Software para el Análisis de Datos. Trabajo del grupo X.

José Ángel Fernández-Caballero, Elena Tortosa, Jorge Pulido, Miguel Grau

Enero 15, 2015

Abstract

A partir de un dataset real con datos de pacientes con el virus VIH se busca obtener conclusiones relacionadas con distintas variables como niveles de carga viral y CD4, mutaciones e índices de éxitos/fracasos del tratamiento. El estudio se compone de una descripción de las variables, seguido de un análisis descriptivo donde se aplican distintos métodos estadísticos, y finalmente un análisis gráfico que apoya el análisis descriptivo realizado previamente.

Descripción del trabajo.

1. Conjunto de datos. Elección del dataset y descripción general del mismo.
2. Explicación de las variables del dataset.
3. Preguntas propuestas por cada uno de los miembros del grupo para estudiar el dataset.
4. Ánalisis descriptivo siguiendo las preguntas propuestas.
5. Gráficos de apoyo al análisis descriptivo.

Para realizar el trabajo, se ha hecho uso de un repositorio en github, permitiendo así llevar un control de versiones colaborativo de una manera sencilla y robusta, además de tener en todo momento una version unificada y disponible del estudio. Se puede acceder al repositorio desde aquí.

Conjunto de datos

El conjunto de datos proceden de la base de datos del hospital de Sevilla, en concreto del servicio de Enfermedades Infecciosas. Estos datos recogen variables socio-demográficas de pacientes infectados con VIH. La Base de Datos (BdD) está protegida con clave y dotada de diferentes mecanismos lógicos que impidan la introducción de datos erróneos. Además sólo podrán acceder a la BdD los investigadores implicados en el proyecto. Cabe aclarar que se separan en una segunda BdD y con diferente clave de acceso los datos identificativos de los pacientes y en esta última BdD el acceso estará permitido unicamente al Investigador principal.

Todos los investigadores que accedan a esta base de datos se comprometen a respetar la confidencialidad de los datos de acuerdo a la Ley Orgánica 15/1999, de 13 de Diciembre, sobre la Protección de datos de Carácter Personal y la ley 41/2002 de 14 de Noviembre, ley básica reguladora de la autonomía del paciente y derechos y obligaciones en materia de información y documentación clínica.

Descripción de las variables

LOCALIDAD: 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)

Carga viral (CV): 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 de tipo continuo. Su rango es de 34-10.000.00 copias/ml.

CD4: 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 de tipo continuo cuyo rango es de 2-1.650 cel/ml.

SEXO: Esta variable nos informa del sexo del paciente. Es una variable de tipo cualitativo nominal, y puede ser tomar dos valores diferentes: hombre o mujer.

ESTADO: Descripción del estado clínico actual del paciente (variable de tipo cualitativo). Presenta 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: Esta variable nos informa de la edad de los pacientes que se están estudiando. Se trata de una variable tipo cuantitativa. Aunque podría considerarse de tipo continuo, aquí se trata como una variable de tipo discreta, en la que únicamente se consideran valores de años completos. Para realizar los análisis de los datos dividiremos a los pacientes en tres grupo según su edad: jóvenes (menores de 3 años), adultos (comprendidos entre los 30 y 65 años) y el grupo de la tercera edad (aquellos mayores de 65 años).

NACIONALIDAD: Esta variable nos informa de la nacionalidad de los pacientes estudiados. Se trata de una variable de tipo cualitativo. En estos datos encontramos cinco nacionalidades diferentes: africana, española, francesa, italiana y rusa.

SUBTIPO: Los subtipos de del virus VIH son consecuencia de la alta capacidad reproductiva que tienen los virus. La mayoría de esos subtipos se deben a la introducción de errores por parte de la transcriptasa inversa, generando así una amplia gama de variantes proteínas que conforman el complejo vírico. Las cepas del virus VIH se clasifican en dos tipos (1 y 2), las cuales a su vez se clasifican en subtipos/genotipos según la sección que tengan mutadas. Actualmente, la cepa causante de la pandemia es la variante VIH-1 M subtipo B, el cual es el predominante en el primer mundo. Este hecho, en Europa, puede compensarse por la entrada del genotipo VIH-1 M subtipo A procedente de inmigrantes portadores del virus de origen africano.

RESISTENCIA MUTACIONES: Las mutaciones de resistencia del virus son aquellas que ocasionan que el virus continúe replicándose en el huésped a pesar de la administración de retrovirales para paliar la enfermedad. Las mutaciones de resistencia se producen principalmente a nivel de la transcriptasa inversa del virus y en las proteasas del mismo, principalmente. Los medicamentos antiretrovirales atacan a estas estructuras para impedir la formación de nuevos viriones, al presentar una mutación, el medicamento no lo reconoce y por tanto la carga viral aumenta en el paciente.

Preguntas objetivo

Se plantean un conjunto de preguntas a aplicar sobre el dataset con el objetivo de obtener conclusiones relevantes de éste. Las preguntas planteadas por cada componente del grupo son (en cursiva las elegidas a desarrollar):

José Angel

1. *¿Existe diferencia de carga viral entre los pacientes naives y los fracasos?*
2. ¿Qué subtipo VIH predomina en cada nacionalidad?
3. ¿Cómo 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 más prevalente?
6. *¿Cómo se distribuyen los individuos en los hospitales de procedencia?*

Elena

1. *Con respecto a la carga viral: ¿Hay algún tipo de relación entre la edad o el sexo y la carga viral? ¿Es posible 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?*
3. Con respecto al subtipo: ¿Hay alguna relación entre el subtipo y el estado/CD4/carga viral? 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 mutaciones: ¿Hay alguna mutación que induzca una mayor carga viral?¿Y una mayor respuesta del sistema inmune?

Jorge Pulido

1. ¿Quiénes son más propensos a dejar el tratamiento: hombres o mujeres?
2. *¿Existe prevalencia de una mutación sobre un subtipo?*
3. ¿Relación entre la nacionalidad y el subtipo?
4. ¿Relación entre la nacionalidad y estado del tratamiento?
5. ¿Relación entre la carga viral y los CD4?
6. *¿Existe relación entre la edad del paciente y las mutaciones que se desarrollan?*

Miquel

1. *¿Hay alguna relación entre el subtipo y el CD4/carga viral?*
2. *¿Existe alguna relación entre la carga viral y los CD4?*

Análisis descriptivo

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

Empezamos calculando los valores medios entre los distintos tipos de pacientes:

```
datosCSV <- read.csv("~/tmp/uoc/uoc_SAD/DatosEntrada/tablaMaster_corregida.csv",
                  header=T, dec=".", na.strings="NA",sep="")
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
```

Observamos que la media de carga viral (CV) en el grupo NAIVE (282669) es bastante superior al grupo de FRACASO (41875) pero vamos a ver si esa diferencia es significativa. Para ello, vamos a calcular un t-test:

```
#Previamente, filtramos datosCSV para seleccionar unicamente los casos
#fracaso-naive y ejecutamos 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. No podemos afirmar que la carga viral media en el grupo NAIVE es la misma que el grupo FRACASO.

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

```
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 84. Por el contrario, el hospital que menos pacientes presenta es Andujar.

3. ¿Existe algún tipo de relación entre la carga viral o los niveles de CD4, y el sexo o la edad del paciente?

Esto nos daría una idea de si el virus es capaz de replicarse mejor en hombres o mujeres, o estos grupos responden de manera diferencial a la infección del virus (según los niveles de CD4). Además, podemos ver si la edad del paciente influye en los niveles de replicación del virus o en la respuesta del individuo a la infección.

```
#Cargamos datos
Tabla <- read.csv("~/tmp/uoc/uoc_SAD/DatosEntrada/tablaMaster_corregida.csv",
                  header=T, dec=".", na.strings="NA",sep="")
```

Recodifico la variable edad en una nueva variable categórica llamada **edad_cat**. Se considerará: **jóven** a los menores de 30 años, **adultos** a los comprendidos entre los 30 y 65 años, **tercera edad** a lo mayores de 65 años.

```
library(car)
Tabla <- within(Tabla, {
  edad_cat <- recode(edad,
  '0:34 = "Joven"; 35:65 = "Adulto"; 66:100 = "Tercera edad"',
  as.factor.result=TRUE)
})
```

Analizo el *número* de individuos y la *proporción* de nuestra población que hay en cada **grupo de edad**:

```
local({
  .Table <- with(Tabla, table(edad_cat))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## edad_cat
##       Adulto        Joven Tercera edad
##          175          102            4
##
## percentages:
## edad_cat
##       Adulto        Joven Tercera edad
##        62.28        36.30         1.42
```

Estudio el *número* de individuos y la *proporción* de nuestra población que hay según el **género**:

```
local({
  .Table <- with(Tabla, table(SEXO))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## SEXO
## HOMBRE  MUJER
##    218     64
##
## percentages:
## SEXO
## HOMBRE  MUJER
##   77.3   22.7
```

Sumarizo los principales estadísticos de **carga viral (CV) y CD4** en función del **género**:

```
library(RcmdrMisc)
```

```
## Loading required package: sandwich
```

```
numSummary(Tabla[,"CV"], groups=Tabla$SEXO,
           statistics=c("mean", "sd", "IQR","quantiles"), quantiles=c(0,.25,.5,.75,1))
```

```
##            mean       sd      IQR 0%     25%   50%    75%     100% data:n
## HOMBRE 227837.6 860691.6 113568.2 34 5056.75 38700 118625 10000000    218
## MUJER  114694.9 236683.5 124922.5 30 1327.50 25550 126250  1420000     64
```

```
numSummary(Tabla[,"CD4"], groups=Tabla$SEXO,
           statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

```
##            mean       sd IQR 0%    25% 50%    75% 100% data:n data:NA
## HOMBRE 367.7685 231.2963 279  3 207.25 365 486.25 1650    216       2
## MUJER  356.7206 242.5547 276  2 179.00 340 455.00 1081     63       1
```

Y en función del **grupo de edad**:

```
numSummary(Tabla[,"CV"], groups=Tabla$edad_cat,
           statistics=c("mean", "sd", "IQR","quantiles"), quantiles=c(0,.25,.5,.75,1))
```

```
##                   mean        sd       IQR    0%      25%      50%    75%
## Adulto       178944.82 791660.12 118540.50    30  2709.50  40400.0 121250
## Joven        247750.28 741128.23 106435.00    34  6440.00  28800.0 112875
## Tercera edad  98771.75  51481.93  56134.75 27900 77615.25 112593.5 133750
```

```
##                   100% data:n
## Adulto       10000000    175
## Joven         6060000    102
## Tercera edad   142000      4
```

```
numSummary(Tabla[,"CD4"], groups=Tabla$edad_cat,
           statistics=c("mean", "sd", "IQR", "quantiles"),quantiles=c(0,.25,.5,.75,1))
```

```
##                   mean       sd    IQR 0%    25%   50%    75% 100% data:n
## Adulto       345.9851 240.0014 300.75  2 171.00 346.5 471.75 1650    174
## Joven        396.8200 217.7806 250.50 12 247.75 369.0 498.25 1081    100
## Tercera edad 342.5000 284.7847 221.50  6 226.50 332.0 448.00  700      4
##              data:NA
## Adulto             1
## Joven              2
## Tercera edad       0
```

A continuación voy a comprobar si las diferencias que se observan en CV y CD4 entre los distintos grupos de género y edad son significativas.

Primero compruebo si los datos poseen una distribución normal aplicando el test de Shapiro-Wilk.

```
CD4_hombre <- subset (Tabla, SEXO=HOMBRE)
CD4_mujer <- subset (Tabla, SEXO=MUJER)
shapiro.test(CD4_hombre$CD4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CD4_hombre$CD4
## W = 0.93631, p-value = 1.345e-09
```

```
shapiro.test(CD4_mujer$CD4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CD4_mujer$CD4
## W = 0.93631, p-value = 1.345e-09
```

```
CV_hombre <- subset (Tabla, SEXO=HOMBRE)
CV_mujer <- subset (Tabla, SEXO=MUJER)
shapiro.test(CV_hombre$CV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CV_hombre$CV
## W = 0.23576, p-value < 2.2e-16
```

```r
shapiro.test(CV_mujer$CV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CV_mujer$CV
## W = 0.23576, p-value < 2.2e-16
```

```r
CD4_Joven <- subset (Tabla, edad_cat=Joven)
CD4_Adulto <- subset (Tabla, edad_cat=Adulto)
CD4_Tercera_edad <- subset (Tabla, edad_cat=Tercera_edad)
shapiro.test(CD4_Joven$CD4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CD4_Joven$CD4
## W = 0.93631, p-value = 1.345e-09
```

```r
shapiro.test(CD4_Adulto$CD4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CD4_Adulto$CD4
## W = 0.93631, p-value = 1.345e-09
```

```r
shapiro.test(CD4_Tercera_edad$CD4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CD4_Tercera_edad$CD4
## W = 0.93631, p-value = 1.345e-09
```

```r
CV_Joven <- subset (Tabla, edad_cat=Joven)
CV_Adulto <- subset (Tabla, edad_cat=Adulto)
CV_Tercera_edad <- subset (Tabla, edad_cat=Tercera_edad)
shapiro.test(CV_Joven$CV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CV_Joven$CV
## W = 0.23576, p-value < 2.2e-16
```

```r
shapiro.test(CV_Adulto$CV)
```

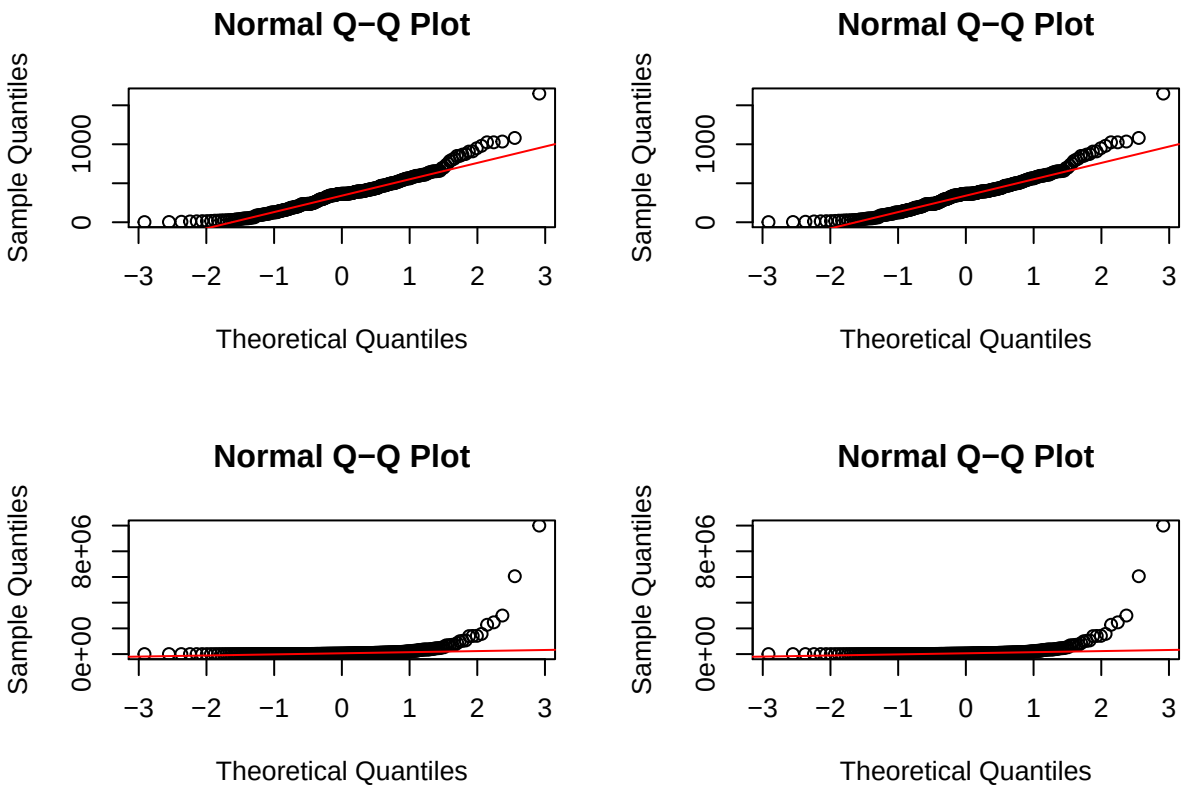
```
##
##  Shapiro-Wilk normality test
##
## data:  CV_Adulto$CV
## W = 0.23576, p-value < 2.2e-16
```

```
shapiro.test(CV_Tercera_edad$CV)
```

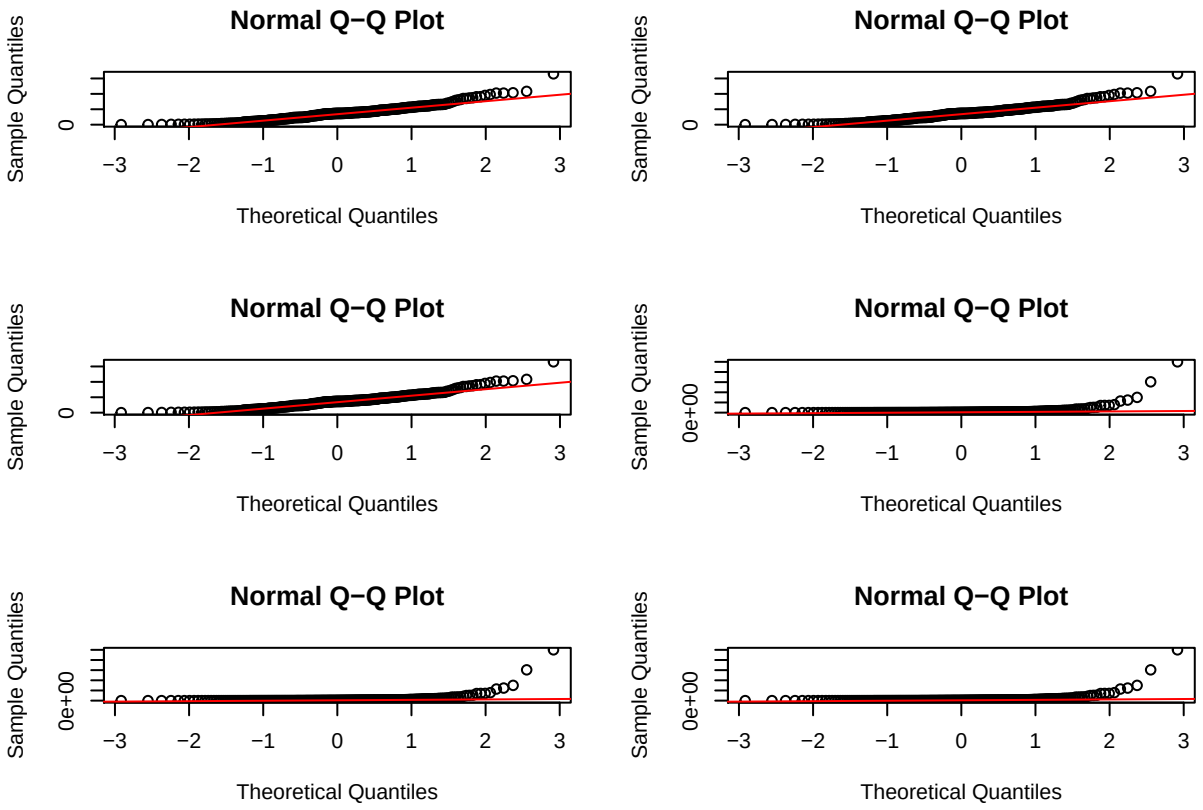
```
##
##  Shapiro-Wilk normality test
##
## data:  CV_Tercera_edad$CV
## W = 0.23576, p-value < 2.2e-16
```

En todos los casos el p-valor es menor que 0.05, por lo que rechazo la hipótesis nula de que los datos poseen una distribución normal. Para comprobarlo de forma visual, represento qqplots.

```
par(mfrow=c(2,2))
qqnorm(CD4_hombre$CD4);qqline(CD4_hombre$CD4, col = 2)
qqnorm(CD4_mujer$CD4);qqline(CD4_mujer$CD4, col = 2)
qqnorm(CV_hombre$CV);qqline(CV_hombre$CV, col = 2)
qqnorm(CV_mujer$CV);qqline(CV_mujer$CV, col = 2)
```



```r
par(mfrow=c(3,2))
qqnorm(CD4_Joven$CD4);qqline(CD4_Joven$CD4, col = 2)
qqnorm(CD4_Adulto$CD4);qqline(CD4_Adulto$CD4, col = 2)
qqnorm(CD4_Tercera_edad$CD4);qqline(CD4_Tercera_edad$CD4, col = 2)
qqnorm(CV_Joven$CV);qqline(CV_Joven$CV, col = 2)
qqnorm(CV_Adulto$CV);qqline(CV_Adulto$CV, col = 2)
qqnorm(CV_Tercera_edad$CV);qqline(CV_Tercera_edad$CV, col = 2)
```



Sabiendo que los datos no siguen una distribución normal aplicaré el test de Wilcoxon cuando comparamos los grupos de género y el test Kruskal Wallis para los grupos de edad:

```r
wilcox.test(CD4 ~ SEXO, data=Tabla)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  CD4 by SEXO
## W = 7105, p-value = 0.5938
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(CV ~ SEXO, data=Tabla)
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  CV by SEXO
## W = 7796.5, p-value = 0.1529
## alternative hypothesis: true location shift is not equal to 0
```

```
kruskal.test(CD4 ~ edad_cat, data=Tabla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CD4 by edad_cat
## Kruskal-Wallis chi-squared = 4.0619, df = 2, p-value = 0.1312
```

```
kruskal.test(CV ~ edad_cat, data=Tabla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CV by edad_cat
## Kruskal-Wallis chi-squared = 1.5302, df = 2, p-value = 0.4653
```

En todos los casos el p-valor es superior a 0.05, por lo que no podemos rechazar la hipótesis nula de que no hay diferencias entre los diferentes grupos de género o edad.

4. ¿Existe prevalencia de una mutación sobre un subtipo?

```
DATOSCSV <- read.csv("~/tmp/uoc/uoc_SAD/DatosEntrada/tablaMaster_corregida.csv",
                 header=T, dec=".", na.strings="NA",sep="")

local({
.Table <- with(DATOSCSV, table(SUBTIPO))
cat("\ncounts:\n")
print(.Table)
cat("\npercentages:\n")
print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## SUBTIPO
##  01_AE  02_AG 06_CPX  14_BG 36_CPX     A1      B      C      D     F1
##      2     22      4      1      2      7    216      5      5      8
##      G      J
##      9      1
##
## percentages:
## SUBTIPO
##  01_AE  02_AG 06_CPX  14_BG 36_CPX     A1      B      C      D     F1
##   0.71   7.80   1.42   0.35   0.71   2.48  76.60   1.77   1.77   2.84
##      G      J
##   3.19   0.35
```

Como puede observarse el subtipo B es el predominante en la población española con un 76% seguido por las forma recombinante URF de las cuales la mayoritaria de este tipo es la 02_AG siendo un 8% del total y un 71% de las formas URF. Todos estos subtipos corresponden a la variante vih-1, la cual se encuentra principalmente en América y Europa. A la vista de los resultados resulta obvio que España tenga predominancia del subtipo B debido a efectos migratorios. Sin embargo, como se denota por los resultados la variante 02_AG, se está incrementando en España debido a la migración africana que recibe nuestro país, pero a la contra es factible preveer que personas que hayan migrado desde África y libres del virus queden infectados por la variante Europea.

```
local({
.Table <- with(DATOSCSV, table(MUTACIONES1))
cat("\ncounts:\n")
print(.Table)
cat("\npercentages:\n")
print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## MUTACIONES1
##  NO  SI
## 183  99
##
## percentages:
## MUTACIONES1
##    NO    SI
## 64.89 35.11
```

De las 282 personas tratadas y documentadas para este tipo, solamente en 99 de ellas se les detectó una o varias mutaciones de resistencia frente a los fármacos anti retrovirales. El estudio de las diferentes mutaciones es complejo ya que pueden darse circunstancias de que un paciente presente un único tipo de mutación (proteasas o retrotranscriptasas) o que presente una colección de ellas. Dentro de las mutaciones las más comunes son las que afectan a las retrotranscriptasas siendo la 103N, la 184V y la 138K las más comunes. Para determinar si existe relación entre ambas variables atenderemos al siguiente contraste de hipótesis: Ho: Las dos variables están relacionadas Ha: Las dos variables son independientes. Para determinar esto se realizara un test de independencia Chi Square

```
local({
.Table <- xtabs(~MUTACIONES1+SUBTIPO, data=DATOSCSV)
cat("\nFrequency table:\n")
print(.Table)
.Test <- chisq.test(.Table, correct=FALSE)
print(.Test)})
```

```
##
## Frequency table:
##             SUBTIPO
## MUTACIONES1 01_AE 02_AG 06_CPX 14_BG 36_CPX  A1   B   C   D  F1   G   J
##          NO     1    18      3     1      1   4 136   3   3   7   5   1
##          SI     1     4      1     0      1   3  80   2   2   1   4   0
##
##  Pearson's Chi-squared test
##
```

```
## data:  .Table
## X-squared = 7.1991, df = 11, p-value = 0.7827
```

En este caso se rechaza la Ho debido a que el p-valor es mayor de 0.05. Esto indica que las variables son independientes de modo que no importa que subtipo del virus presente el paciente podrá darse o no mutaciones en el mismo.

5. ¿Existe relación entre la edad del paciente y las mutaciones que se desarrollan?

Para determinar si existe relación entre ambas variables atenderemos al siguiente contraste de hipótesis:

Ho: Las dos variables estan relacionadas.

Ha: Las dos variables son independientes.

```
local({
.Table <- xtabs(~MUTACIONES1+SEXO, data=DATOSCSV)
cat("\nPercentage table:\n")
print(rowPercents(.Table))
prop.test(.Table, alternative='two.sided', conf.level=.95, correct=FALSE)
})
```

```
##
## Percentage table:
##             SEXO
## MUTACIONES1 HOMBRE MUJER Total Count
##          NO   79.8  20.2   100   183
##          SI   72.7  27.3   100    99
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  .Table
## X-squared = 1.8222, df = 1, p-value = 0.1771
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03473182  0.17581478
## sample estimates:
##    prop 1    prop 2
## 0.7978142 0.7272727
```

En este caso aceptamos la Ho debido a que el p-valor es mayor de 0.05. Esto indica que las variables son independientes de modo que no importa que los hombres son más propensos a desarrollar mutaciones como que no desarrollarlas.

6. ¿Hay alguna relación entre el subtipo y el CD4/carga viral?

El objetivo que buscamos con esta pregunta es saber si hay algun subtipo más agresivo que otro (reflejándose en mayores niveles de carga viral(CV)) o si algun subtipo produce una mayor respuesta del sistema inmune (niveles de CD4).

En primer lugar, comprobamos cómo de diferentes son los valores para cada subtipo.

En cuanto a CV:

```
numSummary(Tabla[,"CV"], groups=Tabla$SUBTIPO,
statistics=c("mean", "sd", "IQR","quantiles"), quantiles=c(0,.25,1))
```

```
##             mean         sd      IQR     0%       25%     100% data:n
## 01_AE  119250.00   9545.942   6750.0 112500 115875.00   126000      2
## 02_AG  219723.18 401178.082 158525.0    100  13975.00  1420000     22
## 06_CPX 302200.00 278227.772 347300.0  88800  88950.00   674000      4
## 14_BG   27900.00         NA      0.0  27900  27900.00    27900      1
## 36_CPX  25272.50  17780.200  12572.5  12700  18986.25    37845      2
## A1      27954.29  58350.621  17666.0     62    174.00   158000      7
## B      222757.74 862638.677 105709.0     30   3541.00 10000000    216
## C       27403.40  32860.491  47132.0    349   2368.00    74200      5
## D      115868.20 137219.174 177965.0    406   9035.00   322000      5
## F1     129774.12 198669.585 116660.2   7100  12639.75   525000      8
## G       50500.22  63340.234  80980.0     52   4620.00   169000      9
## J      129000.00         NA      0.0 129000 129000.00   129000      1
```

Sí que vemos diferencias dependiendo del subtipo. El subtipo con mayor CV es *06_CPX* con diferencia (302200). El que menos, el subtipo *C* (27403).

Y respecto a CD4:

```
numSummary(Tabla[,"CD4"], groups=Tabla$SUBTIPO,
statistics=c("mean", "sd", "IQR","quantiles"), quantiles=c(0,.25,1))
```

```
##            mean       sd   IQR  0%    25% 100% data:n data:NA
## 01_AE  288.5000 191.6259 135.5 153 220.75  424      2       0
## 02_AG  282.0476 217.7187 274.0  28 128.00  906     21       1
## 06_CPX 222.0000 128.2342 183.0 103 122.50  373      4       0
## 14_BG  310.0000       NA   0.0 310 310.00  310      1       0
## 36_CPX 381.0000 165.4630 117.0 264 322.50  498      2       0
## A1     437.7143 375.6105 497.5  50 124.00 1081      7       0
## B      375.9364 233.2343 257.0   2 233.00 1650    214       2
## C      352.6000 244.8567 355.0  56 146.00  640      5       0
## D      312.2000 199.9942 270.0 117 126.00  595      5       0
## F1     330.5000 248.6242 275.5  15 155.75  786      8       0
## G      402.5556 246.7150 351.0  84 279.00  817      9       0
## J      346.0000       NA   0.0 346 346.00  346      1       0
```

En este caso no hay grandes diferencias con lo que podemos concluir que los distintos subtipos no parecen indicar un mayor o menor número de linfocitos T-CD4.

Para comprobar si estas diferencias son significativas, podemos utilizar la prueba de Kruskal-Wallis:

```
kruskal.test(CD4 ~ SUBTIPO, data=Tabla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CD4 by SUBTIPO
## Kruskal-Wallis chi-squared = 6.7964, df = 11, p-value = 0.8153
```

```
kruskal.test(CV ~ SUBTIPO, data=Tabla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CV by SUBTIPO
## Kruskal-Wallis chi-squared = 14.983, df = 11, p-value = 0.1833
```

Se analizan los datos con el fin de comprobar si existen diferencias sustanciales entre ellos (o dicho de otra manera, si todos los datos pueden pertener a una misma población). En este caso, dado que p-valor es superior a 0.05 podemos afirmar que no existen diferencias importantes entre el subtipo y el CD4 o CV, es decir aceptamos la Hipotesis nula.

Como último test, utilizamos el algortimo Chi-cuadrado.

```
Table <- xtabs(CV~SUBTIPO, data=Tabla)
Table
```

```
## SUBTIPO
##    01_AE    02_AG   06_CPX    14_BG   36_CPX       A1        B        C
##   238500  4833910  1208800    27900    50545   195680 48115671   137017
##        D       F1        G        J
##   579341  1038193   454502   129000
```

```
Test <- chisq.test(Table, correct=FALSE)
Test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Table
## X-squared = 435900000, df = 11, p-value < 2.2e-16
```

En este caso, el p-valor es inferior a 0.05 por lo que podemos rechazar la hipotesis nula y aceptar la alternativa, es decir, que sí que esta relacionado el nivel de carga viral con el subtipo. Este resultado se contradice con el de Kruskal-Wallis por lo que no estoy completamente seguro de si el Chi-cuadrado esta aplicado correctamente. El algoritmo debe ser aplicado sobre variables cualitativas y en nuestro caso SUBTIPO sí lo es pero no CV.

Fuentes: http://es.slideshare.net/navarroenrique/gua-contraste-de-hiptesis-blog

7. ¿Existe alguna relación entre la carga viral y los CD4?

Comprobamos el índice de correlación:

```
cor(Tabla[,"CV"], Tabla[,"CD4"], use="complete")
```

```
## [1] -0.1539821
```

El valor es negativo próximo a 0 por lo que ambas variables estan negativamente relacionadas (es decir, que cuando una se incrementa la otra decrece) pero débilmente (el valor es próximo a 0).

Si agrupamos los valores de CD4 para encontrar una relación con la carga viral (CV):

```
Tabla <- within(Tabla, {
CD4_group <- recode(CD4,
'0:200 = "Muy Bajo"; 200:400 = "Bajo"; 400:600 = "Medio"; 600:800 = "Alto"; 800:1100 = "Muy alto"',
as.factor.result=TRUE)
})

local({
.Table <- with(Tabla, table(CD4_group))
cat("\ncounts:\n")
print(.Table)
cat("\npercentages:\n")
print(round(100*.Table/sum(.Table), 2))
})
```

```
##
## counts:
## CD4_group
##     1650     Alto     Bajo    Medio Muy alto Muy Bajo
##        1       20      102       71       14       71
##
## percentages:
## CD4_group
##     1650     Alto     Bajo    Medio Muy alto Muy Bajo
##     0.36     7.17    36.56    25.45     5.02    25.45
```

La mayoría de muestras contienen unos niveles inferiores a 800 cel/ml.

```
numSummary(Tabla[,"CV"], groups=Tabla$CD4_group,
           statistics=c("mean", "sd", "IQR","quantiles"), quantiles=c(0,.25,.5,.75,1))
```

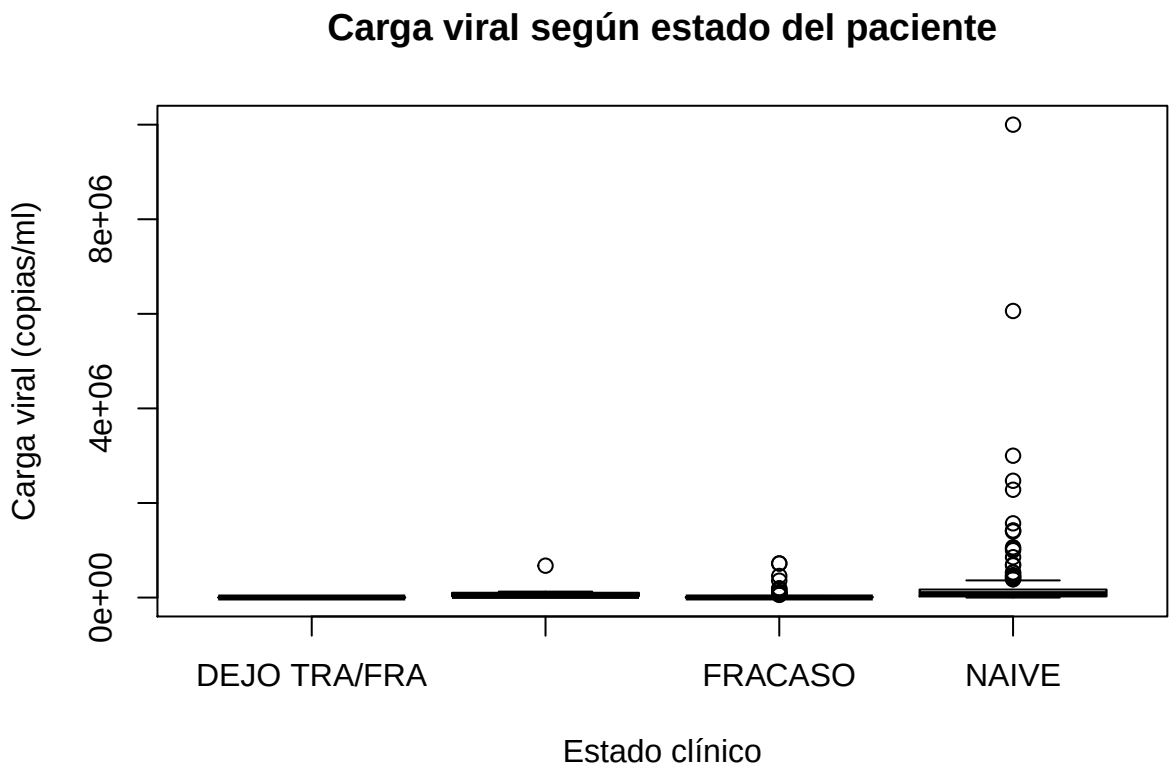
```
##                mean         sd       IQR 0%      25%   50%       75%
## 1650         86.000         NA      0.00 86    86.00    86     86.00
## Alto      47144.750  116185.59  27730.75 62   219.25  3210  27950.00
## Bajo     282750.176 1064054.64 118275.00 34 12475.00 58658 130750.00
## Medio     65060.197  145340.77  51437.00 54  2713.00 24800  54150.00
## Muy alto   5417.571   13111.23   1020.50 30   125.75   371   1146.25
## Muy Bajo 313643.606  800306.66 220650.00 58 14350.00 90400 235000.00
##              100% data:n
## 1650           86      1
## Alto       460000     20
## Bajo     10000000    102
## Medio      853336     71
## Muy alto    48000     14
## Muy Bajo  6060000     71
```

Efectivamente, comprobamos que los niveles de CV son menores a medida que los niveles de CD4 son mayores. Entre un nivel medio-alto (400-800 cel/mel) no hay mucha diferencia pero si nos vamos a los extremos la diferencia es considerable.

Gráficos

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

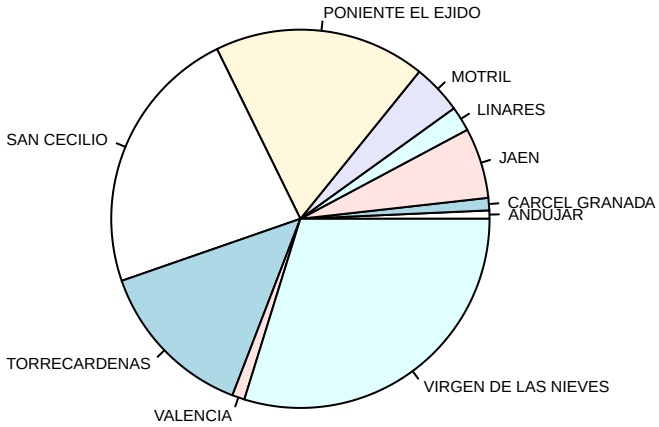
```
boxplot(CV~ESTADO, data=datosCSV, id.method="y", ylab="Carga viral (copias/ml)",
        main= "Carga viral según estado del paciente", xlab="Estado clínico")
```



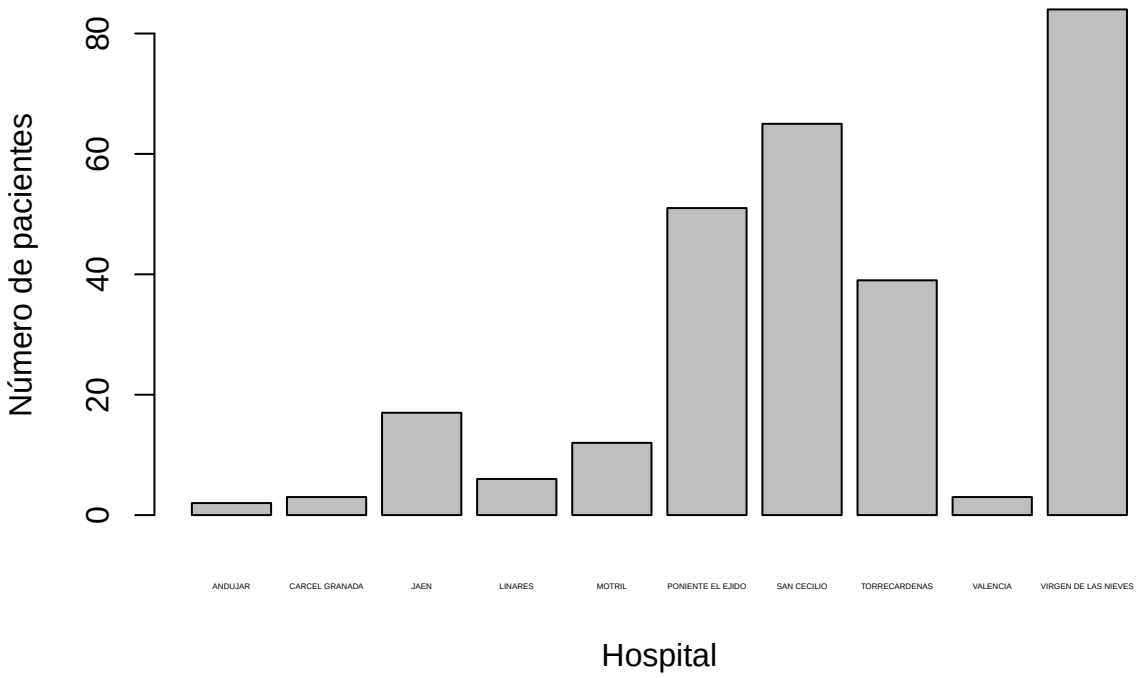
2. ¿ Como se distribuyen los individuos en los hospitales de procedencia?

```
with(datosCSV, pie(table(LOCALIDAD), labels=levels(LOCALIDAD),
                   main="Pacientes según hospital",cex=0.5))
```

Pacientes según hospital

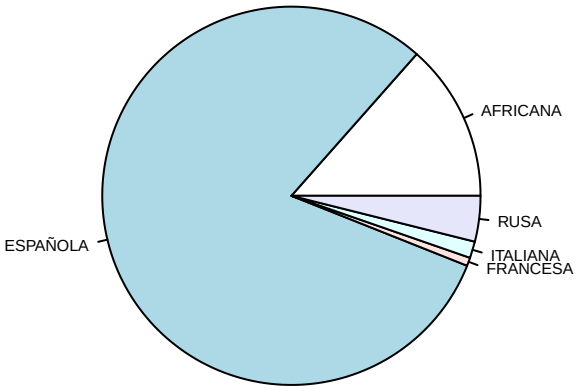



```
plot(datosCSV[1], cex.names =0.29, ylab = "Número de pacientes", xlab = "Hospital")
```



Para saber más acerca de cómo están distribuidas las nacionalidades de los pacientes VIH podemos hacer la siguiente gráfica, donde se observa (como era de esperar) que la nacionalidad Española es la más común.

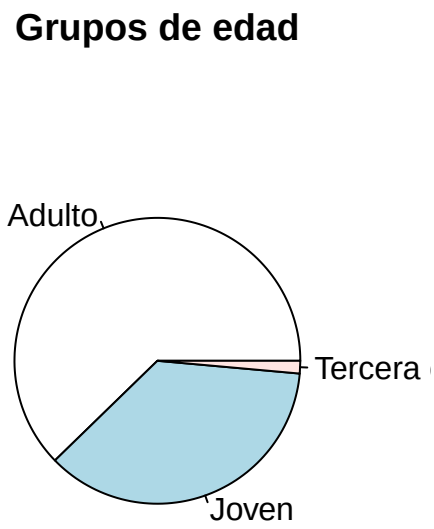
```
with(datosCSV, pie(table(nacionalidad), labels=levels(nacionalidad),
                   main="Frecuencia pacientes según nacionalidad",cex=0.5))
```

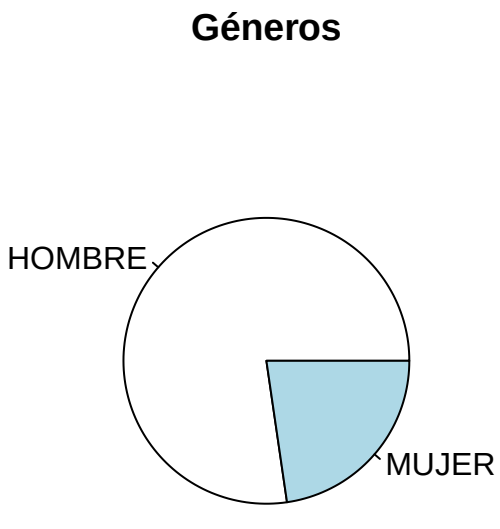


3. ¿Existe algún tipo de relación entre la carga viral o los niveles de CD4, y el sexo o la edad del paciente?

Represento las proporciones de la población de los grupos según el **grupo de edad** y el **género**:

```
layout ( matrix (c(1 ,2) , 1, 2, byrow = TRUE ))
pie(table(Tabla$edad_cat), labels=levels(Tabla$edad_cat),
    main="Grupos de edad")
pie(table(Tabla$SEXO), labels=levels(Tabla$SEXO), main="Géneros")
```





Para representar el resto de los gráficos usamos el paquete lattice.

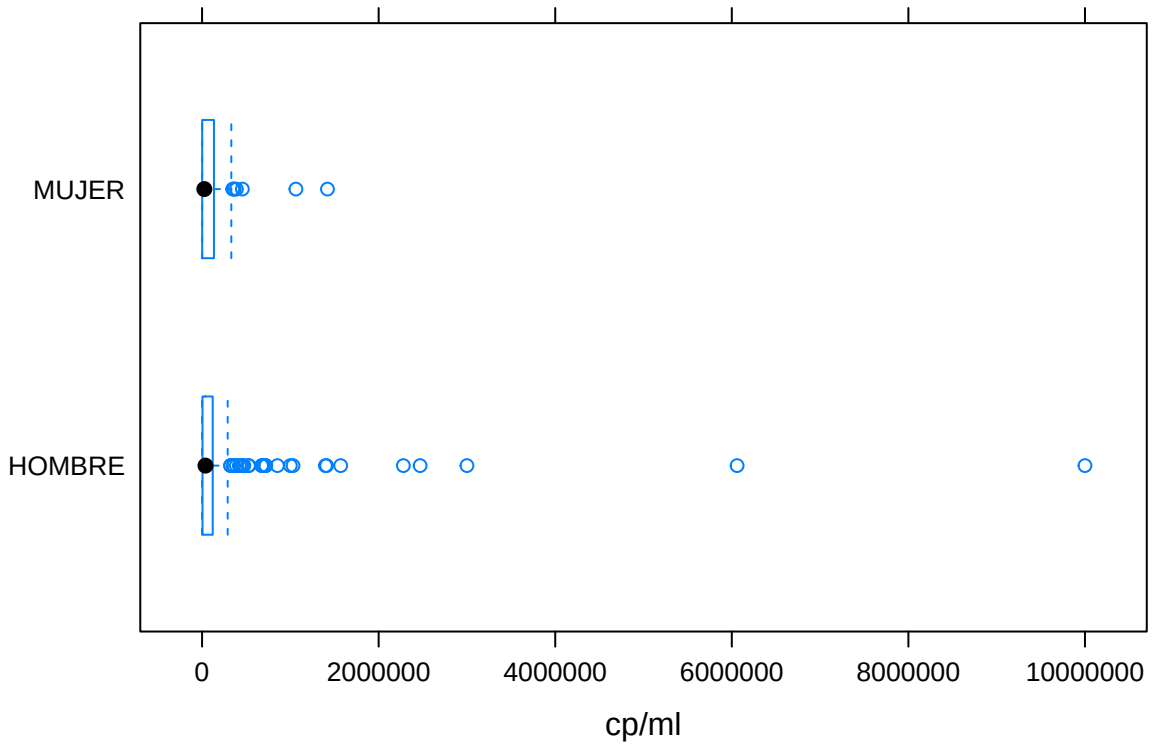
```
library(lattice)
```

1. CV y CD4 según género.

Represento los niveles de **CV** en función del **género**.

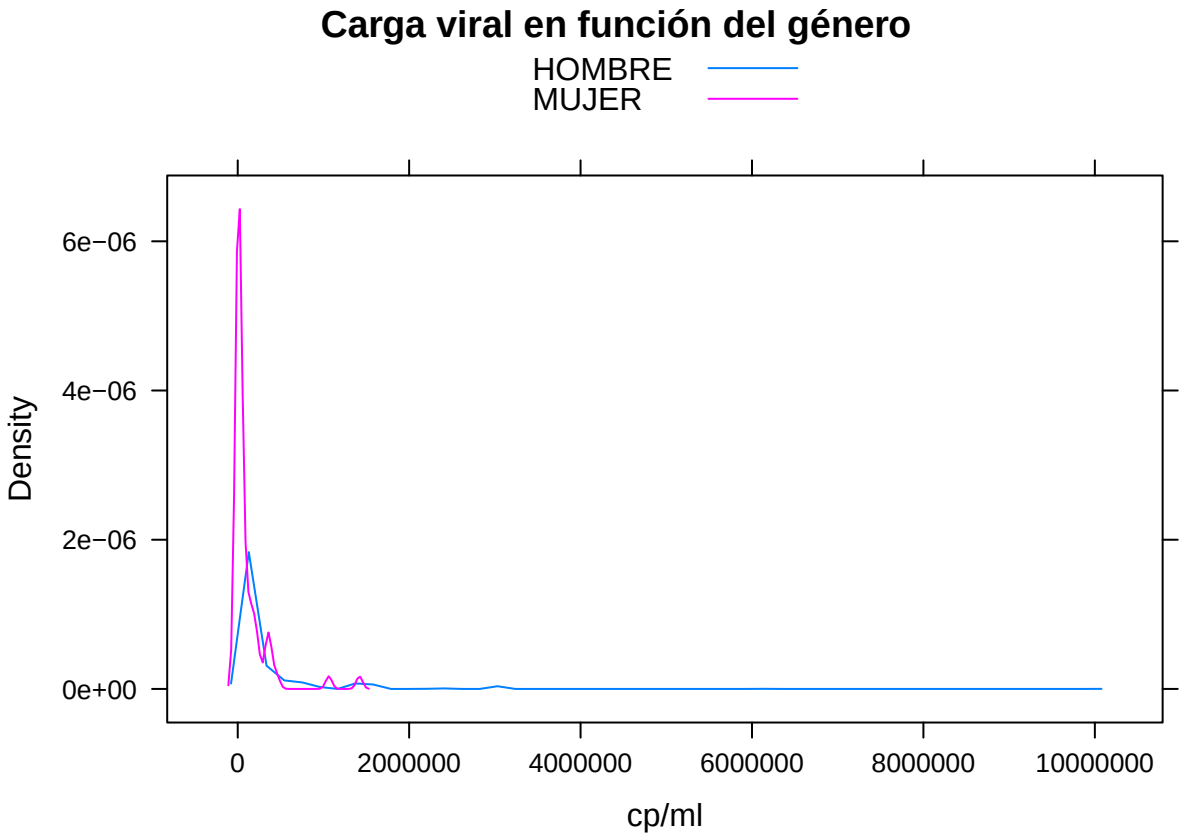
```
> Diagrama de cajas
```

```
bwplot(factor(SEXO) ~ CV, data = Tabla, xlab="cp/ml",
       main = "Carga viral en función del género")
```



> Función de densidad

```
densityplot(~ CV, data = Tabla, groups = SEXO, plot.points=FALSE,
            auto.key=TRUE, main= "Carga viral en función del género",  xlab="cp/ml" )
```

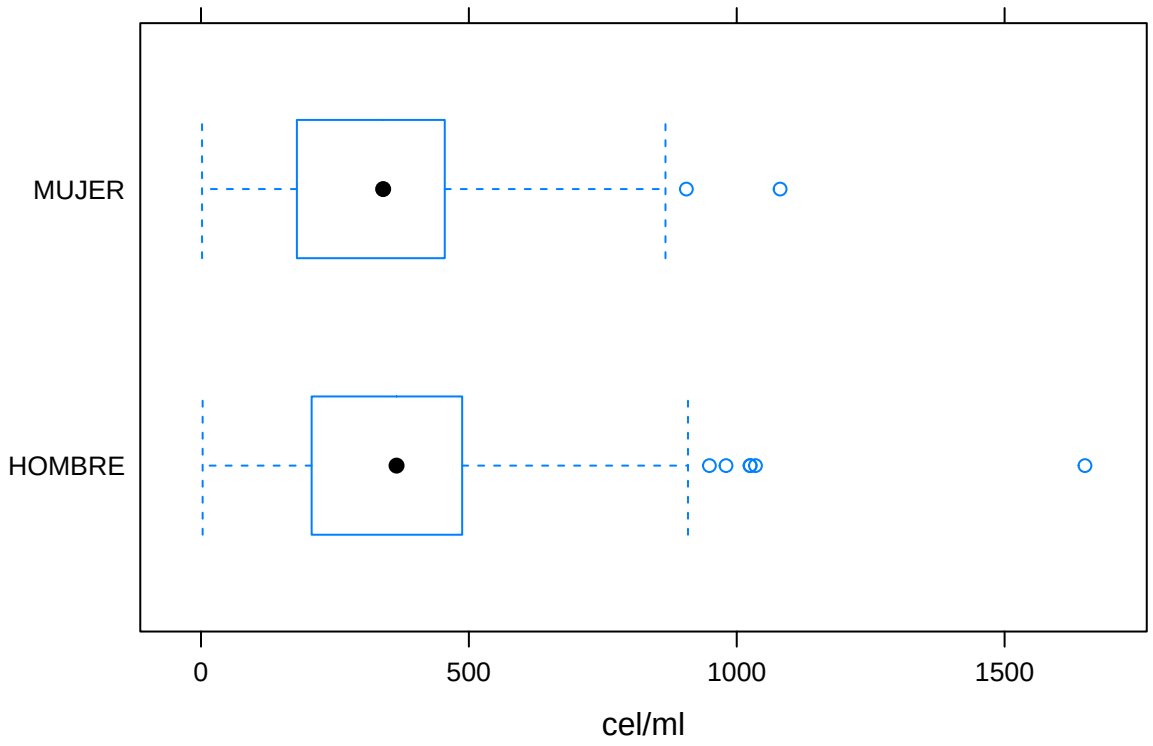


Se puede ver como, a pesar de no haber grandes diferencias en las medias de los diferentes grupos, la distribución varía un poco. En mujeres la población es más homogénea, habiendo un mayor número de casos alrededor de la media. Por el contrario, en hombres se puede ver que hay una mayor variabilidad, encontrando casos con un mayor número de CV en sangre.

Represento los niveles de **CD4** en función del **género**:

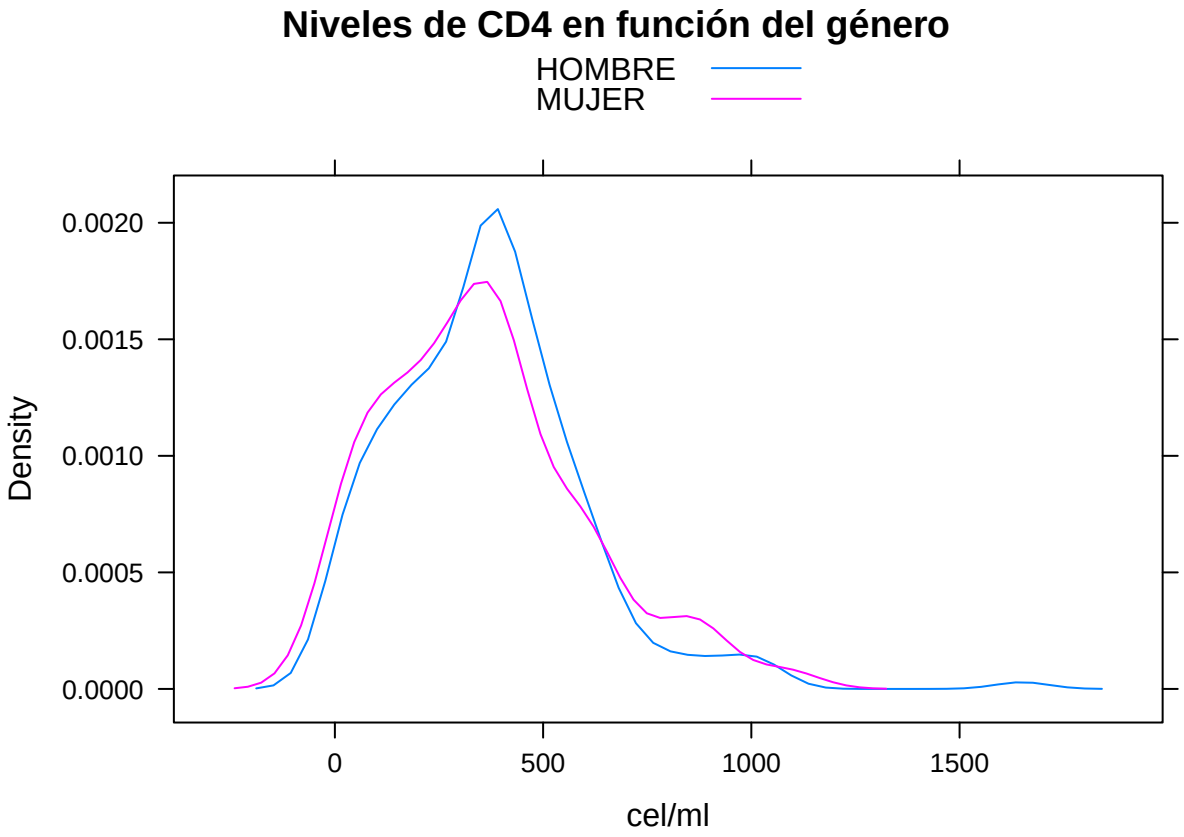
```
> Diagrama de cajas
```

```
bwplot(factor(SEXO)~ CD4, data = Tabla, xlab="cel/ml",
       main = "Niveles de CD4 en función del género")
```



> Función de densidad

```
densityplot(~ CD4, data = Tabla, groups = SEXO, plot.points=FALSE,
            auto.key=TRUE, main= "Niveles de CD4 en función del género",  xlab="cel/ml" )
```



Al igual que ocurre con la CV, los niveles de CD4 no varían mucho entre los distintos sexos. Sin embargo, en este caso las distribuciones de ambos sexos son muy parecidas.

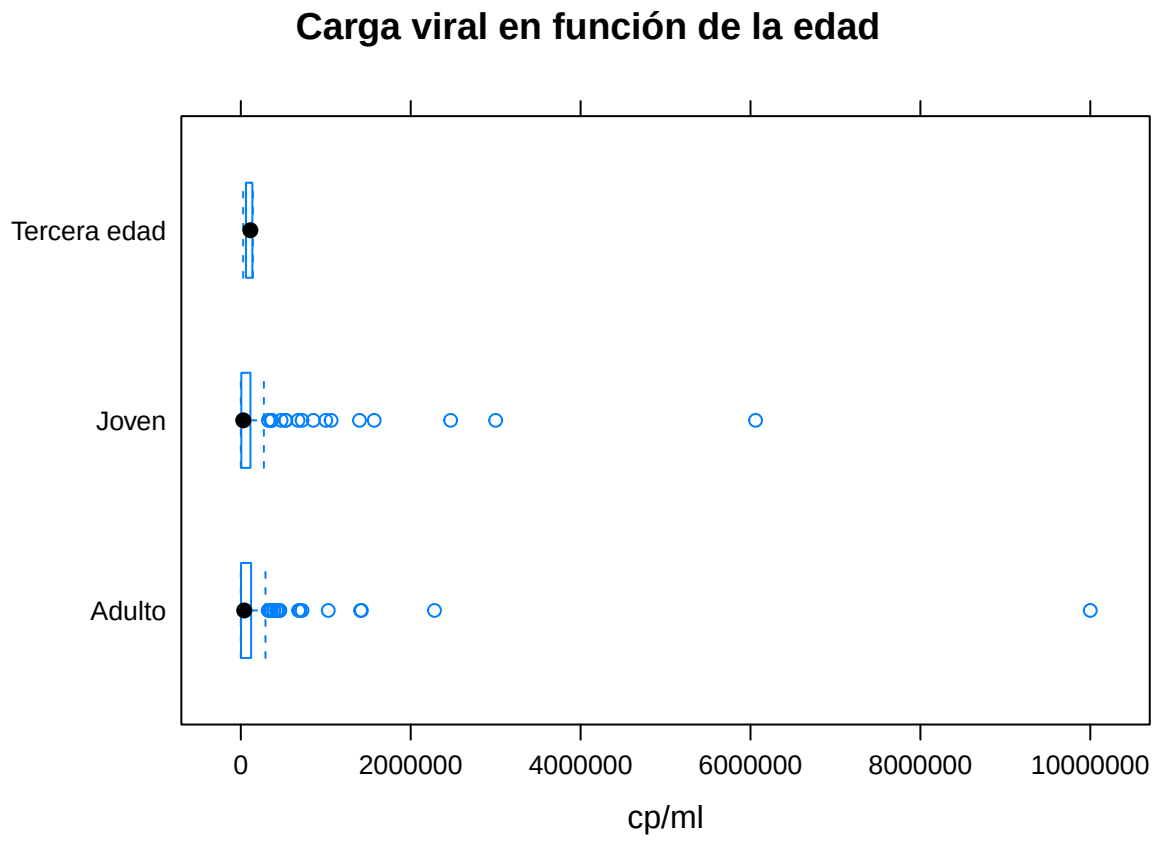
2. CV y CD4 según edad.

Ahora voy a ver si la CV o los niveles de CD4 varían según la edad del paciente:

Represento los niveles de **CV** en función de la **edad**:

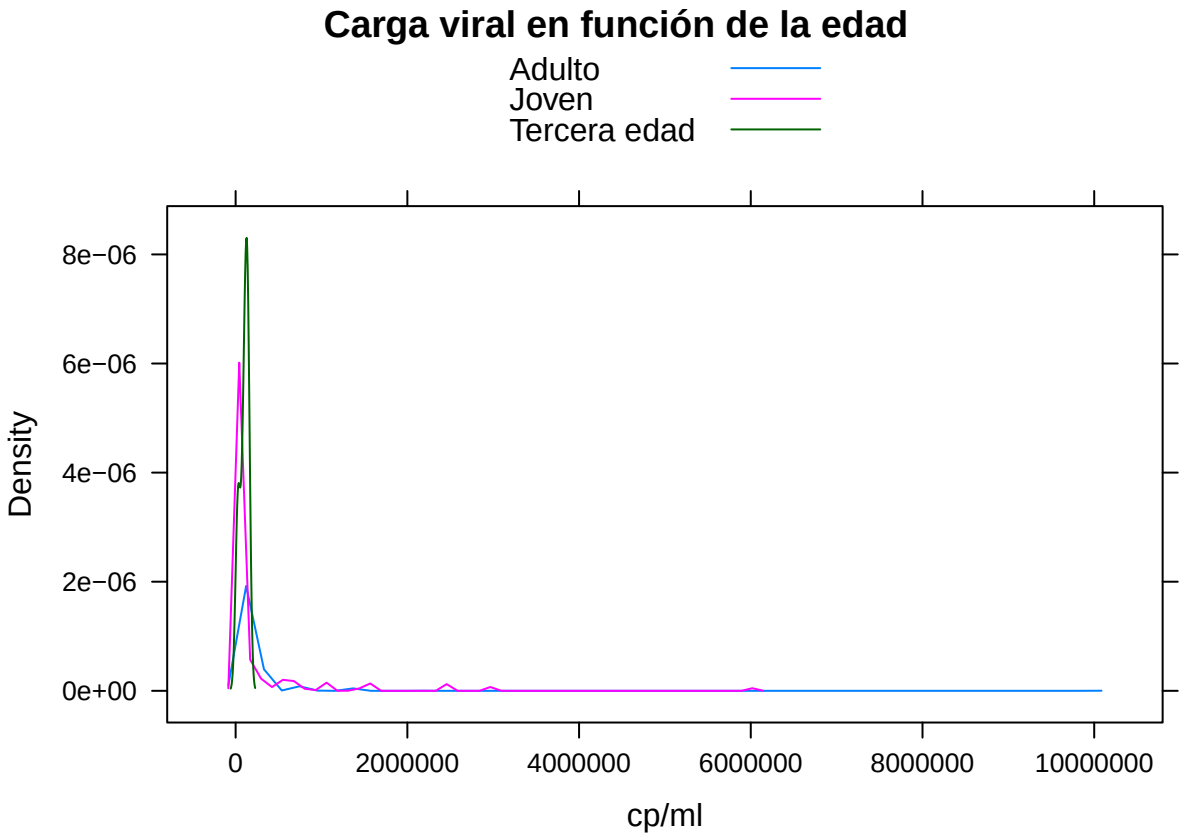
```
> Diagrama de cajas
```

```
bwplot(factor(edad_cat)~ CV , data = Tabla, xlab="cp/ml",
       main = "Carga viral en función de la edad")
```



> Función de densidad

```
densityplot(~ CV, data = Tabla, groups = edad_cat, plot.points=FALSE,
            auto.key=TRUE, main= "Carga viral en función de la edad",  xlab="cp/ml" )
```

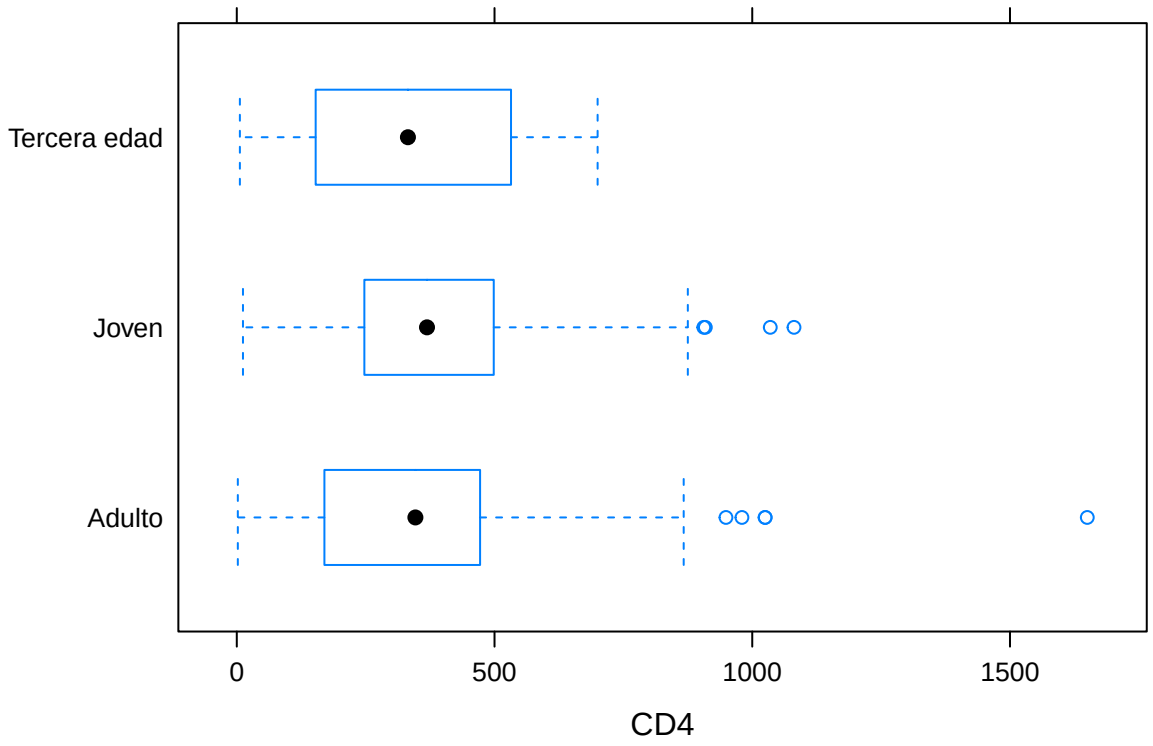


Se puede ver que no hay grandes diferencias en las medias de los diferentes grupos. Sin embargo, la distribución varía. En el grupo de la tercera edad se puede var que los datos son mucho más homogéneos comparados con el grupo de jóvenes. Sin embargo, hay que tener en cuenta que el número de individuos en el grupo de la tercera edad es de únicamente 4. Si comparamos el grupo de jóvenes con el de adultos de puede ver que los valores de CV son mucho más homogéneos en gente joven.

Represento los niveles de **CD4** en función de la **edad**:

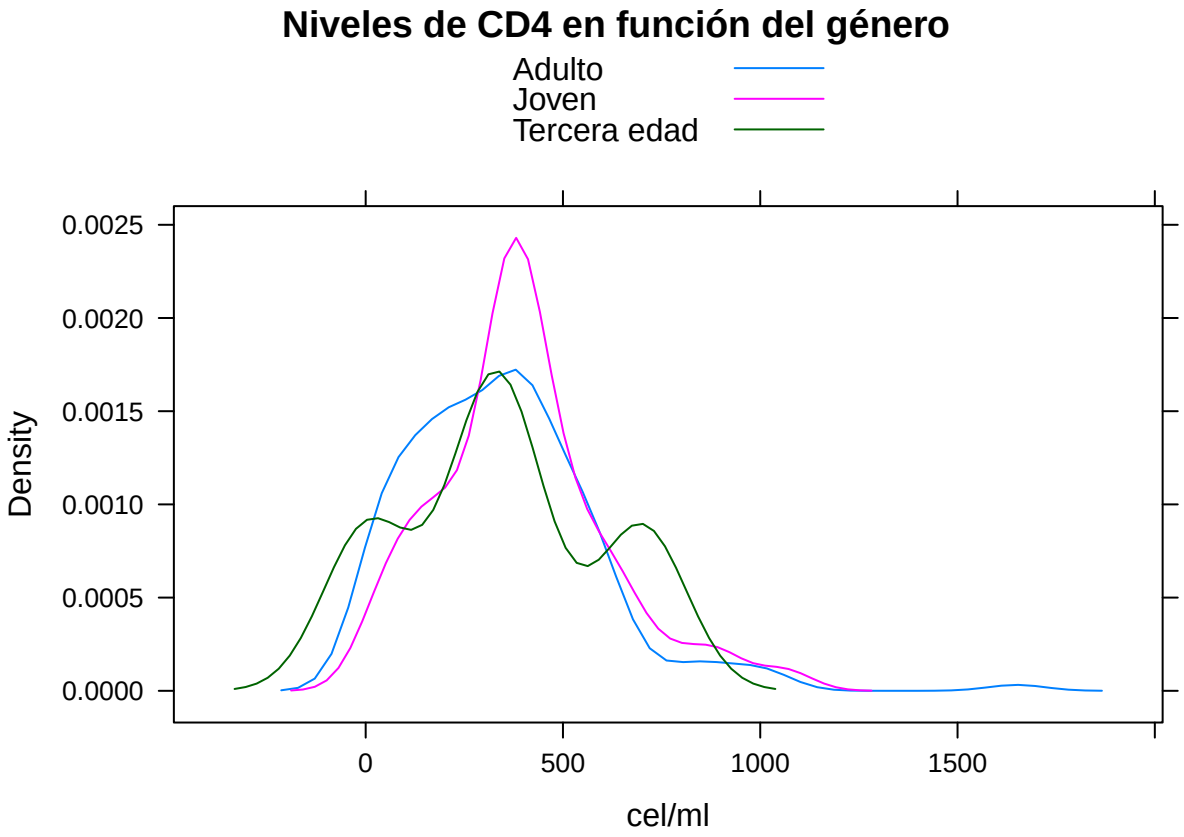
```
> Diagrama de cajas
```

```
bwplot(factor(edad_cat)~ CD4, data = Tabla, main= "Niveles de CD4 en función del género")
```




```
> Función de densidad
```

```
densityplot(~ CD4, data = Tabla, groups = edad_cat, plot.points=FALSE,
            auto.key=TRUE, main= "Niveles de CD4 en función del género",  xlab="cel/ml" )
```



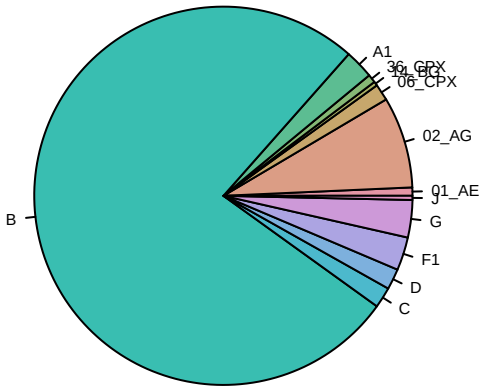
Al igual que ocurre con la CV, los niveles de CD4 no varían mucho entre los distintos grupos de edad. Al igual que ocurría con la CV, las posibles diferencias en los perfiles de distribución de los niveles de CD4 se deban seguramente a las diferencias en el número de individuos de cada grupo. Mientras que en adultos y en jóvenes el número de individuos es bastante elevado (175 y 102, respectivamente), en el grupo de la tercera edad sólo tenemos 4 individuos. Sin embargo, centrándonos únicamente en el grupo de jóvenes y adultos, se puede observar como, al igual que vimos antes con niveles de CV, hay una mayor dispersión de datosreferentes a los niveles de CD4 en el grupo de adultos, comparados con el de jóvenes.

Parece que, mientras que en gente jóven los niveles de CD4 y CV son más homogéneos, en adultos existe una mayor variabilidad.

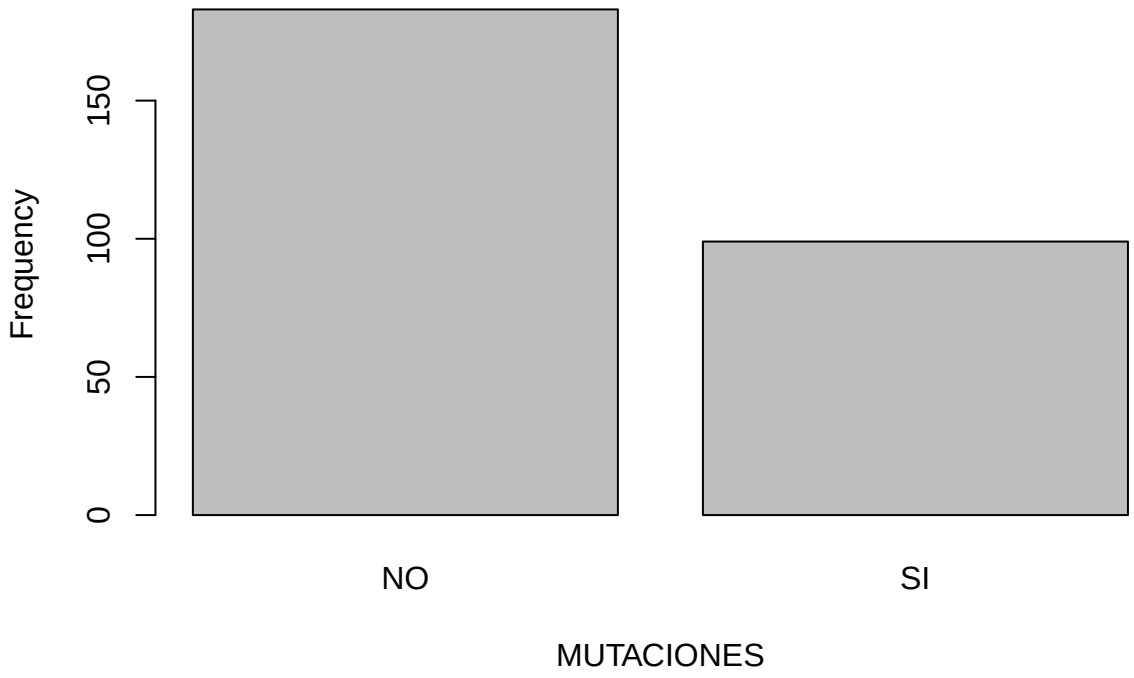
4. ¿Existe prevalencia de una mutación sobre un subtipo?

```
library(colorspace)
with(DATOSCSV, pie(table(SUBTIPO), labels=levels(SUBTIPO),
                   main="Frecuencia de casos agrupados por subtipo",cex=0.5, col=rainbow_hcl(length(lev
```

Frecuencia de casos agrupados por subtipo

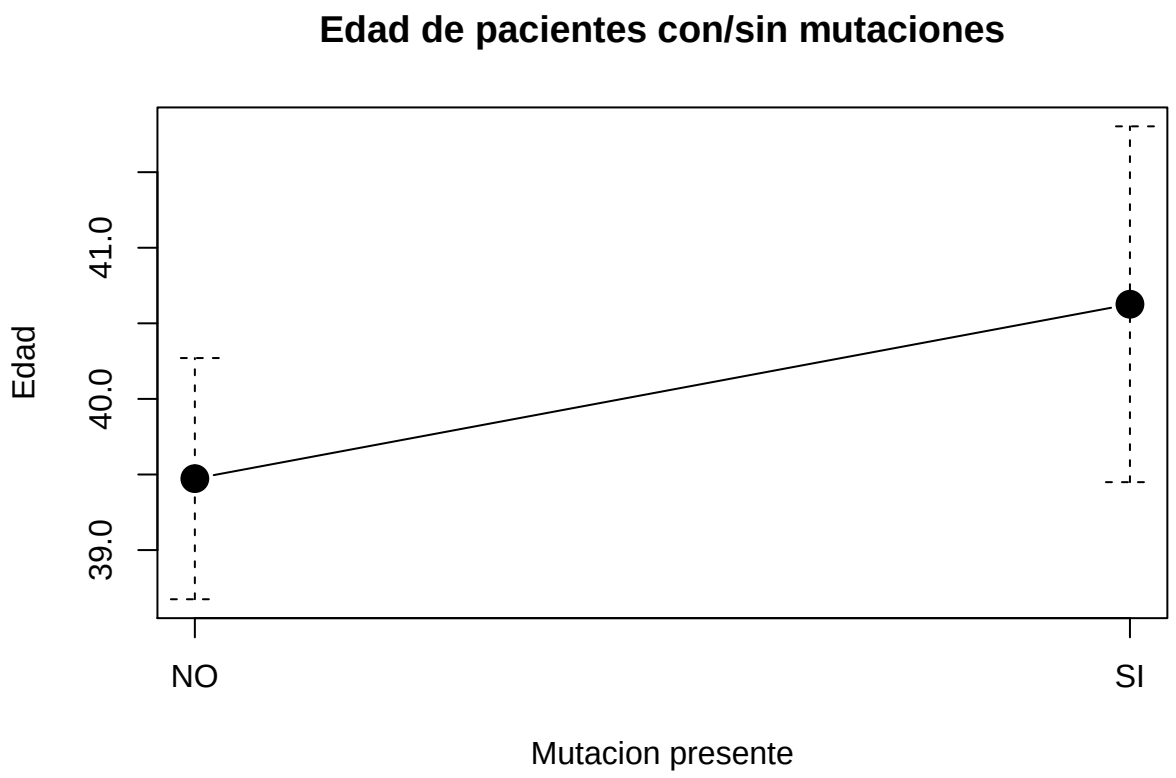



```
with(DATOSCSV, Barplot(MUTACIONES1, xlab="MUTACIONES", ylab="Frequency"))
```



5. ¿Existe relación entre la edad del paciente y las mutaciones que se desarrollan?

```
with(DATOSCSV, plotMeans(edad, MUTACIONES1, error.bars="se", ylab="Edad",
                         xlab="Mutacion presente",
                         main= "Edad de pacientes con/sin mutaciones"))
```

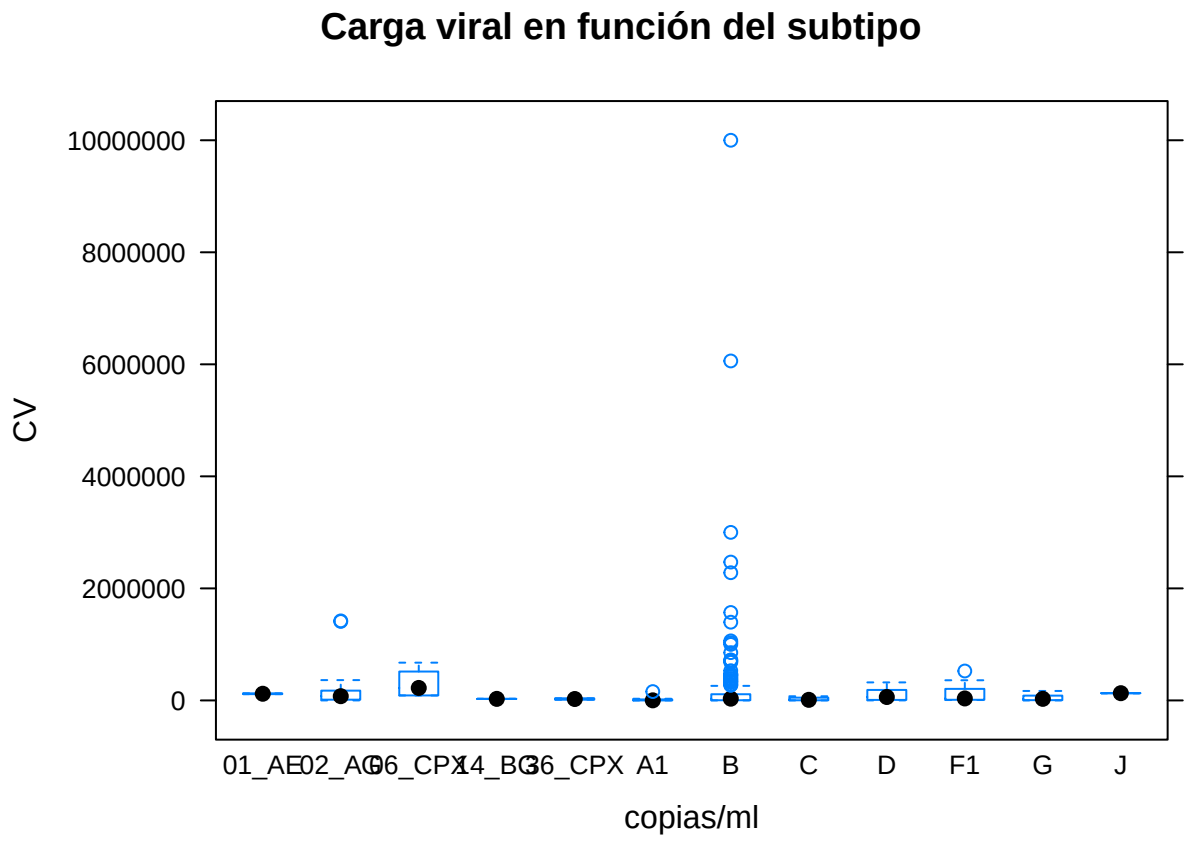


Como puede observarse la media de edad de los pacientes que no presentan mutaciones de resistencia a farmacos es menor que la de aquellos que si presentan variantes del virus con resistencia a ciertos antiretrovirales.

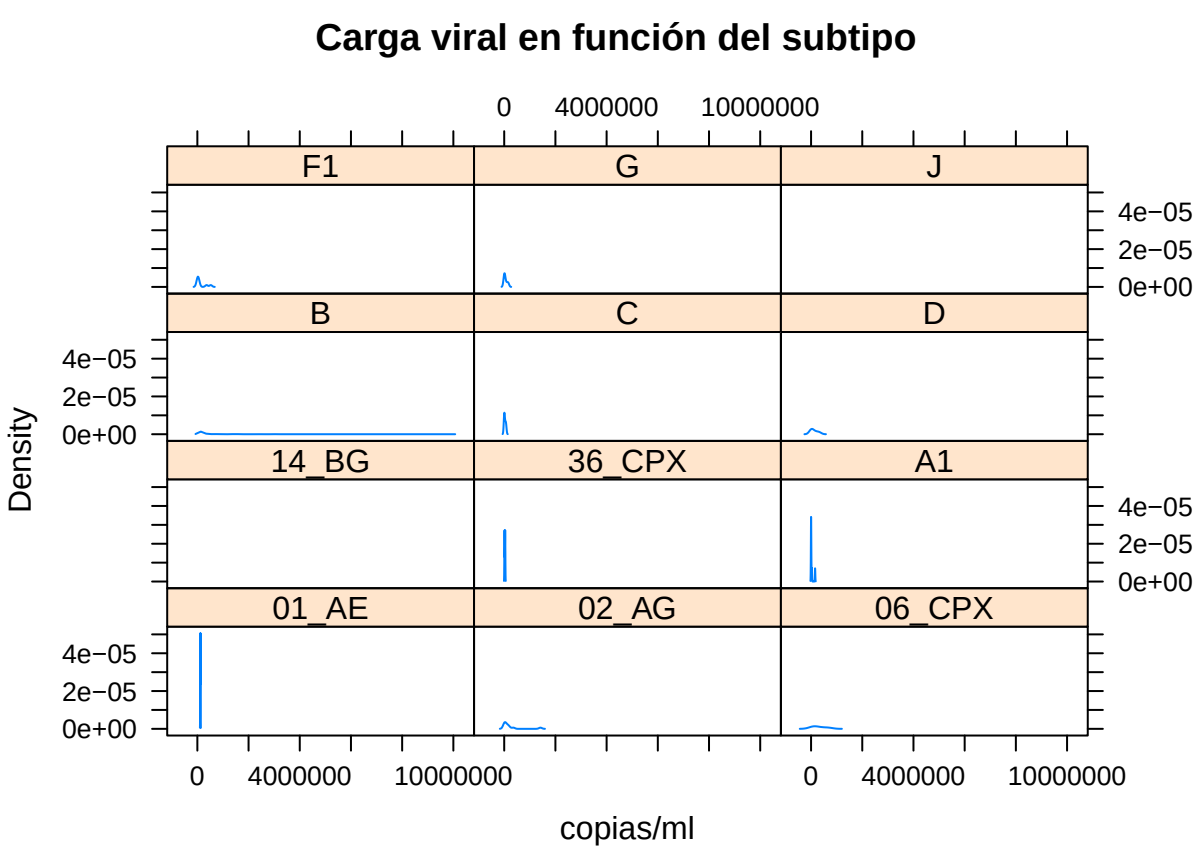
6. ¿Hay alguna relación entre el subtipo y el CD4/carga viral?

Primero comparamos los niveles de carga viral (CV) en distintos subtipos:

```
bwplot( CV ~ factor(SUBTIPO), data = Tabla, xlab="copias/ml",
       main = "Carga viral en función del subtipo")
```



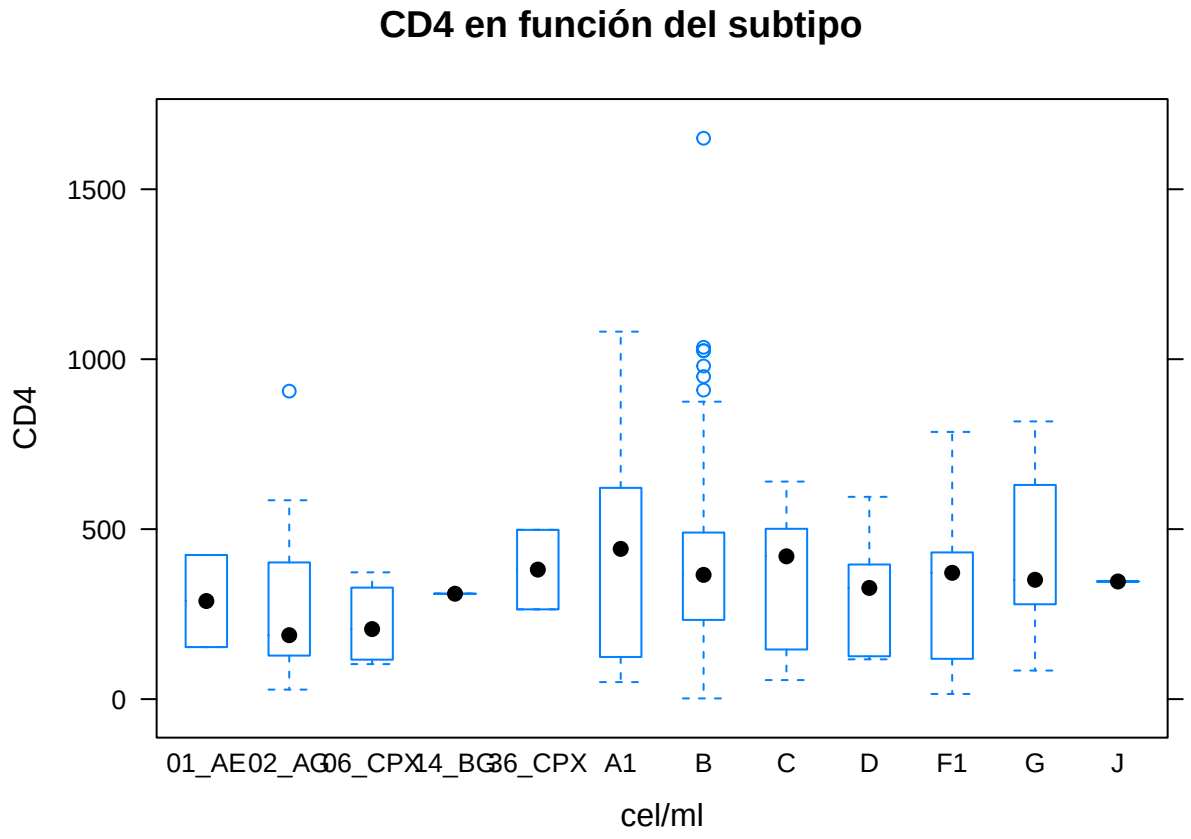

```
densityplot(~ CV|SUBTIPO, data = Tabla, plot.points=FALSE, auto.key=TRUE,
            main= "Carga viral en función del subtipo", xlab="copias/ml",layout=c(3,4))
```



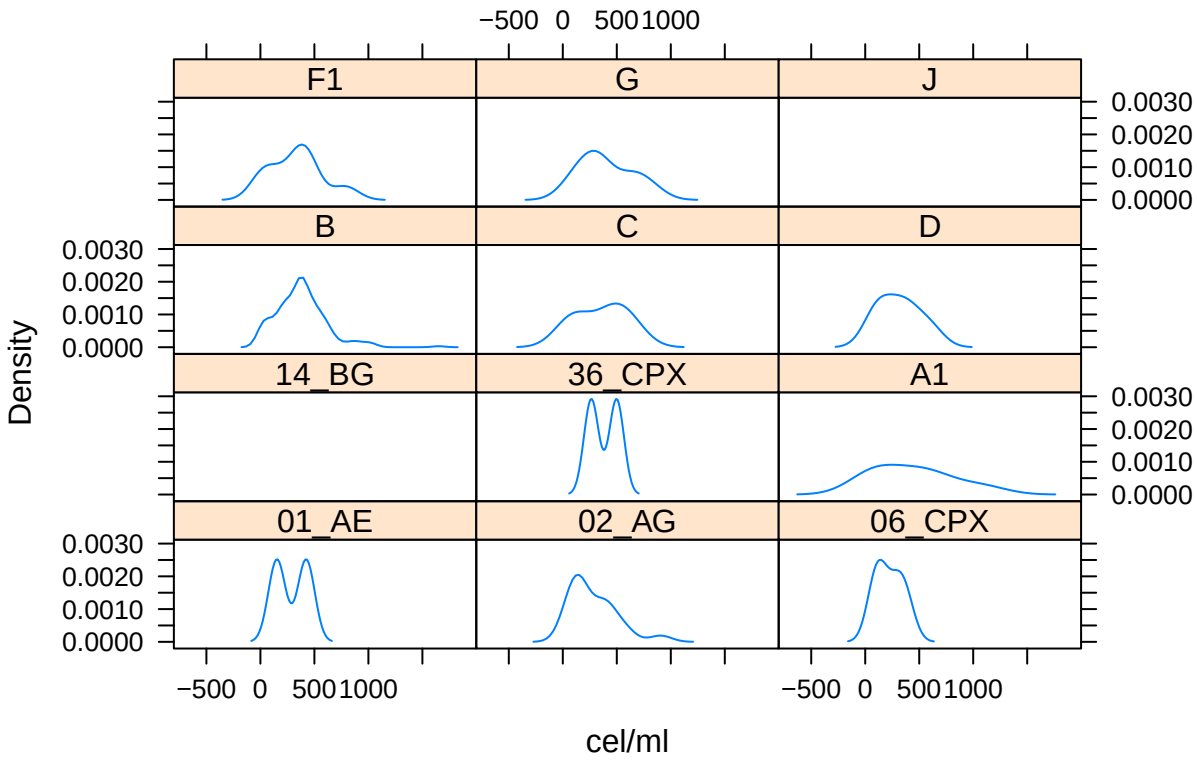
Hay bastante diversidad. Algunos subtipos son más propensos (01_AE, A1, 36_CPX etc) y otros contienen un nivel mucho más alto de CV (V, 02_AG, 06_CPX).

Los mismos analisis respecto al CD4:

```
bwplot(CD4~ factor(SUBTIPO), data = Tabla,
       xlab=list("cel/ml"), main = "CD4 en función del subtipo")
```




```
densityplot(~ CD4|SUBTIPO, data = Tabla, plot.points=FALSE, auto.key=TRUE,
            main= "CD4 en función del subtipo", xlab="cel/ml",layout=c(3,4))
```



Aquí los datos son más homogéneos aunque siguen existiendo diferencias. Hay varios subtipos son datos (J y 14_BG). Los niveles de CD4 con valores bajos son más numerosos en un mismo subtipo (01_AE, 36_CPX, etc) pero también existen subtipos con un número variable de distintos niveles de CD4.

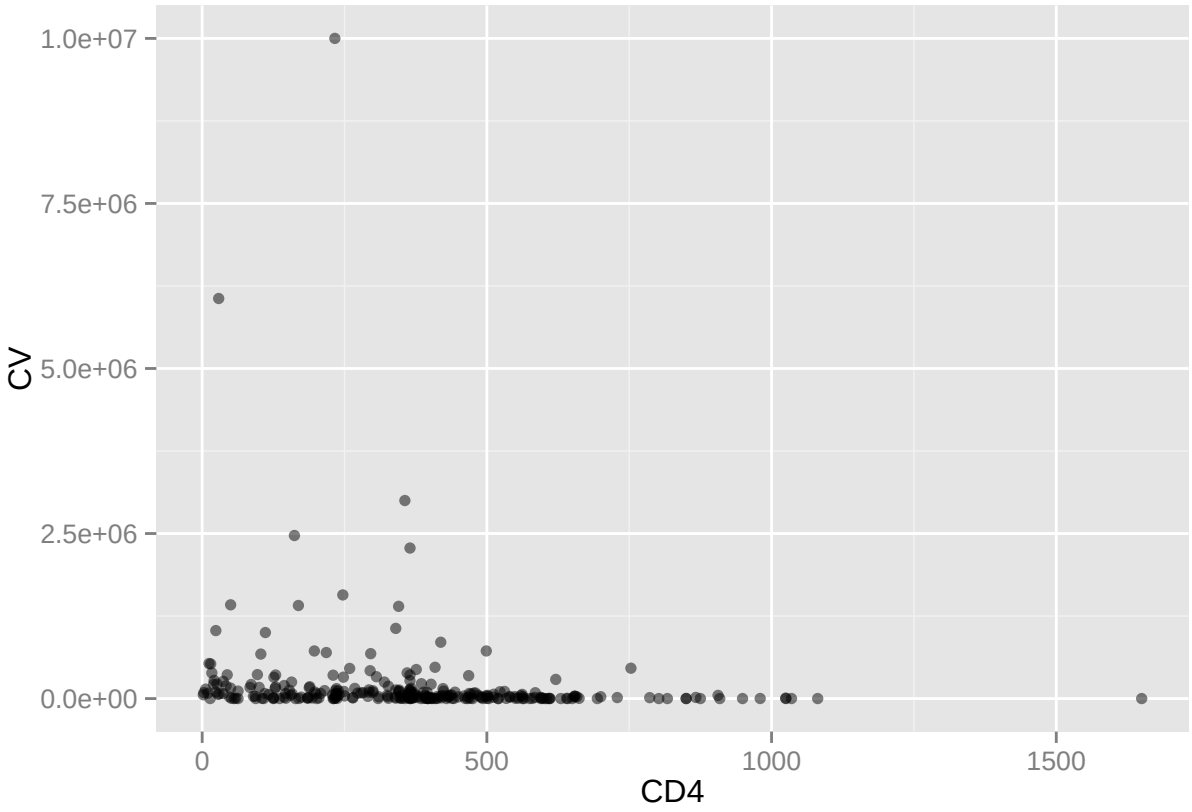
Fuentes: https://cran.r-project.org/web/packages/tigerstats/vignettes/densityplot.html

7. ¿Existe alguna relación entre la carga viral y los CD4?

```
library(ggplot2)

d <- ggplot(Tabla, aes(CD4, CV))
d + geom_point(alpha = 1/2)
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



De entrada, un mayor nivel de CD4 podría indicar un menor nivel de carga viral (CV), ya que no existen muestras en el cuadrante superior derecha a partir de 500 cel/ml de CD4. Con una muestra tan limitada es dificil extraer una conclusión pero podriamos decir que a niveles bajo-medios de CD4 (menor de 500 cel/ml) los niveles de CV, en algunos casos, son bastante altos. Como dijimos en el análisis descriptivo, ambas variables están muy debilmente relacionadas.

Fuentes:

http://docs.ggplot2.org/0.9.3.1/geom_point.html