

ejercicio1

Javier

11/9/2020

##Actividad 1

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

```
datos <- read.table("ejercicio1.txt",head=TRUE)
```

```
head(datos)
```

##	Genero	Edad	Estatura	Nota	Calificacion
## 1	Mujer	21	1.82	5	Aprobado
## 2	Hombre	19	1.83	6	Aprobado
## 3	Mujer	18	1.78	7	Notable
## 4	Mujer	20	1.79	5	Aprobado
## 5	Mujer	23	1.80	9	Sobresaliente
## 6	Hombre	22	1.90	7	Notable

#1. Mostrar las 10 últimas observaciones.

```
tail(datos,n=10)
```

##	Genero	Edad	Estatura	Nota	Calificacion
## 11	Mujer	22	1.73	5	Aprobado
## 12	Hombre	20	1.79	8	Notable
## 13	Hombre	22	1.80	7	Notable
## 14	Mujer	19	1.77	6	Aprobado
## 15	Mujer	19	1.69	3	Suspenso
## 16	Hombre	21	1.75	4	Suspenso
## 17	Mujer	20	1.66	5	Aprobado
## 18	Mujer	21	NA	6	Aprobado
## 19	Mujer	22	1.79	2	Suspenso
## 20	Mujer	23	1.80	8	Notable

#2 ¿Cuál es la estructura de los datos? Indicar dimensión y tipo de variables

```
str(datos) # todo
```

```
## 'data.frame':    20 obs. of  5 variables:
## $ Genero      : chr  "Mujer" "Hombre" "Mujer" "Mujer" ...
## $ Edad       : int   21 19 18 20 23 22 22 20 21 21 ...
## $ Estatura    : num   1.82 1.83 1.78 1.79 1.8 1.9 1.79 1.83 NA 1.65
## ...
## $ Nota        : int    5 6 7 5 9 7 8 3 9 2 ...
## $ Calificacion: chr   "Aprobado" "Aprobado" "Notable" "Aprobado" ...
```

```
summary(datos) # tipos de variable
```

```
##      Genero      Edad      Estatura      Nota
## Length:20      Min.    :18.0      Min.    :1.650      Min.    :2.00
## Class :character 1st Qu.:20.0      1st Qu.:1.755      1st Qu.:4.75
## Mode  :character Median :21.0      Median :1.790      Median :6.00
##                Mean  :20.8      Mean  :1.776      Mean  :5.75
##                3rd Qu.:22.0      3rd Qu.:1.800      3rd Qu.:7.25
##                Max.   :23.0      Max.   :1.900      Max.   :9.00
##                NA's    :2
## Calificacion
## Length:20
## Class :character
## Mode  :character
##
##
##
##

dim(datos) # dimensión

## [1] 20  5

class(datos) # clase de dato que tratamos

## [1] "data.frame"

#3 Calcular la media de las variables univariantes, esto es, de cada
columna (en
# aquellas que se pueda).

colMeans(datos[,2:4],na.rm = TRUE)

##      Edad  Estatura      Nota
## 20.800000  1.776111  5.750000

#4 Crear un nuevo data frame formado únicamente por los alumnos
suspensos. ¿Qué
# dimensión tiene? Guardarlo y exportarlo en formato .txt o .csv

suspensos <- datos[which(datos $Calificacion=="Suspenso"),]
dim(suspensos)

## [1] 5 5

write.table(suspensos,file="suspensos.txt")
write.csv(suspensos,file="suspensos.csv")
write.csv2(suspensos,file="suspensos.csv")

#5 Para el dataset completo obtener el valor más frecuente, o moda de
cada
# distribución, para las variables 'Edad', 'Estatura' y 'Nota'. ¿Hay
alguna
# bimodal?
```

```
# EDAD #
```

```
table(datos$Edad)
```

```
##
```

```
## 18 19 20 21 22 23
```

```
##  1  3  4  5  5  2
```

```
library(modeest)
```

```
mfv(datos$Edad)
```

```
## [1] 21 22
```

```
moda_Edad <- as.numeric(names(table(datos$Edad)==max(table(datos$Edad))))
```

```
moda_Edad
```

```
## [1] 18 19 20 21 22 23
```

```
# ESTATURA #
```

```
datos <- datos[!is.na(datos$Estatura),]
```

```
moda_Estatura <-
```

```
as.numeric(names(table(datos$Estatura)==max(table(datos$Estatura))))
```

```
moda_Estatura
```

```
## [1] 1.65 1.66 1.69 1.73 1.75 1.77 1.78 1.79 1.80 1.82 1.83 1.90
```

```
# NOTA #
```

```
moda_Nota <- as.numeric(names(table(datos$Nota)==max(table(datos$Nota))))
```

```
moda_Nota
```

```
## [1] 2 3 4 5 6 7 8 9
```

#6 Reordenar el data frame en función de la variable 'Nota', de menor a mayor, y

mostrar las seis primeras filas.

```
datos <- datos[with(datos, order(datos$Nota)),]
```

```
head(datos)
```

```
##      Genero Edad Estatura Nota Calificacion
```

```
## 10  Mujer   21     1.65    2      Suspenso
```

```
## 19  Mujer   22     1.79    2      Suspenso
```

```
##  8  Hombre  20     1.83    3      Suspenso
```

```
## 15  Mujer   19     1.69    3      Suspenso
```

```
## 16  Hombre  21     1.75    4      Suspenso
```

```
##  1  Mujer   21     1.82    5      Aprobado
```

#7 Realizar una tabla de frecuencias absolutas y otra de frecuencias relativas

*# para la variable 'Calificación'. Almacenar las tablas anteriores en dos
variables llamadas 'absolutas' y 'relativas'.*

ABSOLUTAS

```
absolutas <- table(datos$Calificacion)
absolutas
```

```
##
##      Aprobado      Notable Sobresaliente      Suspenso
##           6           6           1           5
```

RELATIVAS

```
relativas <- prop.table(absolutas)
relativas
```

```
##
##      Aprobado      Notable Sobresaliente      Suspenso
## 0.33333333 0.33333333 0.05555556 0.27777778
```

*#8 Representar la variable 'Calificación' mediante un diagrama de barras
y un*

*# diagrama de sectores. Incluir un título adecuado para cada gráfico y
colorear*

Las barras y los sectores de colores diferentes.

DIAGRAMA DE BARRAS

```
barplot(absolutas,col=c("red","blue","yellow","green"),main = "Diagrama  
de barras")
```

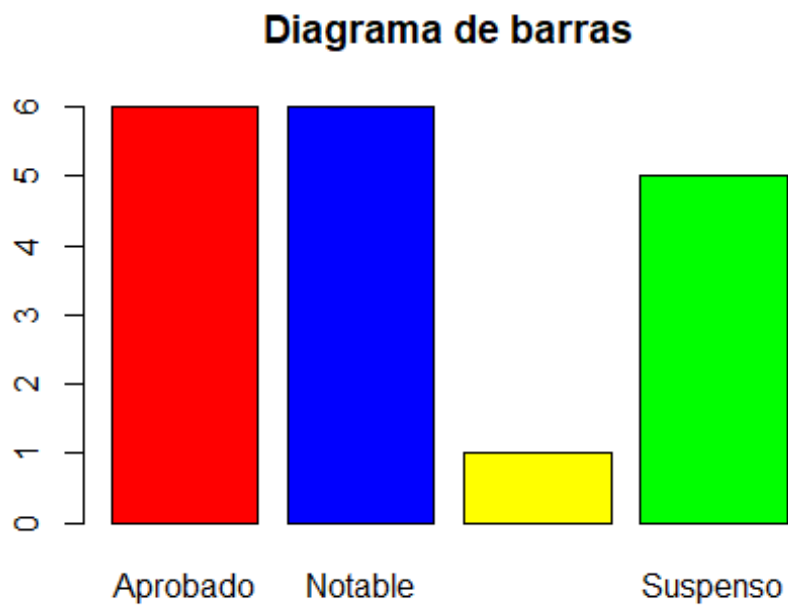


DIAGRAMA DE SECTORES

```
pie(absolutas,col=c("red","blue","yellow","green"),main="Diagrama de sectores")
```

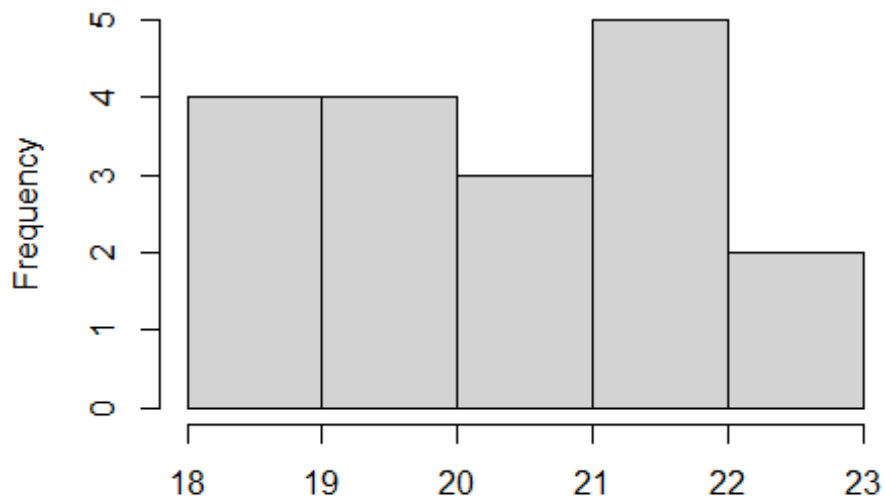
Diagrama de sectores



*#9 Para la variable 'Edad', realizar un histograma y un diagrama de caja
considerandola opción range = 1.5. Incluir un título apropiado para
cada
gráfico. ¿Existe algún valor atípico en esta variable? Reduce el valor
del
argumento range hasta 0.5. ¿Aparece algún atípico? ¿A qué observación
corresponde?*

```
hist(datos$Edad,main = "Histograma para la variable edad",xlab="")
```

Histograma para la variable edad



*#10 Realizar un resumen de la variable 'Nota' con el comando summary. Comprobar
que las medidas que proporciona summary coinciden con las medidas calculadas
de forma individual usando su función específica.*

```
summary(datos$Nota)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      2.000   4.250   5.500   5.556   7.000   9.000
```

MINIMO

```
min(datos$Nota)
```

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

MAXIMO

```
max(datos$Nota)
```

```
## [1] 9
```

CUANTIL

```
quantile(datos$Nota, probs=c(0.25, 0.75))
```

```
## 25% 75%
## 4.25 7.00

# MEDIA #

mean(datos$Nota)

## [1] 5.555556

# MEDIANA #

median(datos$Nota)

## [1] 5.5

#11 Calcular la estatura media de los estudiantes y proporcionar, al
menos, dos
# medidas que indiquen la dispersión de esta variable.

mean(datos$Estatura, na.rm = TRUE)

## [1] 1.776111

var(datos$Estatura, na.rm = TRUE)

## [1] 0.003836928

IQR(datos$Estatura, na.rm = TRUE)

## [1] 0.045

#12 ¿Qué variable es más homogénea: La 'Edad' o La 'Estatura'? Para
determinar
# la homogeneidad de una variable, esto es, la representatividad de su
media, se
# calcula el Coeficiente de Variación de Pearson definido como el
cociente entre
# la desviación típica y la media de la variable.
#
# 
$$CV(x) = \frac{sd(x)}{E(x)}$$

#
# EDAD #

CV_edad<-sd(datos$Edad)/mean(datos$Edad)
CV_edad

## [1] 0.07110963

# ESTATURA #

CV_estatura<-
```



```

sd(datos$Estatura,na.rm=TRUE)/mean(datos$Estatura,na.rm=TRUE)
CV_estatura

## [1] 0.0348756

#13 Obtener la asimetría y curtosis de las variables. ¿Puede asegurarse
# variables siguen una distribución normal? ¿Y la variable multivariante?

library(moments)

##
## Attaching package: 'moments'

## The following object is masked from 'package:modeest':
##
##      skewness

# EDAD #

asimetria_edad <- skewness(datos$Edad,na.rm=TRUE)
asimetria_edad

## [1] -0.1700558

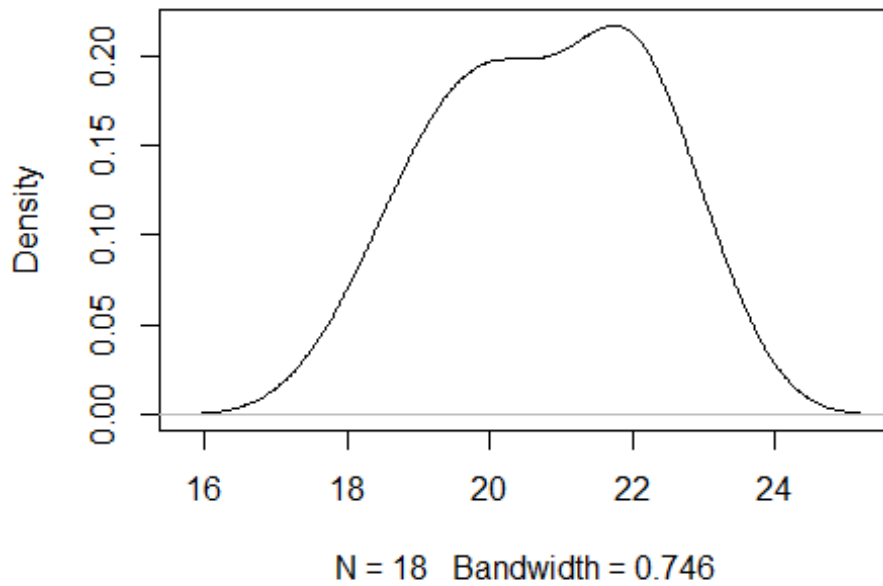
curtosis_edad <- kurtosis(datos$Edad,na.rm=TRUE)
curtosis_edad

## [1] 1.972283

plot(density(datos$Edad))

```

density.default(x = datos\$Edad)



ESTATURA

```
asimetria_estatura <- skewness(datos$Estatura, na.rm=TRUE)
asimetria_estatura
```

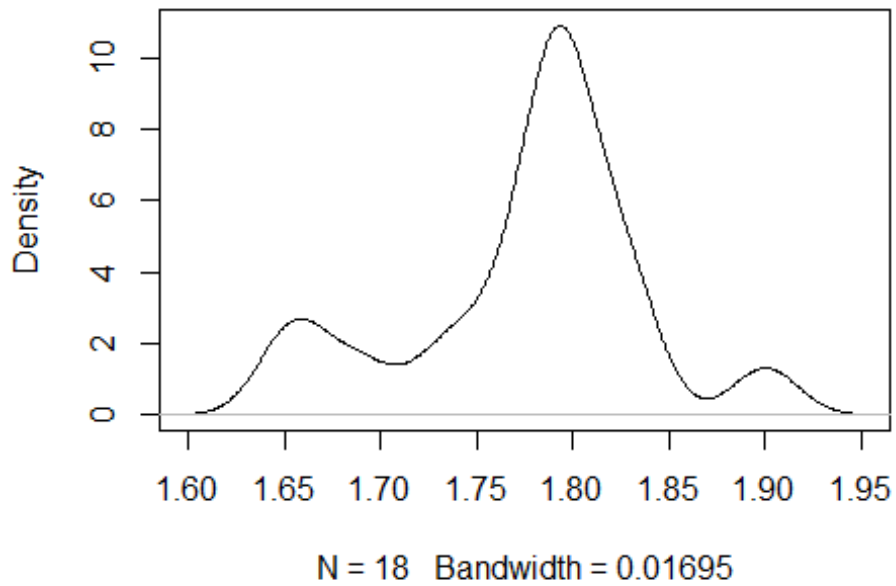
```
## [1] -0.5029718
```

```
curtosis_estatura <- kurtosis(datos$Estatura, na.rm=TRUE)
curtosis_estatura
```

```
## [1] 3.18122
```

```
plot(density(datos$Estatura, na.rm = TRUE))
```

density.default(x = datos\$Estatura, na.rm = TRUE



NOTA

```
asimetria_nota <- skewness(datos$Nota,na.rm=TRUE)
asimetria_nota
```

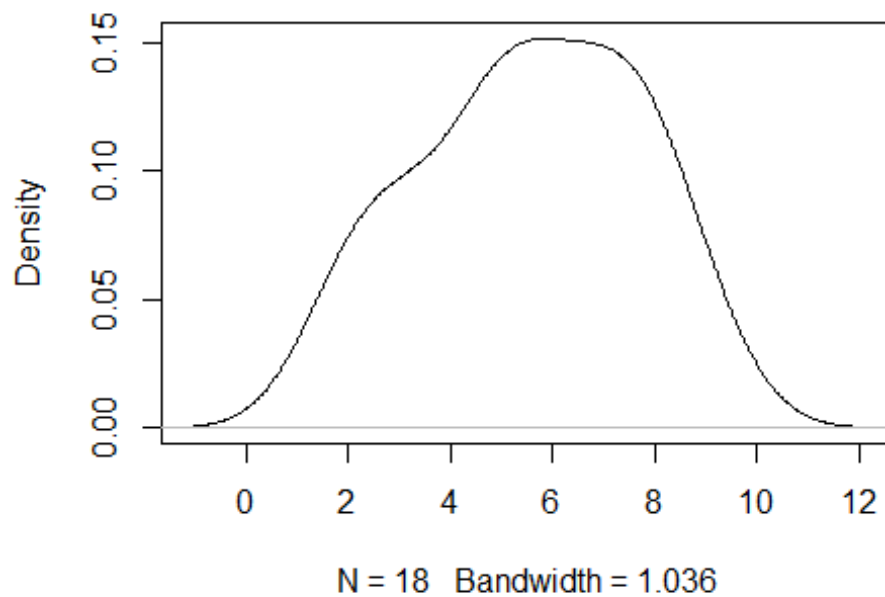
```
## [1] -0.2065658
```

```
curtosis_nota <- kurtosis(datos$Nota,na.rm=TRUE)
curtosis_nota
```

```
## [1] 1.966254
```

```
plot(density(datos$Nota, na.rm = TRUE))
```

```
density.default(x = datos$Nota, na.rm = TRUE)
```

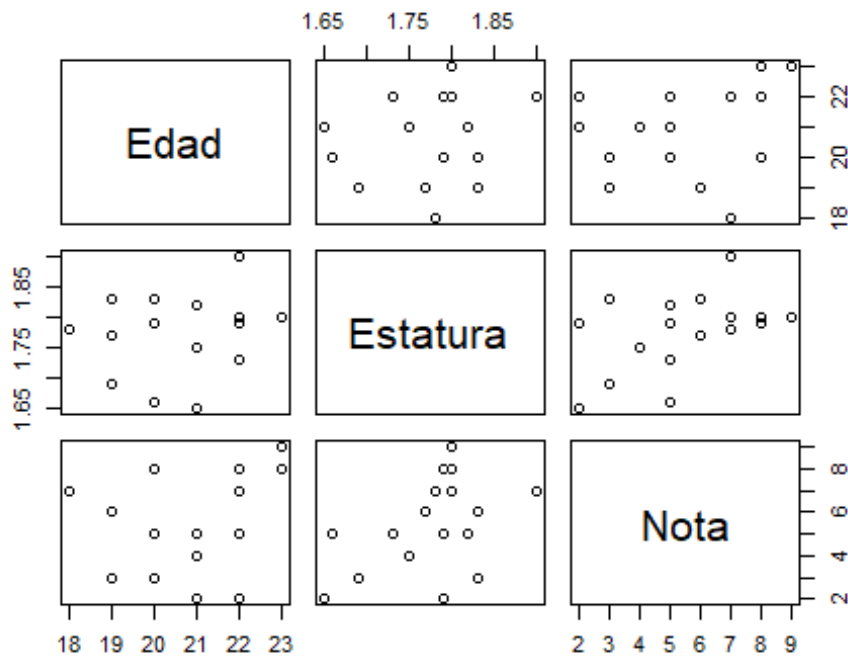


```
#14 ¿Existe alguna correlación entre La edad y La estatura? ¿Y entre el  
sexo y  
# La nota?
```

```
d1 <- na.omit(datos)  
cor(d1[,2:4])
```

```
##           Edad  Estatura    Nota  
## Edad      1.0000000 0.2085317 0.2450594  
## Estatura  0.2085317 1.0000000 0.4548530  
## Nota      0.2450594 0.4548530 1.0000000
```

```
pairs(d1[,2:4])
```



#15 Crea dos dataframes, uno formado sólo por mujeres cuya nota sea superior o igual a 5 y otro formado sólo por hombres con el mismo criterio.

MUJERES

```
aprobados_mujeres <- subset(datos, Genero=="Mujer" & Nota>=5)
aprobados_mujeres
```

```
##      Genero Edad  Estatura  Nota  Calificacion
## 1   Mujer   21    1.82    5    Aprobado
## 4   Mujer   20    1.79    5    Aprobado
## 11  Mujer   22    1.73    5    Aprobado
## 17  Mujer   20    1.66    5    Aprobado
## 14  Mujer   19    1.77    6    Aprobado
## 3   Mujer   18    1.78    7    Notable
## 7   Mujer   22    1.79    8    Notable
## 20  Mujer   23    1.80    8    Notable
## 5   Mujer   23    1.80    9  Sobresaliente
```

HOMBRES

```
aprobado_hombres <- subset(datos, Genero=="Hombre" & Nota>=5)
aprobado_hombres
```

```
##      Genero Edad  Estatura  Nota  Calificacion
## 2   Hombre   19    1.83    6    Aprobado
```

```
## 6 Hombre 22 1.90 7 Notable
## 13 Hombre 22 1.80 7 Notable
## 12 Hombre 20 1.79 8 Notable
```

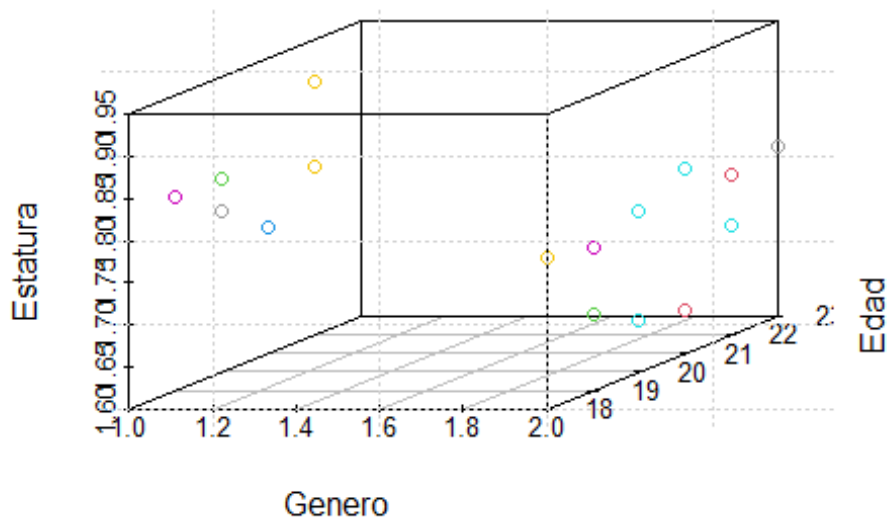
#16 Calcular la nota media por género empleando la función tapply().

```
tapply(datos$Nota, datos$Genero, mean)
```

```
## Hombre Mujer
## 5.833333 5.416667
```

#17 ¿Existe algún atípico multivariante? Representálos en 3D.

```
library(robustbase)
library(scatterplot3d)
scatterplot3d(datos)
d <- na.omit(datos[,2:4])
rownames(d) <- NULL
a <- covMcd(d)
at <- as.factor(a$mcd.wt)
colors <- c("red", "purple")
colors <- colors[as.numeric(at)]
grid()
```



```
scatterplot3d(d, pch=16, type="h", color=colors)
```

