Hoja 2 de problemas y prácticas con R

Estadística Computacional I. Grado en Estadística

Departamento de Estadística e Investigación Operativa. Universidad de Sevilla

Contents

1 Generar una muestra de calificaciones de 50 alumnos con el comando:

`` `sample(c("1S","2A","3N","4SB","5MH"),prob=c(0.3,0.35,0.2,0.1,0.05),50,rep=T)` ``.

```
set.seed(12345)
m1=sample(c("1S","2A","3N","4SB","5MH"),
          prob=c(0.3,0.35,0.2,0.1,0.05),50,rep=T)
m1
```

```
##  [1] "3N"  "4SB" "3N"  "4SB" "1S"  "2A"  "2A"  "1S"  "3N"  "5MH" "2A"  "2A"
## [13] "3N"  "2A"  "1S"  "1S"  "1S"  "1S"  "2A"  "5MH" "1S"  "2A"  "5MH" "3N"
## [25] "1S"  "1S"  "3N"  "1S"  "2A"  "1S"  "3N"  "2A"  "2A"  "3N"  "1S"  "1S"
## [37] "4SB" "4SB" "1S"  "2A"  "3N"  "1S"  "4SB" "3N"  "2A"  "2A"  "2A"  "2A"
## [49] "2A"  "1S"
```

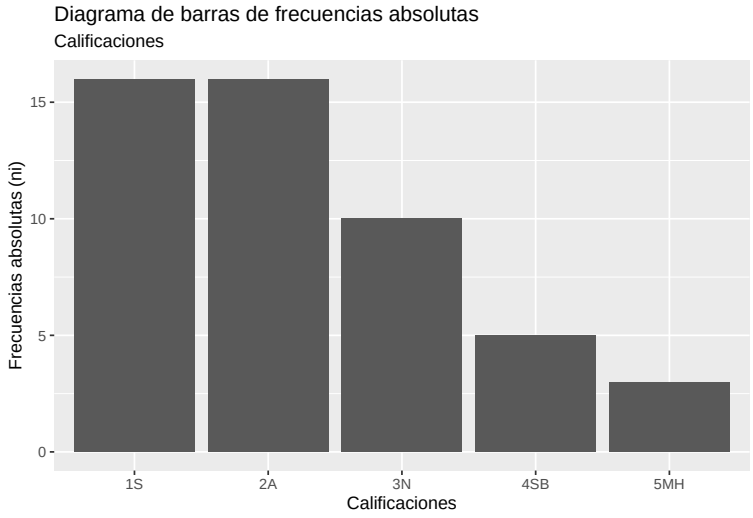
1.1 Resumir los valores obtenidos mediante una tabla de frecuencias incluyendo frecuencias absolutas, frecuencias relativas, porcentajes, y los acumulados de las tres características.

```
tablafre=tibble(valores=m1) %>%
  group_by(valores) %>%
  summarise(
    ni=n() # Frecuencias absolutas
    ) %>%
  mutate(
    fi=ni /length(m1), #Frec rel
    pi=fi*100, #Porcentajes
    Ni=cumsum(ni),
    Fi=cumsum(fi),#Ni/length(m1)
    Pi=cumsum(pi)
    )

tablafre %>%
  kable(booktabs=TRUE) %>%
  kable_styling(latex_options = "striped"  )
```

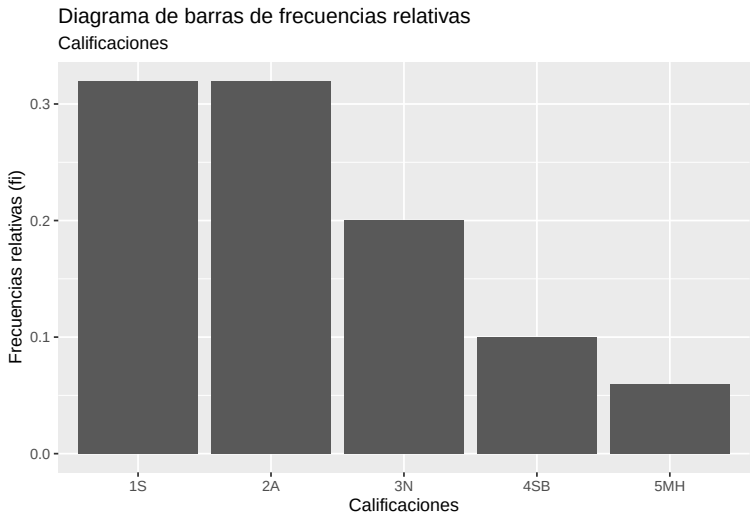
valores	ni	fi	pi	Ni	Fi	Pi
1S	16	0.32	32	16	0.32	32
2A	16	0.32	32	32	0.64	64
3N	10	0.20	20	42	0.84	84
4SB	5	0.10	10	47	0.94	94
5MH	3	0.06	6	50	1.00	100

1.2 Obtener representaciones gráficas adecuadas de las medidas anteriores.

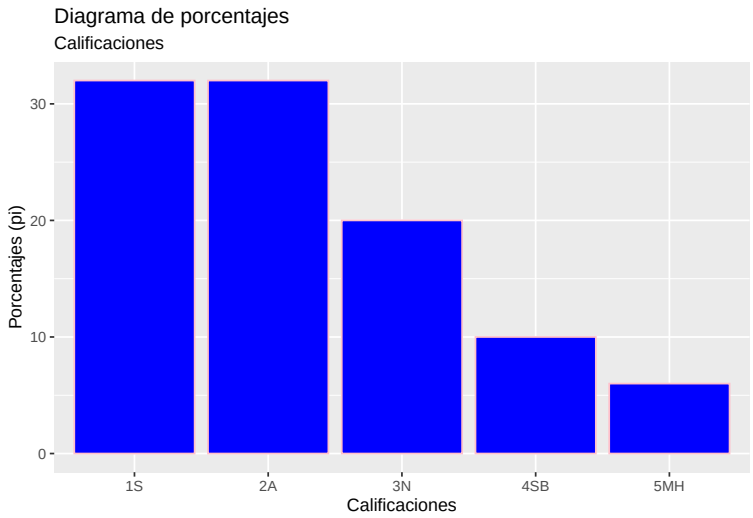
```
tablafre %>%
  ggplot(aes(x=valores , y = ni)) +
  geom_col()+
  labs(
    title = "Diagrama de barras de frecuencias absolutas",
    subtitle = "Calificaciones",
    y="Frecuencias absolutas (ni)",
    x="Calificaciones"
  )
```




```
tablafre %>%
  ggplot(aes(x=valores , y = fi)) +
  geom_col()+
  labs(
    title = "Diagrama de barras de frecuencias relativas",
    subtitle = "Calificaciones",
    y="Frecuencias relativas (fi)",
    x="Calificaciones"
  )
```




```
tablafre %>%
  ggplot(aes(x=valores , y = pi)) +
  geom_col( color = "pink", fill ="blue")+
  labs(
    title = "Diagrama de porcentajes",
    subtitle = "Calificaciones",
    y="Porcentajes (pi)",
    x="Calificaciones"
  )
```




```
tablafre %>%
  ggplot(aes(x=valores , y = Fi, group=1)) +
 # geom_col(color="blue") +
 # geom_line(color="pink")+ # Si no pongo group=1 no me hace la representación gráfica. Es para variabl
  geom_step(col="pink")+ #Variables discretas
  labs(
    title = "Polígono de Frecuencias rel acumuladas",
    subtitle = "Calificaciones",
    y="Frecuencias rel acumuladas (Fi)",
    x="Calificaciones"
  )
```

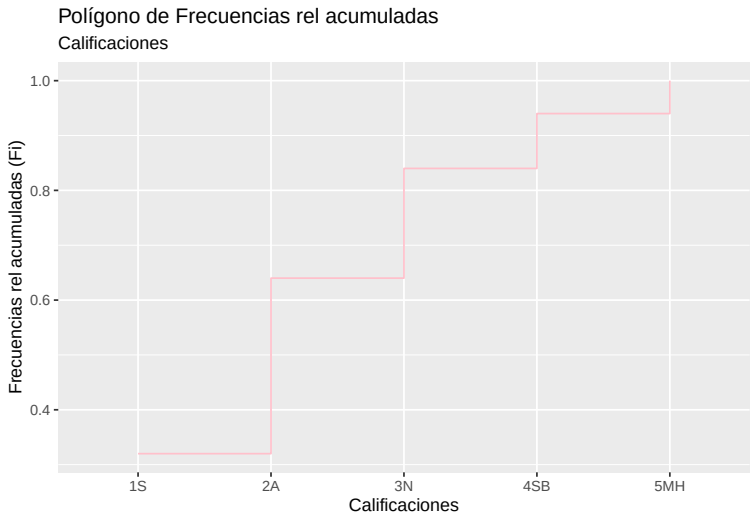


Diagrama de sectores con el sistema base:

```
tablafre$fi
```

```
## [1] 0.32 0.32 0.20 0.10 0.06
```

```
pie(tablafre$ni, labels = tablafre$valores)
```

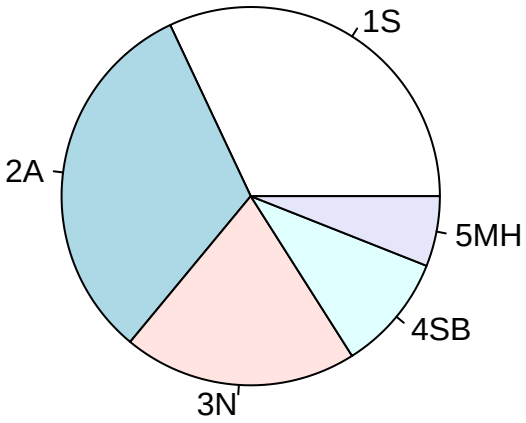
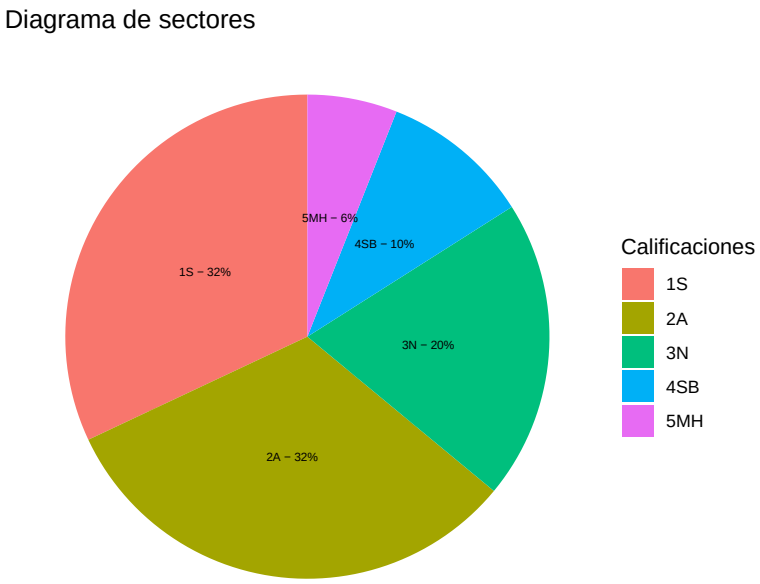


Diagrama de sectores con ggplot2:

```
tablafre %>%
  ggplot(aes(x="",y=pi, fill=factor(valores)))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y",start = 0)+
  theme_void() +
  geom_text(aes(label=paste0(valores," - ",          round(pi,2), "%")),
   position=position_stack(vjust=0.5), size=2 ) +
  labs(
    title = "Diagrama de sectores",
    fill= "Calificaciones"
  )
```



2 Fichero "Familia.txt"

El fichero "Familia.txt" contiene el peso (kgs) y la altura (cms) de los integrantes de una familia.

2.1 Leer en R este fichero y calcular la media y la cuasidesviación típica de cada variable.

```
datos2=read.table(file="Familia.txt",sep=" ")
head(datos2) %>%
  kable(booktabs=TRUE) %>%
kable_styling(latex_options = "striped")
```

	Altura	Peso
Sobrina	120	22
Hijo	172	52
Abuelo	163	71
Hija	158	51
Sobrino	153	51
Abuela	148	60

```
datos2 %>%
  summarise(
    MediaAltura=mean(Altura),
    MediaPeso = mean(Peso),
    SdAltura=sd(Altura),
    SdPeso=sd(Peso)
  ) %>%
  kable(booktabs=TRUE) %>%
kable_styling(latex_options = "striped")
```

MediaAltura	MediaPeso	SdAltura	SdPeso
156.6	54.1	14.93839	13.56835

Otra forma:

```
datos2 %>%
  summarise_each(
   c( sd, mean)) %>%
  kable(booktabs=TRUE) %>%
kable_styling(latex_options = "striped")
```

Altura_fn1	Peso_fn1	Altura_fn2	Peso_fn2
14.93839	13.56835	156.6	54.1

Otra forma:

```
datos2 %>%
  summarise_all(
    list(mean,sd)
  )  %>%
kable(booktabs=TRUE) %>%
  kable_styling(latex_options = "striped")
```

Altura_fn1	Peso_fn1	Altura_fn2	Peso_fn2
156.6	54.1	14.93839	13.56835

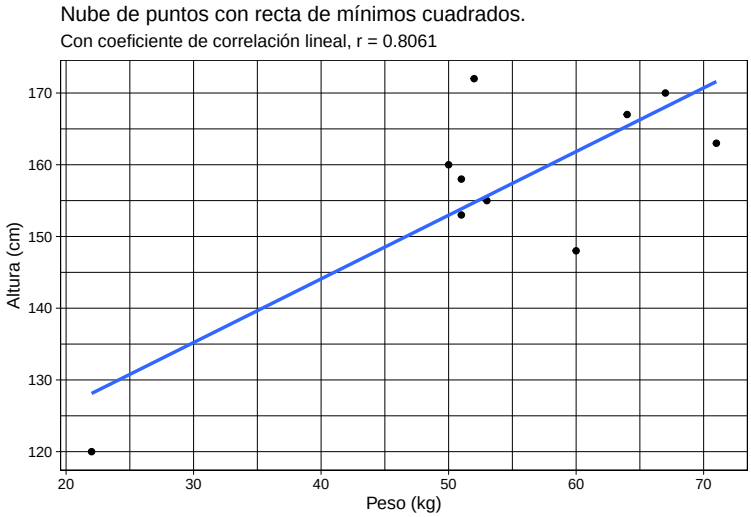
2.2 Nube de puntos y recta de mínimos cuadrados

Dibujar la nube de puntos (Peso, Altura) y superponer la recta de mínimos cuadrados. Calcular el coeficiente de correlación lineal entre ambas variables.

Coeficiente de correlación lineal

```
datos2 %>%
  summarise(
    Ccirlineal=cor(Peso,Altura)
    )-> ccl
```

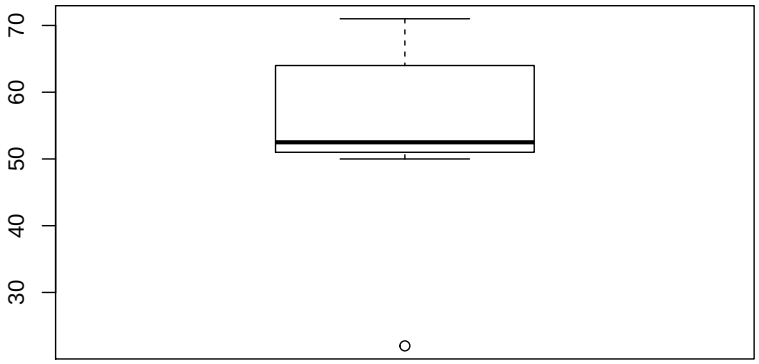
```
ggplot(data=datos2, aes(Peso,Altura)) +
  geom_point()+
  geom_smooth(method=lm, se=FALSE, formula = y~x)+  #Quito los IC
  labs(
    title = "Nube de puntos con recta de mínimos cuadrados.",
    subtitle = paste0("Con coeficiente de correlación lineal, r = " , round(ccl,4)),
    y="Altura (cm)",
    x="Peso (kg)" )+
  theme_linedraw()
```



2.3 Outliers

¿Qué observación es outlier para la variable peso? Repetir el apartado anterior sin esa persona.

```
res2=boxplot(datos2$Peso)
```




```
res2 # Mirlo las estadísticas del diagrama.
```

```
## $stats
##      [,1]
## [1,] 50.0
## [2,] 51.0
## [3,] 52.5
## [4,] 64.0
## [5,] 71.0
## attr(,"class")
##         1
## "integer"
##
## $n
## [1] 10
##
## $conf
##          [,1]
## [1,] 46.00468
## [2,] 58.99532
```

```
##
## $out
## [1] 22
##
## $group
## [1] 1
##
## $names
## [1] "1"
```

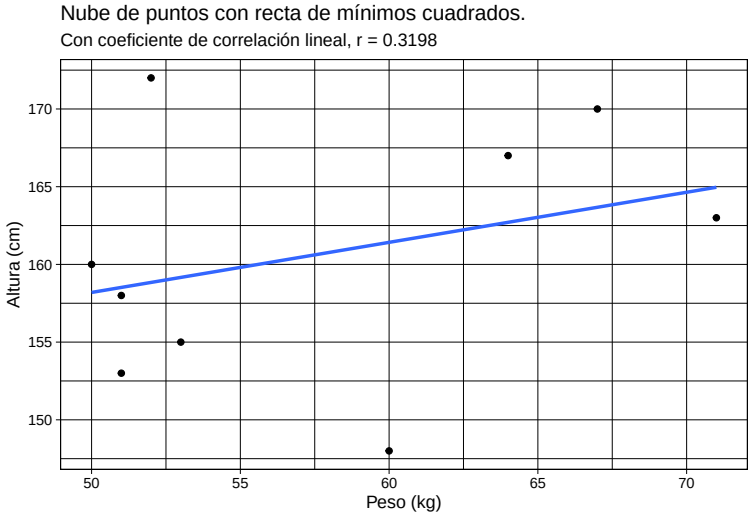
```
datos2 %>%
  arrange(Peso) %>%
  head(1)
```

```
##         Altura Peso
## Sobrina    120   22
```

```
datos2new=datos2[-1,]
```

```
datos2new %>%
  summarise(
    Ccirlineal=cor(Peso,Altura)
    )-> ccl
```

```
ggplot(data=datos2new, aes(Peso,Altura)) +
  geom_point()+
  geom_smooth(method=lm, se=FALSE, formula = y~x)+  #Quito los IC
  labs(
    title = "Nube de puntos con recta de mínimos cuadrados.",
    subtitle = paste0("Con coeficiente de correlación lineal, r = " , round(ccl,4)),
    y="Altura (cm)",
    x="Peso (kg)" )+
  theme_linedraw()
```



2.4 IMC

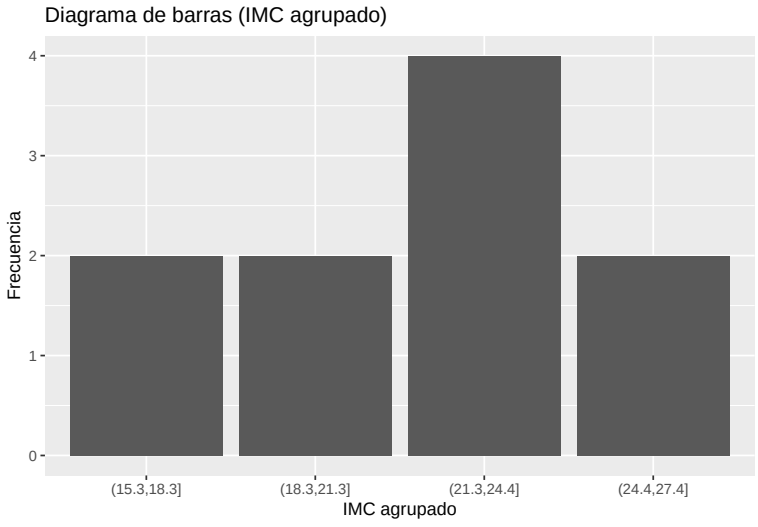
Calcular el Índice de Masa Corporal (IMC), definido como el cociente entre el peso y el cuadrado de la altura (en metros). Representarlo con un gráfico de barras.

```r
dat2imc= datos2 %>%
  mutate(IMC=Peso/((Altura/100)^2))

dat2imc %>%
  kable(booktabs=TRUE) %>%
kable_styling(latex_options = "striped")
```

	Altura	Peso	IMC
Sobrina	120	22	15.27778
Hijo	172	52	17.57707
Abuelo	163	71	26.72287
Hija	158	51	20.42942
Sobrino	153	51	21.78649
Abuela	148	60	27.39226
Tía	160	50	19.53125
Tío	170	67	23.18339
Madre	155	53	22.06035
Padre	167	64	22.94812

Vamos a definir cuatro intervalos para representar mis datos, empleamos el método del rango intercuartílico.

```r
dat2imc %>%
  mutate(
    IMCargu=cut(IMC,breaks=4)
  ) %>% ggplot(aes(x=IMCargu))+
  geom_bar()+
  labs(

  x="IMC agrupado",
  y="Frecuencia",
  title="Diagrama de barras (IMC agrupado)"
  )
```



2.5 Ordenar

Ordenar los familiares de mayor a menor IMC.

```
dat2imc %>%
  arrange(desc(IMC  ) ) %>%
  kable(booktabs=TRUE) %>%
  kable_styling(latex_options = "striped")
```

	Altura	Peso	IMC
Abuela	148	60	27.39226
Abuelo	163	71	26.72287
Tío	170	67	23.18339
Padre	167	64	22.94812
Madre	155	53	22.06035
Sobrino	153	51	21.78649
Hija	158	51	20.42942
Tía	160	50	19.53125
Hijo	172	52	17.57707
Sobrina	120	22	15.27778

3 Librería *ISwR*

Acceder al fichero alkfos de la librería ISwR:

```
library(ISwR)
alkfos %>%
  kable(booktabs=TRUE,longtable=T,caption="Fichero alkfos") %>%
  kable_styling(latex_options = c("striped","repeat_header"))
```

Table 1: Fichero alkfos

grp	c0	c3	c6	c9	c12	c18	c24
1	142	140	159	162	152	175	148
1	120	126	120	146	134	119	116
1	175	161	168	164	213	194	221
1	234	203	174	197	289	174	189
1	94	107	146	124	128	98	114
1	128	97	113	203	NA	NA	NA
1	202	189	208	203	209	200	218
1	190	277	270	171	141	192	190
1	104	117	135	122	112	133	123
1	112	95	114	122	118	119	138
1	160	169	178	208	220	215	232
1	214	211	215	240	227	288	260
1	113	138	112	114	109	106	111
1	237	245	219	213	215	225	228
1	205	213	248	222	225	207	172
1	202	231	236	185	204	226	147
1	137	128	136	146	152	132	150
1	175	163	167	144	168	NA	NA

Table 1: Fichero alkfos *(continued)*

grp	c0	c3	c6	c9	c12	c18	c24
1	174	151	150	133	134	149	146
1	81	81	83	74	82	84	108
1	113	131	298	124	126	140	129
1	104	114	124	102	94	122	125
1	178	172	159	155	157	153	164
2	150	122	103	109	103	87	109
2	173	127	117	124	143	123	144
2	191	174	165	160	177	184	NA
2	191	159	157	161	150	187	215
2	230	150	144	153	125	124	152
2	145	134	167	141	112	212	194
2	128	92	89	78	83	78	80
2	102	86	80	76	82	79	68
2	180	124	116	117	124	NA	NA
2	153	96	97	96	93	156	110
2	115	79	79	79	73	69	72
2	150	113	124	102	100	109	101
2	182	147	156	79	135	NA	162
2	175	146	157	140	143	158	162
2	146	86	81	80	87	89	95
2	92	80	95	95	86	119	NA
2	228	177	185	181	190	182	192
2	178	119	107	NA	102	110	94
2	213	185	152	142	158	178	194
2	161	107	104	107	NA	118	129

3.1 Ver los primeros casos y los últimos.

```
alkfos[c(1,2,3,41,42,43),] %>%
  kable(booktabs=TRUE) %>%
  kable_styling(latex_options = "striped", stripe_index = c(1,2,5:6))
```

	grp	c0	c3	c6	c9	c12	c18	c24
1	1	142	140	159	162	152	175	148
2	1	120	126	120	146	134	119	116
3	1	175	161	168	164	213	194	221
41	2	178	119	107	NA	102	110	94
42	2	213	185	152	142	158	178	194
43	2	161	107	104	107	NA	118	129

3.2 Reformatear los datos a la estructura grp time c id.

3.3 Ordenar el nuevo formato por grp, id y time, y mostrar las variables en el orden (grp, id, time, c).

4 Fichero *"dietas.dat"*

Leer el fichero "dietas.dat", donde se distinguen 4 dietas alimenticias, y se mide el peso durante 6 meses.Cada variable toma valores de 1 a 6, a mayor valor más lejos del peso ideal. Representar gráficamente la evolución de los pesos medios según la dieta.

En este problema se trabajará con el conjunto de datos Boston de la librería MASS.

4.1 Acceder a los datos, en particular, averiguar qué información contiene y cuál es la dimensionalidad de los datos.

4.2 Ordenar las variables según el valor absoluto de su coeficiente de correlación lineal con medv (variable a predecir en este conjunto de datos).

4.3 ¿Destaca algún distrito por su tasa de criminalidad? Similarmente, por los impuestos sobre la propiedad o por la ratio alumnos-profesor.

4.4 ¿Cuántos distritos son limítrofes con el río? Calcular las medias de crim y medv según chas.

4.5 Analizar la relación lineal entre lstat y medv.

5 Comprobar empíricamente el Teorema de Fisher a partir de 5000 muestras de tamaño 10 de una ley N(0,1):

5.1 Analizar la relación lineal entre las medias y las cuasivarianzas.

5.2 Estudiar gráficamente si los cocientes (n-1)*cuasivar/(sigma^2) siguen una ley chi-cuadrado.

6 Comprobar mediante una simulación el ajuste de las distribuciones chi-cuadrado y la distribución F-Snedecor a partir de las cuasivarianzas muestrales para 10000 pares de muestras independientes. En cada par, la primera muestra será de tamaño 10 de la ley N(0,1), y la segunda muestra de tamaño 8 de la ley N(10,3).

7 Fichero *"salarios.txt"*

El fichero "salarios.txt" contiene datos sobre el salario (variable wage) y otras características para 3000 trabajadores.

7.1 Leer en R los datos.

7.2 Representar gráficamente los salarios según las variables age, year y education, y superponer estimaciones de la media del salario según cada variable.

7.3 Dibujar la evolución anual del salario medio según el nivel educativo.

7.4 Calcular los porcentajes de variación interanual del salario medio según nivel educativo.

7.5 Ordenar el fichero de datos según año (creciente) y edad (decreciente).

8 Librería *MASS*

Acceder al data frame painters de la librería MASS.

8.1 Interpretar y resumir la información contenida en este fichero de datos.

8.2 Seleccionar las escuelas del renacimiento y Veneciana para los siguientes apartados.

8.3 Generar en una sola pantalla los diagramas de caja y bigotes según la escuela.

8.4 Construir nubes de puntos en las que se distinga la escuela.

8.5 Comparar mediante gráficos de barras las medias de ambas escuelas.