Ejercicio 3

Cedric Prieels

Octubre 2016

Resumen

Este ejercicio consiste en leer dos ficheros, y en usar diferentes métodos de los que vimos en clase para estudiar si hay correlación o no entre las dos columnas de los ficheros de datos. Después, se estima también mediante el uso del coeficiente de correlación si la correlación que obtenemos en cada caso es fuerte, o débil.

Introducción y métodología

La correlación se define como la relación o dependencia que puede existir entre dos series de datos (en este ejercicio, solo se estudia la correlación de tipo lineal pero existen muchos más tipos). Hay que tener mucho cuidado a la hora de establecer si hay una correlación o no, porque muchos parámetros pueden influyir esta determinación (como los efectos de selección, si todos los puntos no tienen la misma probabilidad de estar medidos).

Para ver si hay correlación o no entre los datos a nuestra disposición, se pueden usar diferentes tests que vimos en clase. Se usan entonces las tres pruebas siguientes para resolver el ejercico :

- Test de Pearson (Test de correlación lineal paramétrico)
- Test de Spearman (Test de correlación lineal no paramétrico por rangos)
- Test de la tau de Kendall (Test de correlación lineal no paramétrico por rangos relativos)

El test de Pearson es un test parámetrico mientras que las otras dos pruebas son no-parámetricas y entonces más robustas en el sentido de que no se basan en la hipotesis que consiste en suponer que las distribuciones de puntos que estamos estudiando son gaussianas.

Como en muchos casos en estádistica, solo podremos calcular el valor de la probabilidad nula H0 que consiste en decir que "no tenemos correlación". Si obtenemos un valor P(H0) muy pequeño, podremos concluir que no existe correlación entre las variables que estamos estudiando.

Lo que nos interesa no es solamente definir si hay correlación o no, pero también de determinar cuanto de fuerte es esta correlación. Para esto, se puede usar un parámetro que se define diferentamente en cada test pero que sigue siempre las mismas reglas : toma valores entre -1 y 1 (correlación linea perfecta), y nos indica directamente si la correlación entre las series de datos es fuerte o débil.

Habrá que tener mucho cuidado con este valor del coeficiente de correlación lineal que calculamos, porque tener un valor de coeficiente cerca a 1 (correlación linea perfecta) no signica que tengamos correlación entre los datos (podríamos tener un caso que tenga r = 1 pero P(H0) = 0).

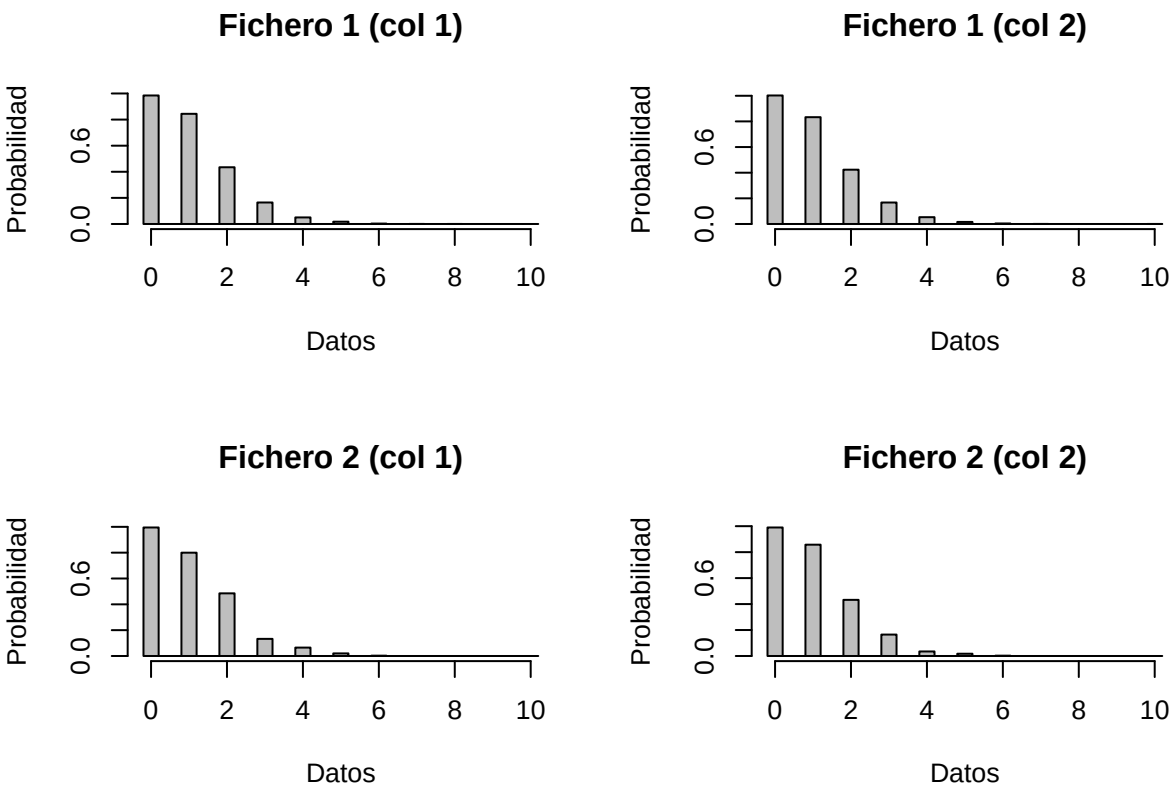
Resultados

Para ver si hay una correlación entre las columnas de los ficheros de datos, lo primero que hay que hacer es por supuesto y como siempre abrir estos ficheros para ver la pinta que tienen las distribuciones y representarlas con histogramas.

```r
rm(list=ls())

setwd('/Users/ced2718/Documents/Universite/Estadistica/Ejercicio_3/')
data1 <- read.table("dat5.dat", header=FALSE)
data2 <- read.table("dat6.dat", header=FALSE)

#Convertimos a unos arrays que contienen cada columna de los datos iniciales
data1Col1 <- data1[,1]
data1Col2 <- data1[,2]
data2Col1 <- data2[,1]
data2Col2 <- data2[,2]

#Calculamos el valor medio de cada columna
meanData1Col1 = mean(data1Col1)
meanData1Col2 = mean(data1Col2)
meanData2Col1 = mean(data1Col1)
meanData2Col2 = mean(data2Col2)
```

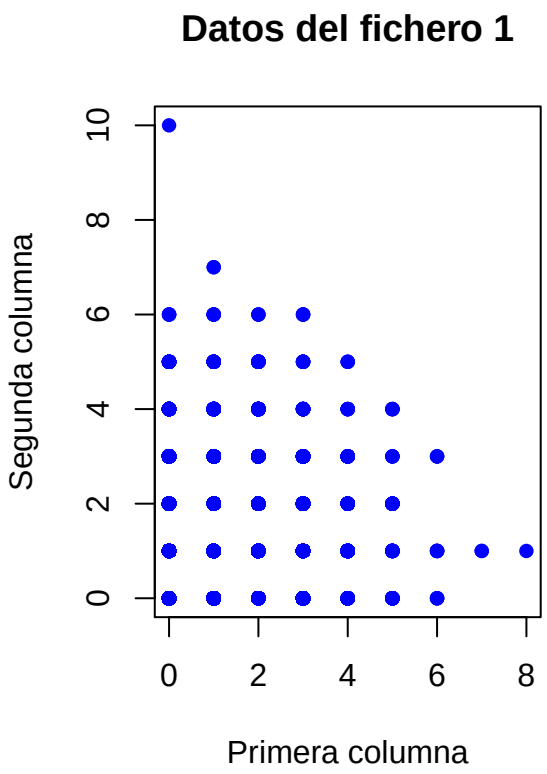
En este caso, vemos directamente que los ficheros de datos solo tienen números enteros positivos en cada columna. Cada fichero tiene además un número de lineas diferente. Pintamos ahora estos datos en histogramas para estudiarlos un poco más en detalles, antes de empezar a estudiar la correlación. Primero, se representa cada columna sola, aunque no sea quizá el más interesante a pintar en este caso de estudio de correlación.

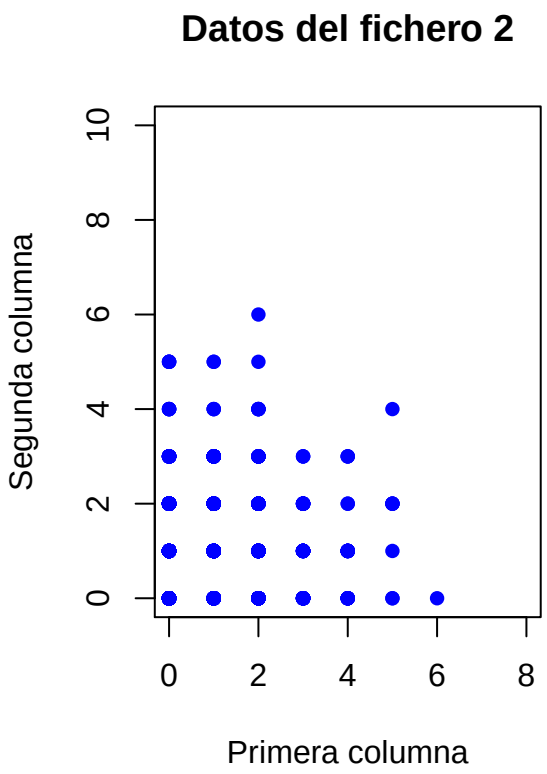
```r
par(mfrow=c(2,2))
bines <- c(-0.2, 0.2, 0.8, 1.2, 1.8, 2.2, 2.8, 3.2, 3.8, 4.2, 4.8,
           5.2, 5.8, 6.2, 6.8, 7.2, 7.8, 8.2, 8.8, 9.2, 9.8, 10.2)
hist(data1Col1, breaks=bines, freq = F, main="Fichero 1 (col 1)",
     xlab="Datos", ylab="Probabilidad", lwd=1, pch=".", col="grey")
hist(data1Col2, breaks=bines, freq = F, main="Fichero 1 (col 2)",
     xlab = "Datos", ylab="Probabilidad", lwd=1, pch=".", col="grey")
hist(data2Col1, breaks=bines, freq = F, main="Fichero 2 (col 1)",
     xlab = "Datos", ylab="Probabilidad", lwd=1, pch=".", col="grey")
hist(data2Col2, breaks=bines, freq = F, main="Fichero 2 (col 2)",
     xlab = "Datos", ylab="Probabilidad", lwd=1, pch=".", col="grey")
```



Después, se representa lo más interesante, cada columna frente a la otra por cada fichero. Ponemos la misma escala en los dos casos para poder comparar mejor los datos que tenemos.

```
par(mfrow=c(1,2))
plot(data1Col1, data1Col2, col="blue", pch=16, main="Datos del fichero 1",
     xlab = "Primera columna", ylab = "Segunda columna")
plot(data2Col1, data2Col2, col="blue", pch=16, main="Datos del fichero 2",
     xlab = "Primera columna", ylab = "Segunda columna", xlim = c(0,8), ylim=c(0,10))
```





Ahora que sabemos un poco más que pinta tienen las distribuciones, podemos empezar a estudiar las posibles correlaciones que existen.

Test de Pearson

El valor de la probabilidad nula P(H0) se puede calcular directamente usando la función cor.test de R, que toma como parámetros los dos arrays de datos y un string igual al test que estamos usando.

```
pearsonData1 <- cor.test(data1Col1, data1Col2, method="pearson")
pearsonData2 <- cor.test(data2Col1, data2Col2, method="pearson")

#Ya podemos calcular el valor de P(H0) y del coeficiente de correlación lineal r
probaPearsonData1 <- pearsonData1$p.value
probaPearsonData2 <- pearsonData2$p.value

rPearsonData1 <- pearsonData1$estimate
rPearsonData2 <- pearsonData2$estimate
```

El valor del coeficiente de correlación lineal r también se puede calcular a mano, usando las formulas vistas en clase. Esto nos permite comprobar el resultado que acabamos de obtener.

```
#Fichero 1
sumData1 <- sum((data1Col1-meanData1Col1)*(data1Col2-meanData1Col2))
sumData1Col1 <- sum((data1Col1-meanData1Col1)**2)
sumData1Col2 <- sum((data1Col2-meanData1Col2)**2)
rPearson2Data1 <- sumData1/sqrt(sumData1Col1*sumData1Col2)

#Fichero 2
```

```
sumData2 <- sum((data2Col1-meanData2Col1)*(data2Col2-meanData2Col2))
sumData2Col1 <- sum((data2Col1-meanData2Col1)**2)
sumData2Col2 <- sum((data2Col2-meanData2Col2)**2)
rPearson2Data2 <- sumData2/sqrt(sumData2Col1*sumData2Col2)
```

Comparamos ahora con una tabla los resultados obtenidos con los dos métodos.

```
resultadosPearsonR <- matrix(c(rPearsonData1, rPearsonData2,
                               rPearson2Data1, rPearson2Data2),ncol=2)

colnames(resultadosPearsonR) <- c("Coeficiente r (con R)", "Coeficiente r (a mano)")
rownames(resultadosPearsonR) <- c("Fichero 1","Fichero 2")

rtab <- as.table(resultadosPearsonR)
head(rtab)
```

```
##           Coeficiente r (con R) Coeficiente r (a mano)
## Fichero 1            0.07988428             0.07988428
## Fichero 2            0.03963437             0.03963302
```

Obtenemos resultados iguales calculando el valor del coeficiente de correlación linea a mano y con R. Los resultados obtenidos hasta ahora tienen entonces buena pinta.

Test de Spearman

Esta prueba es una prueba más robusta que el test de Pearson porque no supone que tengamos ficheros de entrada con datos que siguen una distribución gaussiana. Este test no se aplica directamente a los datos pero se aplica a los rangos de los datos (lo que está bien también para que un eventual punto muy fuera de la distribución no impacte tanto los resultados). Lo primero que hay que hacer es entonces calcular estos rangos, como lo hicimos en el ejercicio 2.

```
data1rangosCol1 <- rank(data1Col1, ties.method="average")
data1rangosCol2 <- rank(data1Col2, ties.method="average")

data2rangosCol1 <- rank(data2Col1, ties.method="average")
data2rangosCol2 <- rank(data2Col2, ties.method="average")
```

Después, como para el primer test, calculamos el valor de la probabilidad nula y del coeficiente de correlación lineal ρ, de dos maneras diferentes (a mano y con las funciones de R).

```
spearmanData1 <- cor.test(data1Col1, data1Col2, method="spearman")
```

```
## Warning in cor.test.default(data1Col1, data1Col2, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
spearmanData2 <- cor.test(data2Col1, data2Col2, method="spearman")
```

```
## Warning in cor.test.default(data2Col1, data2Col2, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
#Ya podemos calcular el valor de P(HO) y del coeficiente de correlación lineal rho
probaSpearmanData1 <- spearmanData1$p.value
probaSpearmanData2 <- spearmanData2$p.value

rhoSpearmanData1 <- spearmanData1$estimate
rhoSpearmanData2 <- spearmanData2$estimate

#Cálculo de rho a mano
meanRangosData1Col1 <- mean(data1rangosCol1)
meanRangosData1Col2 <- mean(data1rangosCol2)
meanRangosData2Col1 <- mean(data2rangosCol1)
meanRangosData2Col2 <- mean(data2rangosCol2)

#Fichero 1
sumRangosData1 <- sum((data1rangosCol1-meanRangosData1Col1)*
                        (data1rangosCol2-meanRangosData1Col2))
sumRangosData1Col1 <- sum((data1rangosCol1-meanRangosData1Col1)**2)
sumRangosData1Col2 <- sum((data1rangosCol2-meanRangosData1Col2)**2)
rhoSpearman2Data1 <- sumRangosData1/sqrt(sumRangosData1Col1*sumRangosData1Col2)

#Fichero 2
sumRangosData2 <- sum((data2rangosCol1-meanRangosData2Col1)*
                        (data2rangosCol2-meanRangosData2Col2))
sumRangosData2Col1 <- sum((data2rangosCol1-meanRangosData2Col1)**2)
sumRangosData2Col2 <- sum((data2rangosCol2-meanRangosData2Col2)**2)
rhoSpearman2Data2 <- sumRangosData2/sqrt(sumRangosData2Col1*sumRangosData2Col2)
```

Se pueden resumir otra vez todos los resultados en una tabla para comparar y analizarlos.

```
resultadosSpearmanRho <- matrix(c(rhoSpearmanData1, rhoSpearmanData2,
                                  rhoSpearman2Data1, rhoSpearman2Data2),ncol=2)

colnames(resultadosSpearmanRho) <- c("Coeficiente rho (con R)", "Coeficiente rho (a mano)")
rownames(resultadosSpearmanRho) <- c("Fichero 1","Fichero 2")

rtab <- as.table(resultadosSpearmanRho)
head(rtab)
```

```
##           Coeficiente rho (con R) Coeficiente rho (a mano)
## Fichero 1              0.08252166               0.08252166
## Fichero 2              0.04898227               0.04898227
```

Los coeficientes obtenidos con los dos métodos son los mismos, lo que nos puede confortar en el método usado.

Test de la tau de Kendall

Esta prueba es también una prueba no parámetrica que no necesita suponer que tengamos distribuciones gaussianas, lo que no interesa mucho en este caso (porque las distribuciones que tenemos tienen más una pinta de seguir una distribución de Poisson).

```r
kendallData1 <- cor.test(data1Col1, data1Col2, method="kendall")
kendallData2 <- cor.test(data2Col1, data2Col2, method="kendall")

#Ya podemos calcular el valor de P(H0) y del coeficiente de correlación lineal tau
probaKendallData1 <- kendallData1$p.value
probaKendallData2 <- kendallData2$p.value

tauKendallData1 <- kendallData1$estimate
tauKendallData2 <- kendallData2$estimate
```

Resultados obtenidos

Ahora podemos empezar a mirar y a analyzar un poco más todos los resultados que hemos obtenido hasta ahora. Primero, se representan todos los valores de probabilidad de la hipotésis nula y de los coeficientes de correlación lineal en una table.

Fichero	$P_{H_0}(Pearson)$	$P_{H_0}(Spearman)$	$P_{H_0}(Kendall)$
Fichero 1	$1.24 \cdot 10^{-15}$	$1.39 \cdot 10^{-16}$	$1.43 \cdot 10^{-16}$
Fichero 2	$2.10 \cdot 10^{-1}$	$01.22 \cdot 10^{-1}$	$1.22 \cdot 10^{-1}$

Vemos claramente una diferencia entre los valores de probabilidades con los dos ficheros : el fichero 1 (dat5.dat) tiene correlación entre sus dos columnas (porque P(H0) tiene un valor muy cerca a 0) mientras que es dificíl de sacar una conclusión exacta sobre el fichero 2, puesto que la probabilidad que obtenemos no tiene valor muy cerca a 0 o muy cerca a 1. Podemos observar también que en el primer caso, los tres métodos diferentes nos dan valores de probabilidad muy cercas mientras que el método de Pearson se aleja bastante de los otros dos en el caso del fichero 2 (probablemente porque es la única prueba parámetrica de las 3, que supone que las dsitribuciones que tenemos son gaussianas).

También se puede estudiar los diferentes valores de coeficiente de correlación linear obtenidos.

r (con R)	r (a mano)	ρ (con R)	ρ (a mano)	τ (con R)	τ (a mano)
Fichero 1	0.07988428	0.07988428	0.08252166	0.08252166	0.07084642
Fichero 2	0.03963437	0.03963302	0.04898227	0.04898227	0.04204475

En este caso, lo primero que podemos observar es que el cálculo de este coeficiente a mano o con R siempre da casi el mismo valor, lo que da confianza en el método usado. Después, se observa también que cada prueba da un valor de coeficiente muy similar por cada fichero. Por fin, vemos que este coeficiente es más grande en el caso del fichero 1, lo que significa que tenemos un tipo de correlación muy fuerte, mientras que la correlación es menos fuerte en el fichero 2.

Conclusión

En conclusión, como ya explicado, podemos concluir que el fichero 1 (dat5.dat) presenta una correlación evidente entre sus dos columnas (todos los métodos usados devuelven este mismo resultado, que la probabilidad de que no exista correlación es nula). Sin emebargo, la correlación obtenida es bastante débil, porque los tres coeficientes de correlación lineal obtenidos toman valores entre 0.070 y 0.082.

Lo mismo pasa con el fichero 2 (dat6.dat), aunque en este caso la correlación es menos evidente (en este caso, la probabilidad de obtener esos resultados y de no haber correlación es pequeña, pero diferente de 0). No podemos descartar la posibilidad de existencia de correlación. Conviene indicar que el test de Pearson

devuelve un valor de probabilidad diferente de las otras dos pruebas (más fiables) y que este fichero tiene solamente 1.000 números, mientras que el primer fichero tenía 10.000 datos. En este caso, el valor de los coeficientes de correlación lineal es incluso más pequeño que en el caso precidente, tomando valores en el intervalo [0.039-0.048].

Podemos concluir diciendo que en este caso se han estudiado dos distribuciones de datos que tienen correlaciones débiles (coeficiente de correlación lineal cerca a 0) pero muy significativas (valor de probabilidad de la hipotésis nula cerca a 0 también).

Bibliografía

R Markdown, *Markdown basics*, http://rmarkdown.rstudio.com/authoring_basics.html. Consultado por última vez el 29 de octubre 2016.

R Development Core Team (2008). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.