del `summary(sol_aov)` os damos la salida a falta de algunos de los valores

```
> summary(sol_aov)
             Df Sum Sq Mean Sq F value Pr(>F)
tecnologia   --    674  -----   -----  0.592
Residuals   ---  94142  -----
```

```
pairwise.t.test(nota,tecnologia,p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  nota and tecnologia
##
##        Mathematica Python
## Python 0.49        -
## R      0.32        0.75
##
## P value adjustment method: none
```

Se pide

1. ¿Podemos asegurar que la muestras son normales en cada grupo? ¿y son homocedásticas? Sea cual sea la respuesta justificad que parte del código la confirma.
2. La función `aov` que test calcula. Escribid formalmente la hipótesis nula y la alternativa.
3. Calcula la tabla de ANOVA y resuelve el test.
4. ¿Qué contrates realiza la función `pairwise.t.test`? Utilizando los resultados anteriores aplicad e interpretad los contrates del apartado anterior utilizando el ajuste de Holm.

1.6 Problema 6 : Comparación de las tasas de interés para la compra de coches entre seis ciudades.

Consideremos el `data set newcar` accesible desde https://www.itl.nist.gov/div898/education/anova/newcar.dat de *Hoaglin, D., Mosteller, F., and Tukey, J. (1991). Fundamentals of Exploratory Analysis of Variance. Wiley, New York, page 71.*

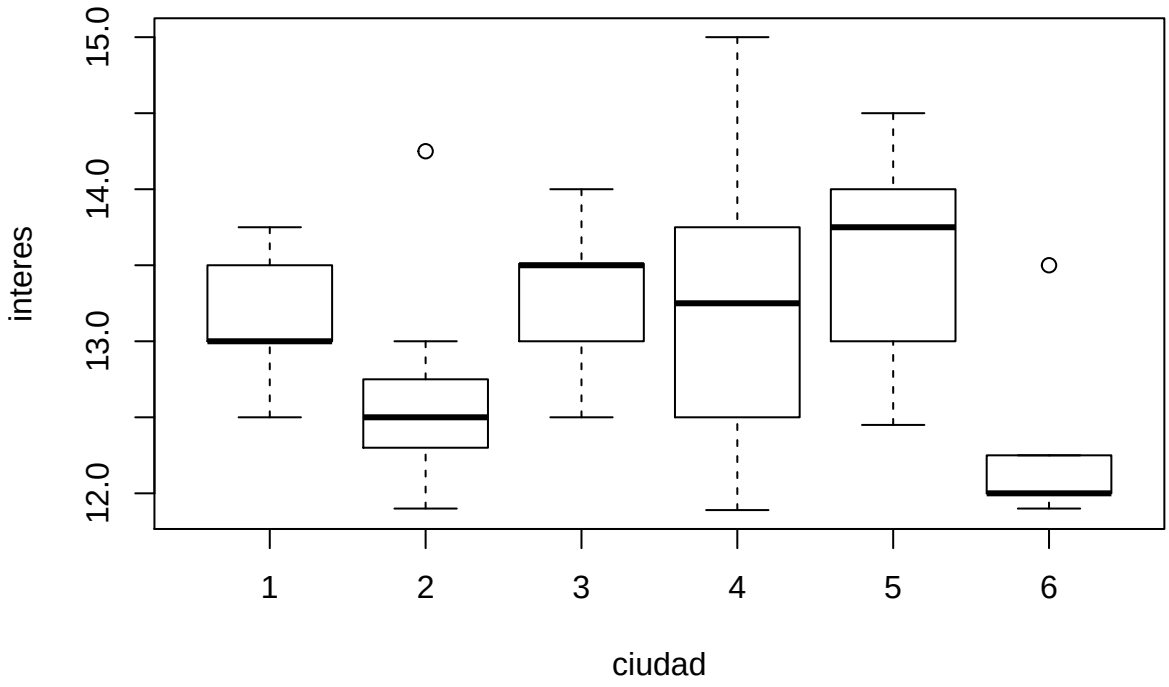
Este data set contiene dos columnas:

- Rate (interés): tasa de interés en la compra de coches a crédito
- City (ciudad) : la ciudad en la que se observó la tasa de interés para distintos concesionarios (codificada a enteros). Tenemos observaciones de 6 ciudades.

```
datos_interes=read.table(
  "https://www.itl.nist.gov/div898/education/anova/newcar.dat",
  skip=25)
# salto las 25 primeras líneas del fichero,son un preámbulo qiue explica los datos.
names(datos_interes)=c("interes","ciudad")
str(datos_interes)
```

```
## 'data.frame':    54 obs. of  2 variables:
##  $ interes: num  13.8 13.8 13.5 13.5 13 ...
##  $ ciudad : int  1 1 1 1 1 1 1 1 1 2 ...
```

```
boxplot(interes~ciudad,data=datos_interes)
```



Se pide:

1. Comentad el código y el diagrama de caja.
2. Se trata de contrastar si hay evidencia de que la tasas medias de interés por ciudades son distintas. Definid el ANOVA que contrasta esta hipótesis y especificar qué condiciones deben cumplir las muestras para poder aplicar el ANOVA.

3. Comprobad las condiciones del ANOVA con un test KS y un test de Levene (con código de `R`). Justificad las conclusiones.

4. Realizad el contraste de ANOVA (se cumplan las condiciones o no) y redactar adecuadamente la conclusión. Tenéis que hacedlo con funciones de `R`.

5. Se acepte o no la igualdad de medias realizar las comparaciones dos a dos con ajustando los p-valor tanto por Bonferroni como por Holm al nivel de significación $\alpha = 0.1$. Redactad las conclusiones que se obtienen de las mismas.

1.7 Problema 7: Cuestiones cortas

- Cuestión 1: Supongamos que conocemos el p-valor de un contraste. Para que valores de nivel de significación α RECHAZAMOS la hipótesis nula.
- Cuestión 2: Hemos realizado un ANOVA de un factor con 3 niveles, y hemos obtenido un p-valor de 0.001. Suponiendo que las poblaciones satisfacen las condiciones para que el ANOVA tenga sentido, ¿podemos afirmar con un nivel de significación $\alpha = 0.05$ que las medias de los tres niveles son diferentes dos a dos? Justificad la respuesta.
- Cuestión 3: Lanzamos 300 veces un dado de 6 caras de parchís, queremos contrastar que los resultados son equiprobables. ¿Cuáles serian las frecuencias esperadas o teóricas del contraste?
- Cuestión 4: En un ANOVA de una vía, queremos contrastar si los 6 niveles de un factor definen poblaciones con la misma media. Sabemos que estas seis poblaciones son normales con la misma varianza $\sigma = 2$. Estudiamos a 11 individuos de cada nivel y obtenemos que $SS_{Total} = 256.6$ y $SS_{Tr} = 60.3$. ¿Qué vale SS_E. ¿Qué valor estimamos que tiene σ^2?
- Cuestión 6: Calculad la correlación entre los vectores de datos $x = (1, 3, 4, 4)$, $y = (2, 4, 12, 6)$.
- Cuestión 7: De estas cuatro matrices, indicad cuáles pueden ser matrices de correlaciones, y explicad por qué.

$$A = \begin{pmatrix} 1 & 0.8 \\ -0.8 & 1 \end{pmatrix}, B = \begin{pmatrix} 0.8 & 0.6 \\ 0.6 & 0.8 \end{pmatrix}, C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, D = \begin{pmatrix} 1 & 1.2 \\ 1.2 & 1 \end{pmatrix}.$$