Software para el Análisis de Datos. Trabajo grupo X.

Jose Angel Fernandez-Caballero Rico, Elena Tortosa Binacua, Jorge Pulido Lozano, Miguel Grau López

Diciembre 20, 2015

Abstract

Aquí podemos hacer una despripción general de nuestros datos/objetivos.

Descripción del trabajo.

1. Buscad un conjunto de datos relacionado con la Bioestadística. Puede ser generado por simulación, pueden ser datos incorporados en un paquete de R (no se usarán los paquetes IRIS, IRIS3) u otros repositorios. Tenéis que explicar la procedencia de estos datos.
2. Explicad las variables de vuestro fichero, tipo, clasificación, etc... ¡Todo aquello que creáis que es relevante!
3. Pensad un mínimo de cuatro de preguntas objetivo que queráis contestar con estos datos.
4. Haced el análisis descriptivo de los datos. Todo aquello que creáis que es necesario. ¡Resumid los datos a vuestro criterio pero con sentido! Pensad siempre en las preguntas objetivo que queréis contestar. Si debéis recodificar variables.
5. Generad los gráficos que creáis necesarios para resumir de forma gráfica la información que tenéis.

Descripción de las variables

LOCALIDAD (Jose Angel Fernandez-Caballero).

Esta variable nos informa del Hospital de procedencia donde se encuentra el paciente. Se trata de una variable tipo cualitativo nominal. El fichero de datos cuenta con 11 localidades distintas (VIRGEN DE LAS NIEVES, Poniente El Ejido, Torrecardenas, San Cecilio, Jaen, Motril, Andujar, LINARES, CARCEL GRANADA, JAEN y VALENCIA)

CV (Jose Angel Fernandez-Caballero).

Mediante esta variable podemos ver la cuantificación de la infección por virus VIH, se calcula por estimación de la cantidad de partículas virales en los fluidos corporales. Se trata de una variable cuantitativa. Su rango es de 34-10.000.00 copias/ml.

CD4 (Jose Angel Fernandez-Caballero).

Los linfocitos T-CD4 son un tipo de células que constituyen una parte esencial del sistema inmunitario. Su función principal es la de activar al propio sistema alertándole de la presencia de patógenos o de una replicación errónea de células humanas, para que pueda hacerles frente y corregir la situación. Se trata también de una variable cuantitativa cuyo rango es de 2-1.650 cel/ml.

SEXO: Sexo del paciente. Hombre-Mujer.

ESTADO: Descripción del estado clínico actual del paciente. Tres posibles opciones:

1. Naive: Paciente que todavía no han comenzado ningún tratamiento o que empieza a tratarse por primera vez.
2. Dejó tratamiento: Paciente que ha abandonado el tratamiento. El motivo puede ser de distinta índole (efectos secundarios, desisten por agotamiento etc)

3. Fracaso: Paciente bajo tratamiento en los que no se ha conseguido frenar la replicación del virus. Las razones pueden ser varias, por ejemplo una mutación de resistencia o sencillamente que el paciente haya dejado de tomar el fármaco.

EDAD (Elena Tortosa)

NACIONALIDAD (Elena Tortosa)

SUBTIPO (Jorge Pulido)

MUTACIONESPAÑA

RESISTENCIA MUTACIONESPAÑA (Jorge Pulido)

Preguntas objetivo

José Angel

1. ¿ Existe diferencia de Carga viral entre los pacientes naives y los fracasos?
2. ¿ Que subtipo VIH predomina en cada nacionalidad?
3. ¿ Como se distribuye en % la infeccion VIH entre mujeres y hombres?
4. ¿ Las mutaciones de resistecia se da en pacientes naives o en fracasos?
5. ¿ Que mutación es la mas prevalente?
6. ¿ Como se distribuyen los individuos en los hospitales de procedencia?

Jorge Pulido

1. ¿quien son mas propensos a dejar el tto: hombre o mujeres? (facilona)
2. ¿existe prevalencia de una mutacion sobre un subtipo? (ya propuesta)
3. ¿relacion entre la nacionalidad y el subtipo? (ya propueta)
4. ¿relacion entre la nacionalidad y estado del tto? (ver si ciertas nacionalidades quedan excluidas o no)
5. ¿relacion entre la cv y los CD4?

Elena

1. Con respecto a la carga viral: ¿Hay algún tipo de relación entre la edad o el sexo y la carga viral? No sé si puede tener sentido que el virus se replique más en hombras o mujeres, o en gente joven o más mayor.

2. Con respecto a los niveles de CD4: ¿Hay algún tipo de relación entre la edad o el sexo y los niveles de cd4? ¿Responde mejor el sistema inmune de hombres o mujeres frente a la infección por el virus?¿Hay alguna diferencia entre gente joven o mayor respecto a los niveles de CD4? Para la edad podríamos hacer disferentes categorias como menores de 25, entre 25-40, entre 40 y 65 y mayores de 65. O algo así...

3. Con respecto al subtipo: ¿Hay alguna relación entre el subtipo y el estado/CD4/CV? A lo mejor algún subtipo es más agresivo que otro e induce una mayor carga viral, una mayor respuesta del sistema inmune o un fracaso en el tratamiento.

4. Con respecto a las mutacions: ¿Hay alguna mutación que induzca una mayor carga viral?¿Y una mayor respuesta del sistema inmune?

Análisis descriptivo

1. ¿ Existe diferencia de Carga viral entre los pacientes naives y los fracasos?

```
datosCSV <- read.csv("/home/miquel/uoc_SAD/DatosEntrada/TABLA\ MASTER.csv", header=T, dec=".",sep="\t")
CV<-datosCSV[2]
ESTADO<-datosCSV[5]
with(datosCSV, tapply(CV, list(ESTADO), mean, na.rm=TRUE))
```

```
##     DEJO TRA/FRA DEJO TRATAMIENTO          FRACASO            NAIVE
##          1390.00        140985.71         41875.72        282669.35
```

```
#original (selecciona columna sexo en lugar de ESTADO real)
#ESTADO<-datosCSV[4]
#   with(datosCSV, tapply(CV, list(ESTADO), mean, na.rm=TRUE))
```

Observamos que la media de CV en el grupo NAIVE (282669) es bastante superior al grupo de FRACASO (41875) pero vamos a ver si esa diferencia es significativa.

```
#original
#(da error, como si t.test solo aceptara una lista con 2 posibles valores
# y ESTADO contiene DEJO TRATAMIENTO, FRACASO y NAIVE (3))
#t.test(CV~ESTADO, alternative='two.sided', conf.level=.95, var.equal=FALSE,
#data=datosCSV_fracaso_naive)

#Si previamente, filtramos datosCSV para seleccionar unicamente los casos
#fracaso-naive, sí que podemos ejecutar el t.test
datosCSV_fracaso_naive <- subset(datosCSV,
                          datosCSV$ESTADO == "FRACASO" | datosCSV$ESTADO == "NAIVE")
t.test(CV~ESTADO, alternative='two.sided', conf.level=.95,
       var.equal=FALSE, data=datosCSV_fracaso_naive)
```

```
##
##  Welch Two Sample t-test
##
## data:  CV by ESTADO
## t = -3.4522, df = 197.42, p-value = 0.0006802
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -378345.9 -103241.4
## sample estimates:
## mean in group FRACASO   mean in group NAIVE
##              41875.72             282669.35
```

Obtenemos un p valor inferior a 0.05 por tanto, no existen evidencias significativas para aceptar la hipótesis nula de igualdad de medias, por tanto, no podemos afirmar que la carga viral media en el grupo NAIVE es la misma que el grupo FRACASO.

2. ¿ Como se distribuyen los individuos en los hospitales de procedencia?

```r
table(datosCSV[1])
```

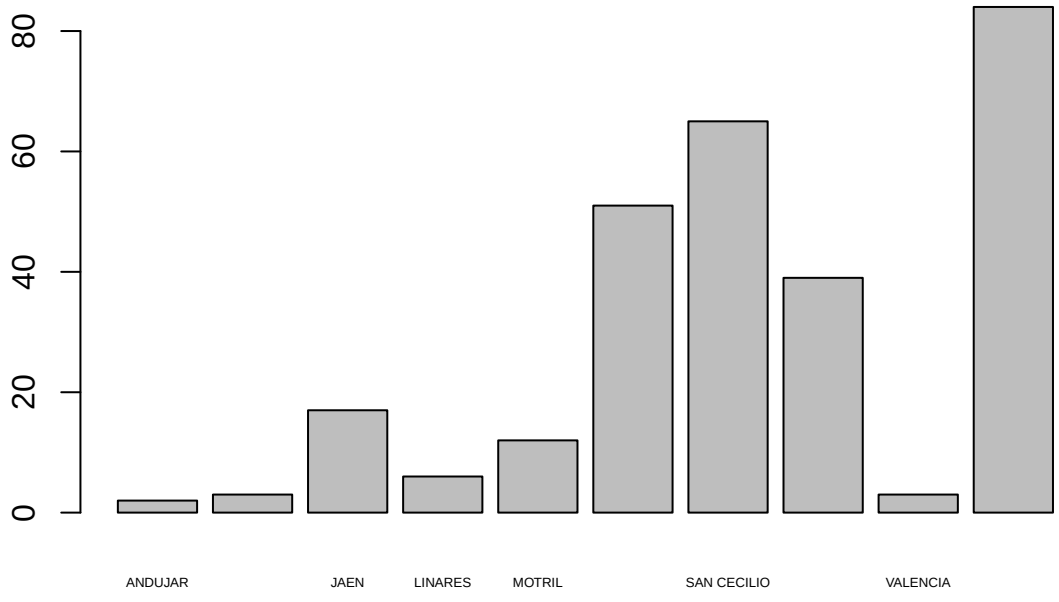
```
##
##              ANDUJAR       CARCEL GRANADA                 JAEN
##                    2                    3                   17
##              LINARES              MOTRIL    PONIENTE EL EJIDO
##                    6                   12                   51
##          SAN CECILIO       TORRECARDENAS             VALENCIA
##                   65                   39                    3
## VIRGEN DE LAS NIEVES
##                   84
```

El hospital con más pacientes es el Virgen de las Nieves con un total de 80.

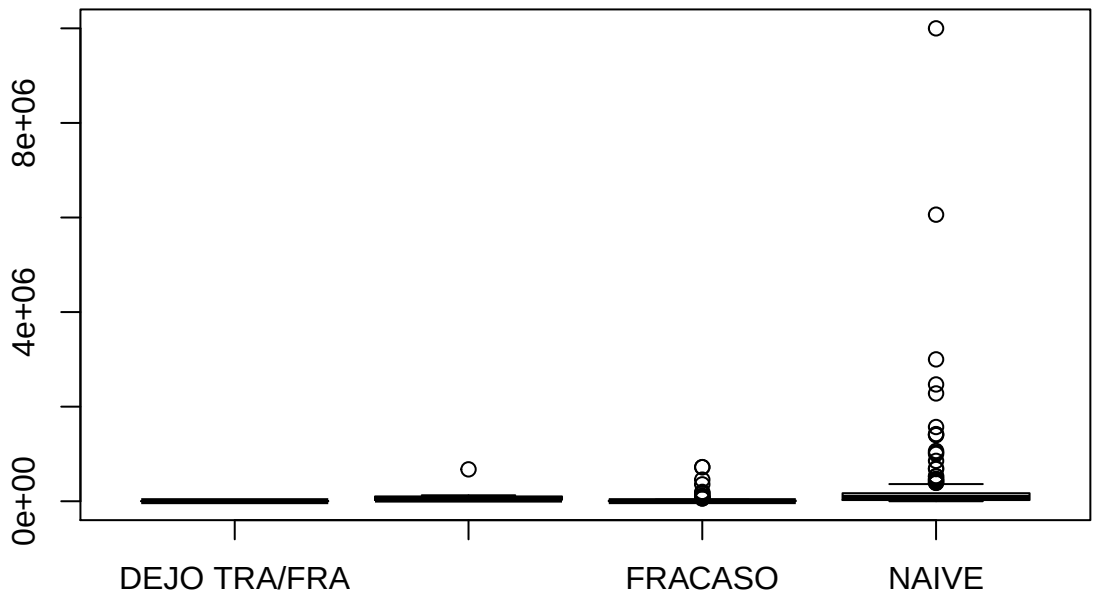
Gráficos

1. ¿ Existe diferencia de Carga viral entre los pacientes naives y los fracasos?

```r
plot(datosCSV[1], cex.names =0.4)
```

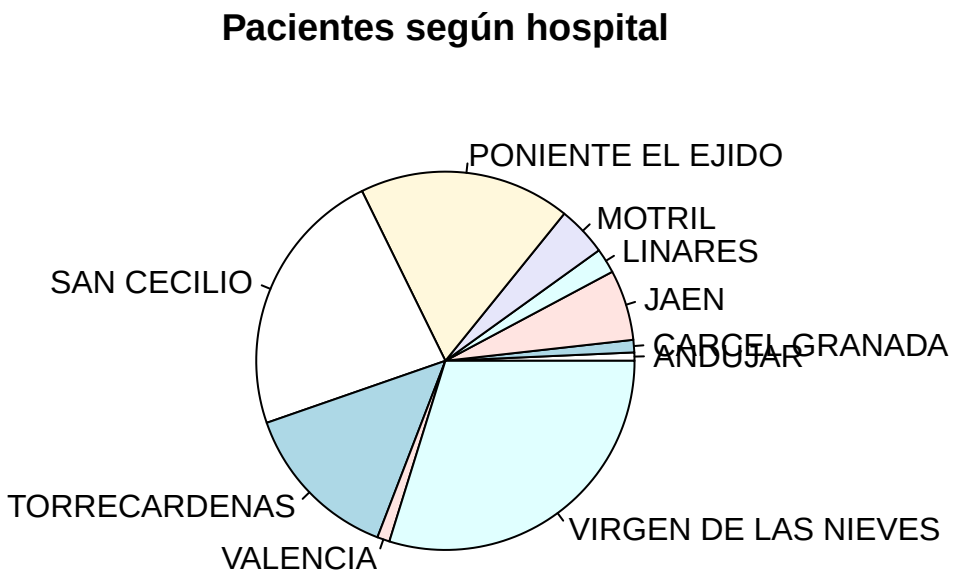



```r
boxplot(CV~ESTADO, data=datosCSV, id.method="y")
```



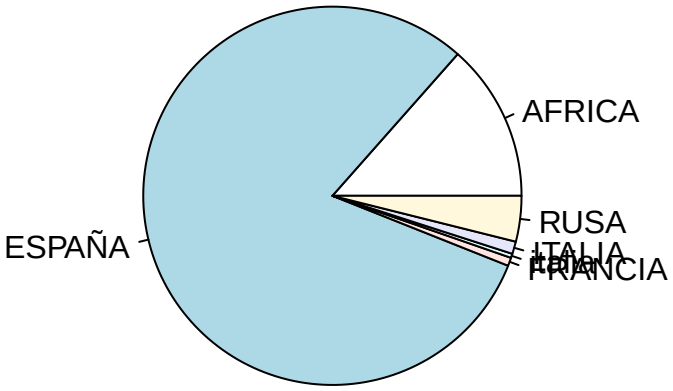
2. ¿ Como se distribuyen los individuos en los hospitales de procedencia?

```
with(datosCSV, pie(table(LOCALIDAD), labels=levels(LOCALIDAD), xlab="",
  ylab="", main="Pacientes según hospital"))
```



Para saber mas acerca de como estan distribuidas las nacionalidades de los pacientes VIH podemos hacer la siguiente grafica. Donde se observa (como era de esperar) que la nacionalidad Española es la mas comun.

```
with(datosCSV, pie(table(nacionalidad), labels=levels(nacionalidad), xlab="",
                   ylab="", main="Frecuencia pacientes según nacionalidad"))
```

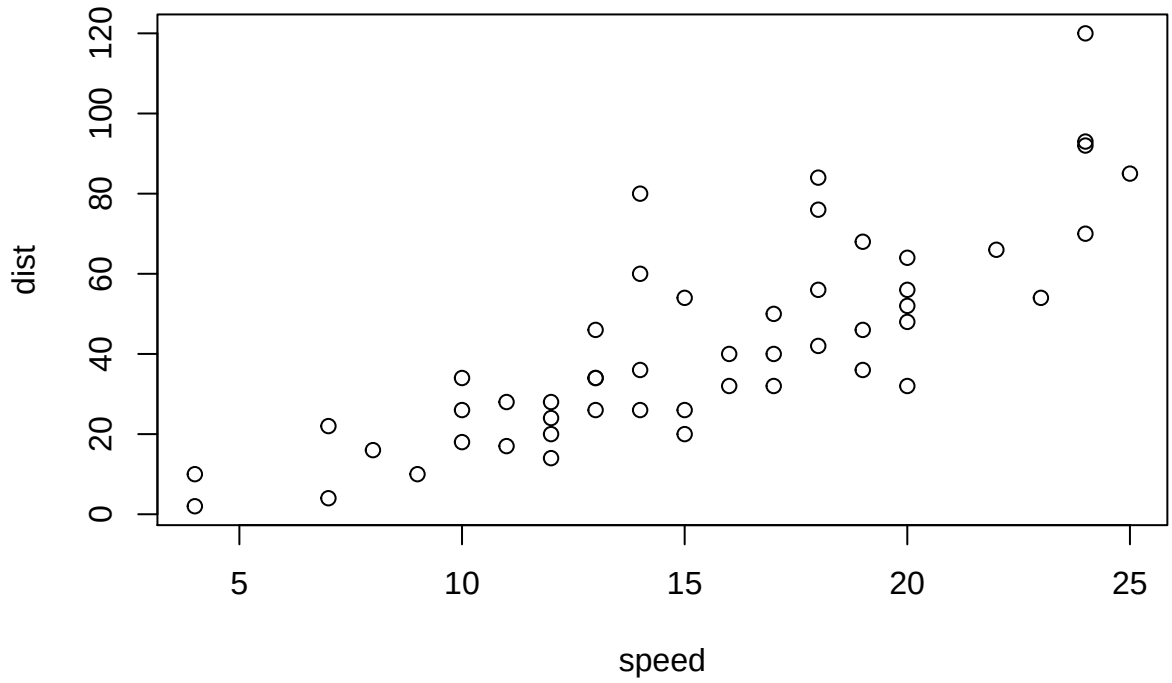


Ejemplo de como agregar gráficos en el informe:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Por ejemplo:



`echo = FALSE` evita mostrar el código que genera el gráfico.