

Estadística Descriptiva

Cotos-Yáñez, T.R., Pérez-González, A.

2020-10-20

Contents

Prólogo	5
1 Introducción. Elementos notables de Estadística Descriptiva	7
1.1 Introducción	7
1.2 Conceptos Generales.	8
2 Estadística Descriptiva Unidimensional	11
2.1 Tablas de distribución de frecuencias:	11
2.2 Representaciones Gráficas	15
2.3 Principales características numéricas	32
2.4 Medidas de posición de tendencia central	32
2.5 Medidas de posición no centrales: los Cuantiles	39
2.6 Medidas de Dispersión	41
2.7 Medidas de Dispersión relativas	44
2.8 Medidas de Forma	46
3 Ejercicios resueltos de Estadística descriptiva unidimensional con R y R-Commander	51
4 Ejercicios propuestos de Estadística descriptiva unidimensional para realizar con R-Commander	65
5 Estadística Descriptiva Bidimensional	73
5.1 Introducción	73
5.2 Tablas de frecuencias de doble entrada	74
5.3 Distribuciones marginales	75
5.4 Distribuciones condicionadas	76
5.5 Momentos de una distribución bidimensional numérica	79
5.6 Independencia Estadística	82
6 Ejercicios resueltos con R de Análisis Descriptivo Bidimensional	83

Prólogo

Esta documentación contiene tanto la parte Teórica de la materia como su aplicación práctica usando el lenguaje R a través del plugin R-Commander y el plugin RcomanderPlugin.InferenciaEstadistica.

Está escrito en R-Markdown empleando el paquete `bookdown` y está disponible en el repositorio Github: ¿?¿?.

Este obra está bajo licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional.

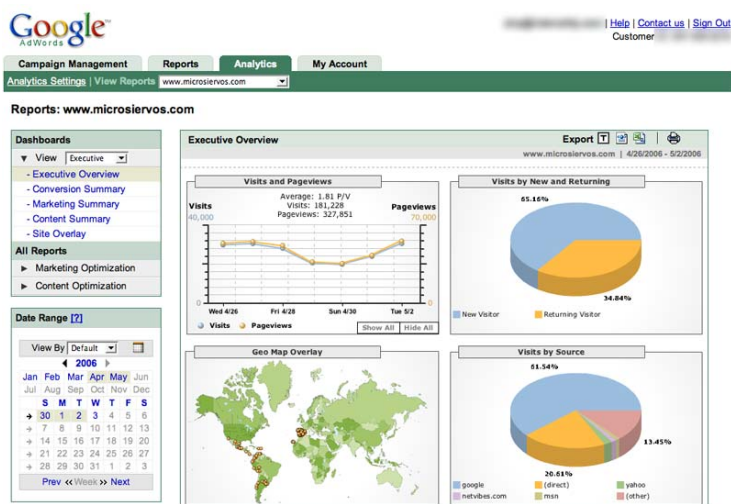
Chapter 1

Introducción. Elementos notables de Estadística Descriptiva

1.1 Introducción

Un primer paso en cualquier análisis de un problema o experimento es la observación de la información disponible. Esta información debe ser tratada inicialmente como un todo y es conveniente establecer mecanismos que permitan, no solo describirla completamente, sino también resumirla. Dicho de otra forma, dar características que permitan describir la información que se dispone. La estadística descriptiva es la parte de la estadística que se encarga de ordenar, analizar, representar y resumir un conjunto de datos, tratándolos como un todo.

Un claro ejemplo pueden ser los análisis descriptivos que se hacen de las visitas a portales web (buscar en google *estadística web* y pinchar por ejemplo en *Imágenes de estadística web*).



1.2 Conceptos Generales.

- **Población.** Cualquier conjunto de personas, objetos, ideas o acontecimientos.
- **Muestra.** Subconjunto de individuos pertenecientes a una población determinada. Es deseable que sea un subconjunto *representativo* de la población.
- **Variables:** En los elementos de una población se pueden observar uno o varios caracteres:
 - *Variable cualitativa o atributo:* Es un carácter de una población que no es susceptible de ser medido numéricamente. Se clasifican en nominales y ordinales. La v. cualitativas ordinales se pueden ordenar.

Ejemplos de variables cualitativas nominales son: el color de ojos, la raza, el sexo, la inclinación política, . . . , ejemplos de v. ordinales son: el nivel de estudios, la escala de grises (de blanco a negro), . . .

- *Variable cuantitativa:* Es cualquier carácter de una población susceptible de tomar valores numéricos. Se clasifican en discretas y continuas. Mientras que las discretas toman únicamente un número finito o infinito numerable de valores, las continuas pueden tomar todos los valores de un determinado intervalo.

Ejemplos de variables discretas: número de trabajos en cola en un servidor, número de hijos, número de caras obtenidas al lanzar dos monedas, . . .

Ejemplos de variables continuas: salario, peso, estatura, . . .

Ejercicio 1.1. Identifica que tipo de variables son:

- tipo de procesador,
- velocidad de un procesador (MHz),
- modelo de impresora,
- número de páginas impresas por minuto,
- variable representando un byte,
- la altura y el peso de una persona,
- la parte entera de la medición de la altura de una persona,
- ¿la variable número de núcleos de un procesador es ordinal o cuantitativa discreta?

Describe situaciones donde la variable *Edad* pueda ser considerada como v.a. continua, discreta e incluso como v.a. cualitativa ordinal.

Chapter 2

Estadística Descriptiva Unidimensional

Sea X la variable de interés, y sean x_1, x_2, \dots, x_k los valores distintos que toma dicha variable medidos en una muestra de tamaño N .

- **Frecuencia absoluta:** número de veces que se repite cada valor de la variable.
- **Frecuencia relativa:** frecuencia absoluta dividida por el número total de datos, N .
- **Frecuencia absoluta acumulada** de un valor de la variable (ordenados de menor a mayor) es el número de datos menores o iguales que dicho valor (no calculable para v. cualitativas nominales).
- **Frecuencia relativa acumulada** es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos.

Fr. Absoluta de x_i n_i	Fr. Relativa de x_i f_i
Fr. Abs. Acumulada de x_i $N_i = n_1 + \dots + n_i = \sum_{k=1}^i n_k$	Fr. Rel. Acumulada de x_i $F_i = f_1 + \dots + f_i = \sum_{k=1}^i f_k = \frac{N_i}{N}$

2.1 Tablas de distribución de frecuencias:

- Variables cualitativas y cuantitativas discretas: Datos sin agrupar.

Valores variable	Fr. abs.	Fr. rel.	Fr. abs. ac.	Fr. rel. ac.
x_i	n_i	$f_i = \frac{n_i}{N}$	$N_i = \sum_{j=1}^i n_j$	$F_i = \frac{N_i}{N}$
x_1	n_1	f_1	N_1	F_1
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	n_k	f_k	$N_k = N$	$F_k = 1$
	$N = \sum_{i=1}^k n_i$	1		

Las frecuencias acumuladas sólo tienen sentido cuando hay una ordenación de los datos, es decir, no tienen sentido cuando hay variables cualitativas nominales.

- **Variables cuantitativas continuas:** Datos agrupados. Si el tamaño N de la muestra es grande, suelen agruparse los datos en casillas o intervalos de clase de la forma $[L_{i-1}, L_i)$ donde $i = 1, 2, \dots, k$, sustituyendo cada valor de la variable por el punto medio del intervalo $x_i = \frac{L_{i-1} + L_i}{2}$ que llamaremos **marca de clase**.

Int. clase	M. clase	Fr. abs.	Fr. rel.	Fr. abs. ac.	Fr. rel. ac.	Amp.	Densidad
$[L_{i-1}, L_i)$	x_i	n_i	f_i	N_i	F_i	a_i	d_i
$[L_0, L_1)$	x_1	n_1	f_1	N_1	F_1	a_1	d_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$[L_{k-1}, L_k]$	x_k	n_k	f_k	$N_k = N$	$F_k = 1$	a_k	d_k
		N	1				

Se define **amplitud del intervalo** de clase como $a_i = L_i - L_{i-1}$. Los intervalos no tienen por qué tener la misma amplitud.

Se define **densidad de frecuencia** relativa como la frecuencia relativa por unidad de la variable, $d_i = \frac{f_i}{a_i}$.

Ejemplo 2.1. El conjunto de datos *iris* del paquete *datasets* (*Estadística Básica/Datos/Conjunto de datos en paquetes/Leer conjunto de datos en paquete adjunto...*) contiene mediciones en centímetros de diferentes variables de 150 flores (ver ayuda del conjunto de datos).

Responder a las siguientes cuestiones:

- Clasifica estadísticamente las variables del conjunto de datos *iris*.
 - *iris\$Sepal.Length*, *iris\$Sepal.Width*, *iris\$Petal.Length*, *iris\$Petal.Width* son v. continuas.
 - *iris\$Species* es una v. cualitativa nominal.

- Calcula a distribución de frecuencias completa da variable *Species*.
- Obtén la distribución de frecuencias agrupada de la variable *Sepal.Length* en 10 grupos de longitud 0.5.

Solución:

```
> ### Menú en: Estadística Básica/Estadística Descriptiva/Distribuciones de frecuencia
> ### variables cualitativas
> calcular_frecuencia(df.nominal=iris["Species"], ordenado.frec=TRUE, df.ordinal=NULL,
+   cuantil.p=0.5, iprint = TRUE)
```

Variables nominales:

Variable: Species

	ni	fi
setosa	50	0.333
versicolor	50	0.333
virginica	50	0.333

N= 150

```
> ### Primero: Recodificamos la variable usando el menú de segmentar
> ### Estadística Básica/Datos/Segmentar variable numérica
> iris$Sepal.Length.agrup2 <- as.ordered(with(iris, binVariable(Sepal.Length, bins=10,
+   method='intervals', labels=NULL)))
>
> ### Segundo: La v. generada es una v.a continua agrupada, para obtener la distribución
> ### de frecuencias completa, podemos usar el menú de distribución de frecuencias de
> ### v. cualitativas ordinales. Informáticamente el menú anterior clasificó la v.
> ### resultante como v. tipo factor ordenado.
>
> ### Menú: Distribución de frecuencias de v. cualitativas ordinales,
> ### Estadística Básica/Estadística Descriptiva/Distribuciones de frecuencia variables
> ### cualitativas...
> calcular_frecuencia(df.nominal=NULL, ordenado.frec=FALSE, df.ordinal=
+   iris["Sepal.Length.agrup2"], cuantil.p=0.5, iprint = TRUE)
```

Variables ordinales:

Variable: Sepal.Length.agrup2

	ni	fi	Ni	Fi
(4.3,4.66]	9	0.0600	9	0.060

```
(4.66,5.02] 23 0.1533 32 0.213
(5.02,5.38] 14 0.0933 46 0.307
(5.38,5.74] 27 0.1800 73 0.487
(5.74,6.1] 22 0.1467 95 0.633
(6.1,6.46] 20 0.1333 115 0.767
(6.46,6.82] 18 0.1200 133 0.887
(6.82,7.18] 6 0.0400 139 0.927
(7.18,7.54] 5 0.0333 144 0.960
(7.54,7.9] 6 0.0400 150 1.000
N= 150
```

Cuantil: 0.5

```
Variable      Fi
Sepal.Length.agrup2 (5.74,6.1] 0.6333333

> ### La densidad de frecuencia se obtendría dividiendo la columna fi por la amplitud.
> ### En este caso, amplitud constante ai=0.5. Las conclusiones serían idénticas
> ### observando la columna hi e fi.
>
>
> ### Alternativamente se podría usar el menú de Recodificar, aunque los resultados
> ### no tienen porque coincidir.
> ### Primero: Recodificamos en Estadística Básica/Datos/Modificar el conjunto de datos
> ### activo/Recodificar ...
> iris <- within(iris, {
+   Sepal.Length.agrup2 <- Recode(Sepal.Length,
+   '4:4.5 = "[4,4.5)"; 4.5:5 = "[4.5,5)"; 5:5.5 = "[5,5.5)"; 5.5:6 = "[5.5,6)";
+   6:6.5 = "[6,6.5)"; 6.5:7 = "[6.5,7)"; 7:7.5 = "[7,7.5)"; 7.5:8 = "[7.5,8)";
+   as.factor=TRUE)
+ })
> ### Segundo: La v. generada es una v.a continua agrupada, para obtener la
> ### distribución de frecuencias completa, podemos usar el menú de distribución
> ### de frecuencias de v. cualitativas ordinales. Informáticamente el menú
> ### anterior clasificó la v. resultante como v. tipo factor (es decir, v.
> ### cualitativa nominal). Debemos modificar para que se R Commander la
> ### considere como factor ordenado.
> ### Menú: Estadística Básica/Datos/Modificar el conjunto de datos activo
> ### /Cambiar tipo de variables
> iris <- within(iris, {
+   Sepal.Length.agrup2 <- as.ordered(Sepal.Length.agrup2)
+ })
>
> ### Tercero: Distribución de frecuencias de v. cualitativas ordinales,
> ### Estadística Básica/Estadística Descriptiva/Distribuciones de frecuencia
> ### variables cualitativas...
> calcular_frecuencia(df.nominal=NULL, ordenado=frec=FALSE, df.ordinal=
```

```
+ iris["Sepal.Length.agrup2"], cuantil.p=0.5, iprint = TRUE)
```

```
-----
```

Variables ordinales:

Variable: Sepal.Length.agrup2

	ni	fi	Ni	Fi
[4,4.5)	5	0.0333	5	0.0333
[4.5,5)	27	0.1800	32	0.2133
[5,5.5)	27	0.1800	59	0.3933
[5.5,6)	30	0.2000	89	0.5933
[6,6.5)	31	0.2067	120	0.8000
[6.5,7)	18	0.1200	138	0.9200
[7,7.5)	6	0.0400	144	0.9600
[7.5,8]	6	0.0400	150	1.0000

N= 150

Cuantil: 0.5

	Variable	Fi
Sepal.Length.agrup2	[5.5,6)	0.5933333

Ejercicio 2.1. La distribución del tiempo de conexión a la red internet en un edificio viene agrupado en intervalos por la siguiente tabla:

Tiempo (minutos)	Marca Clase	Nº de usuarios
(0 – 20]	10	10
(20 – 28]	24	52
(28 – 32]	30	30
(32 – 48]	40	8

Obtén la distribución de frecuencias completa de la variable agrupada.

Solución. La variable estadística a considerar se puede definir como $X = \text{tiempo de conexión}$, la tabla completa quedaría:

Tiempo (minutos)	Marca Clase	Nº de usuarios	N_i	f_i	F_i	d_i
(0 – 20]	10	10	10	10/100	10/100	$\frac{0.1}{20} = 0.005$
(20 – 28]	24	52	62	52/100	62/100	$\frac{0.52}{8} = 0.065$
(28 – 32]	30	30	92	30/100	92/100	$\frac{0.3}{4} = 0.075$
(32 – 48]	40	8	100	8/100	100/100	$\frac{0.08}{16} = 0.005$

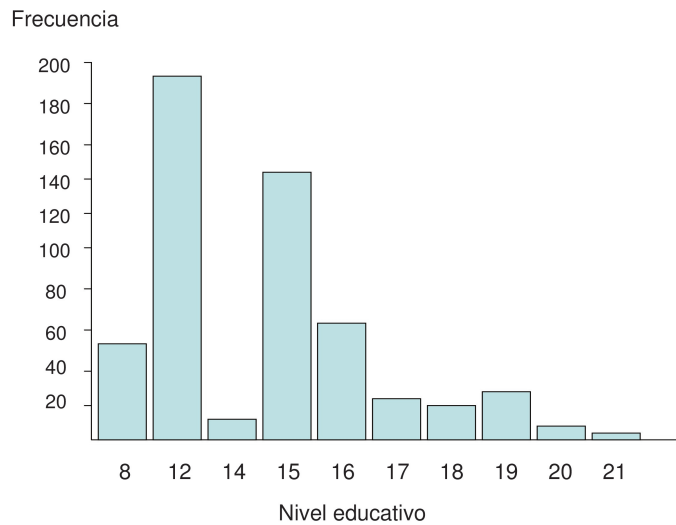
2.2 Representaciones Gráficas

Aunque la tabla estadística encierra toda la información posible disponible, a veces es necesario traducirla o acompañarla de una representación gráfica de

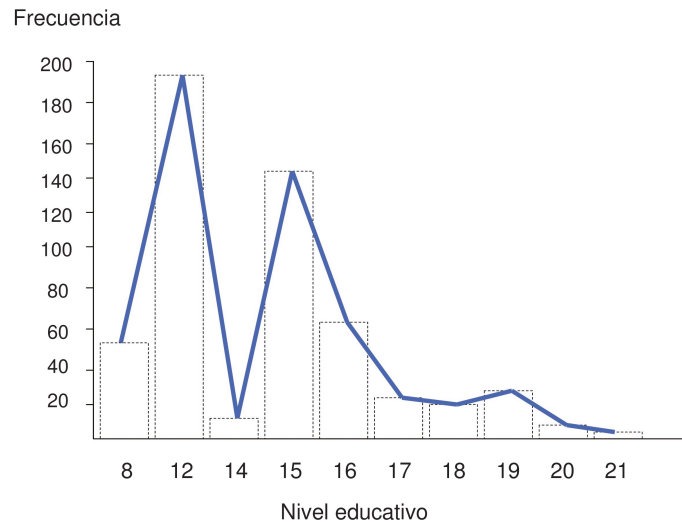
modo que la referencia visual sirva de punto de partida para el análisis estadístico. Los tipos de gráficos más utilizados son:

1. Para fenómenos cualitativos: cartograma, pictograma, nube de palabras, diagrama de sectores y diagrama de rectángulos.
2. Para fenómenos cuantitativos:
 - Distribuciones no agrupadas: diagrama de barras y polígono de frecuencias.
 - Distribuciones agrupadas: histograma de frecuencias, polígono de frecuencias y polígono acumulativo.

Diagrama de barras: En el eje de abscisas, de unos ejes coordenados rectangulares, se señalan los valores de la variable, construyendo sobre estos unos segmentos verticales de longitud igual (o proporcional) a cada una de las frecuencias.



Polígono de frecuencias: Se obtiene uniendo mediante segmentos los puntos $(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)$ ó $(x_1, f_1), (x_2, f_2), \dots, (x_k, f_k)$.



- En distribuciones agrupadas en intervalos.

Histogramas: Se marcan en el eje X los extremos de los intervalos de clase, y sobre cada intervalo se construye un rectángulo cuya base coincide con el intervalo y la altura es igual (o proporcional) a la frecuencia de dicho intervalo.

Generalmente $d_i = \frac{n_i}{a_i}$ ó $d_i = \frac{f_i}{a_i}$.

Pasos en la construcción de un histograma:

- Determinar el rango de los datos. Rango es igual al dato mayor menos el dato menor.
- Obtener el número de clases:

Regla de la raíz : $k = \text{int}(\sqrt{N})$

Regla de Sturges : $h = \text{int}(1 + \log_2(N))$ y $k = \frac{\max(x) - \min(x)}{h}$

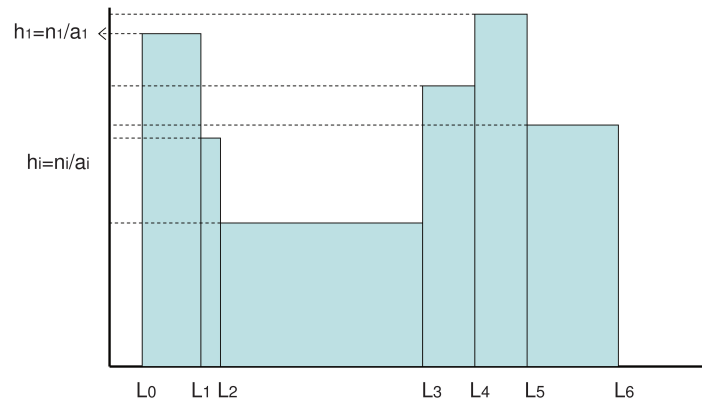
Regla de Scott : $h = 3.49 \frac{S}{\sqrt[3]{N}}$, con S la desviación típica de los datos y $k = \frac{\max(x) - \min(x)}{h}$

Regla de Freedman-Diaconis : $k = 2 \frac{RIQ}{\sqrt[3]{N}}$

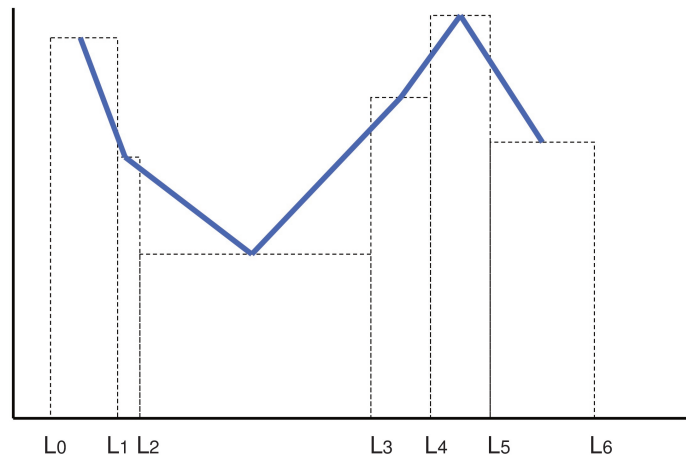
Suposiciones para la regla de Scott: asintóticamente cierta, el exponente -1/3 es para funciones suaves y el coeficiente 3.49 se obtiene suponiendo función de densidad gaussiana.

- Establecer la longitud de clase: es igual al rango entre el número de clases.
- Construir los intervalos de clases: Los intervalos resultan de dividir el rango de los datos en relación al resultado del PASO 2 en intervalos iguales. Obtener la frecuencia de datos dentro de cada intervalo.

- e. Dibujar el histograma: En caso de que las clases sean todas de la misma amplitud, se hace un gráfico de barras unidas, las bases de las barras son los intervalos de clases y altura son la frecuencia (la relativa o la absoluta) de las clases. En caso de intervalos con longitudes desiguales la altura del intervalo es la densidad de frecuencia.



Polígono de frecuencias}: Se obtiene al unir por segmentos los puntos medios de las bases superiores de los rectángulos del histograma.



- Para variables cualitativas}

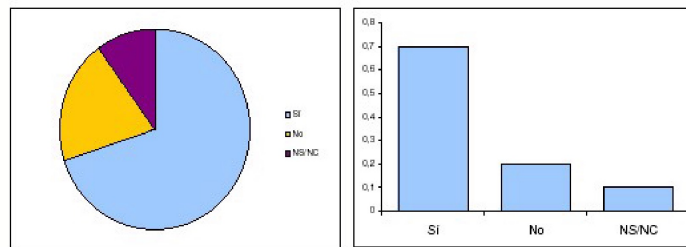
Pictogramas: Los datos se representan gráficamente sustituyendo las barras del diagrama de barras o los rectángulos del histograma por figuras alusivas al carácter estudiado.

Cartogramas: Es la forma de representar sobre el mapa de la región estudiada los caracteres correspondientes, bien empleando distintos colores, tramas de

distinta intensidad, o con una numeración adecuada.

Diagramas sectoriales: Consiste en representar mediante sectores circulares las distintas modalidades de la variable de manera que los sectores han de tener un ángulo central proporcional a la frecuencia correspondiente.

Diagrama de rectángulos: Todos los rectángulos tienen la misma base y sus áreas son proporcionales a la frecuencia.



Nube de palabras: En textos, cuando hay muchas palabras:

```
> library(tm)
> library(wordcloud)
> library(colorspace, pos=18)
> mostra <- sample(x=c(letters, LETTERS, 0:9), size=1000, replace=TRUE)
> tb <- table(mostra)
> wordcloud(attr(tb,"dimnames")[[1]], freq=tb, colors=rainbow(length(tb)) ) )
```



```
> apellidos.nomes <- read.csv("data/2020-2021estudiantes ESEI.csv", header=FALSE)
> apellidos <- unique(apellidos.nomes$V1)
> apel <- unlist(strsplit(as.character(apellidos), split=" "))
> freq <- table(apel)
> wordcloud(attr(freq,"dimnames")[[1]],freq=freq, colors=rainbow(length(freq)))
```

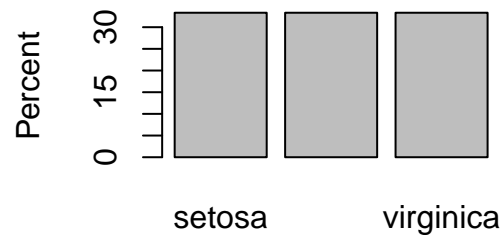


```
> data(iris, package="datasets")
> ### Representa gráficamente la v. cualitativa nominal Species.
> ### Con R Commander.  Menú Gráficas/Diagrama de barras
> with(iris, Barplot(Species, xlab="Frecuencias absolutas", ylab="Frequency"))
```



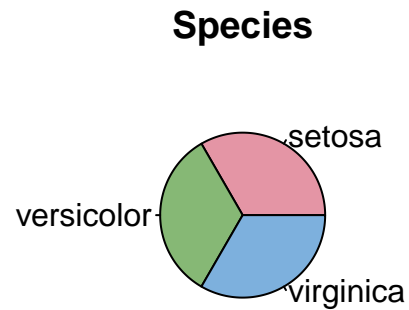
Frecuencias absolutas

```
> with(iris, Barplot(Species, xlab="Porcentajes", ylab="Percent", scale="percent"))
```

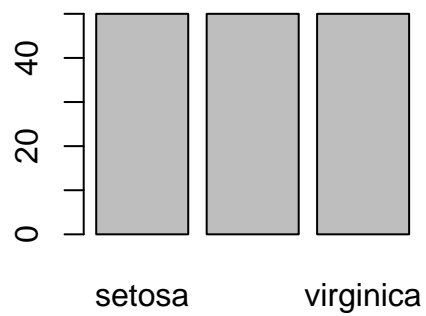


Porcentajes

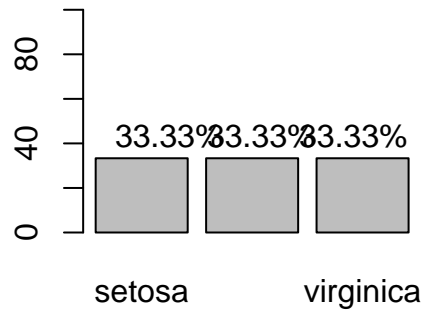
```
> ### Con R Commander. Menú Gráficas/Diagrama de sectores
> with(iris, pie(table(Species), labels=levels(Species), xlab="", ylab="",
+   main="Species", col=rainbow_hcl(3)))
```



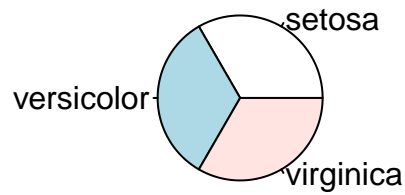
```
> ### Con R
> b <- table(iris$Species)
> barplot(b) #frec. absolutas
```



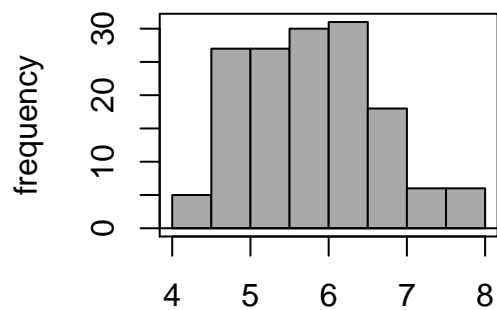
```
> barplot(100*b/sum(b),ylim=c(0,100)) #frec. relativas en tanto por cen
> text(1:3,100*b/sum(b)+10,labels=paste(round(100*b/sum(b),2),"%",sep=""))
```



```
> pie(b) # de sectores
```

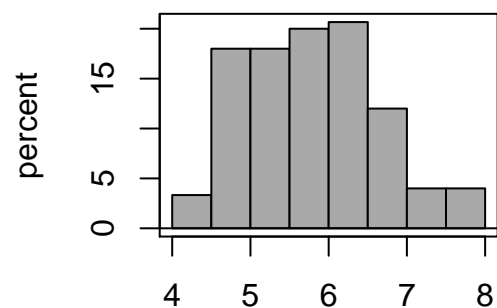


```
> ### Representa gráficamente la v. continua Sepal.Length
> ### Histogramas con R Commander. Menú Gráficas/Histograma ...
> with(iris, Hist(Sepal.Length, scale="frequency", breaks="Sturges",
+             col="darkgray", xlab="Frecuencias absolutas"))
```

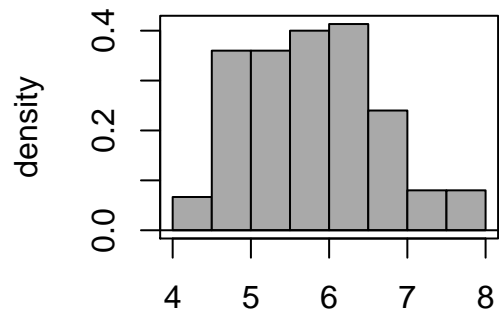
Frecuencias absolutas

```
> with(iris, Hist(Sepal.Length, scale="percent", breaks="Sturges",  
+               col="darkgray", xlab="Porcentajes"))
```



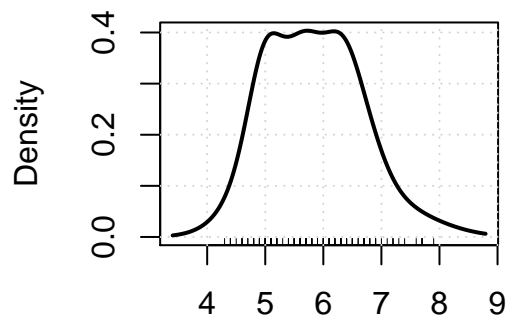
Porcentajes

```
> with(iris, Hist(Sepal.Length, scale="density", breaks="Sturges",  
+               col="darkgray", xlab="Densidad de frecuencia"))
```



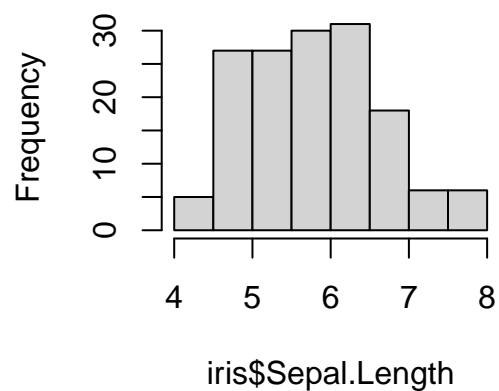
Densidad de frecuencia

```
> ### Estimación no paramétrica de la densidad con R Commander.
> ### Menú Gráficas/Estimar densidad
> densityPlot( ~ Sepal.Length, data=iris, bw=bw.SJ, adjust=1,
+             kernel=dnorm, method="adaptive")
```

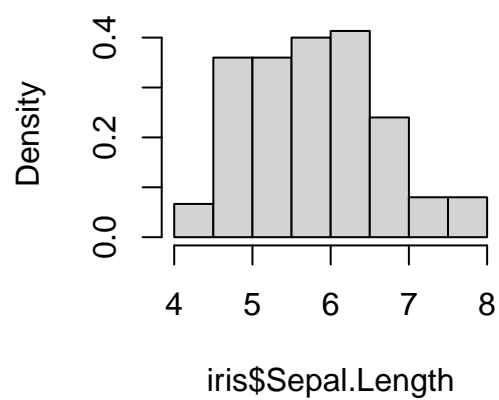


Sepal.Length

```
> ### Histogramas con R
> hist(iris$Sepal.Length) # Frecuencias absolutas
> hist(iris$Sepal.Length, freq=TRUE) #Frecuencias absolutas
```

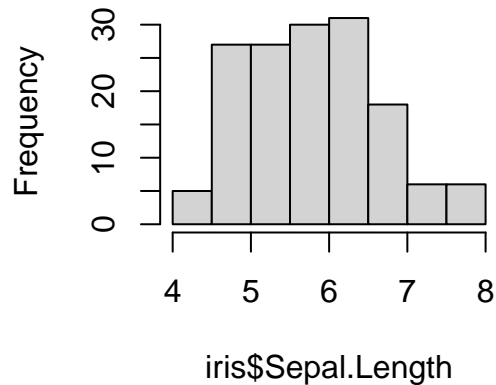
Histogram of iris\$Sepal.Length

```
> hist(iris$Sepal.Length, freq=FALSE) #densidad de frecuencias
```

Histogram of iris\$Sepal.Length

```
> hist(iris$Sepal.Length, breaks=seq(4,8,0.5)) # mod. intervalos
```

Histogram of iris\$Sepal.Length



```
> hist(iris$Sepal.Length, plot=FALSE) # Frecuencias relativas

$breaks
[1] 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0

$counts
[1] 5 27 27 30 31 18 6 6

$density
[1] 0.06666667 0.36000000 0.36000000 0.40000000 0.41333333 0.24000000 0.08000000
[8] 0.08000000

$mids
[1] 4.25 4.75 5.25 5.75 6.25 6.75 7.25 7.75

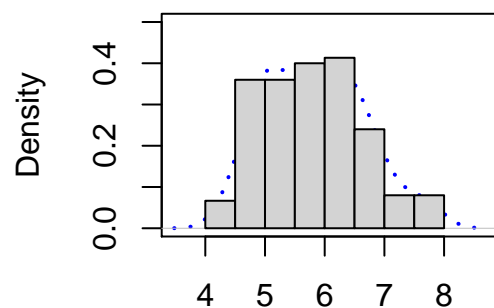
$xname
[1] "iris$Sepal.Length"

$equidist
[1] TRUE

attr("class")
[1] "histogram"

> ### Estimación no paramétrica de la densidad con R.
> plot(density(iris$Sepal.Length), col="blue",lwd=2, lty=3, ylim=c(0,0.5) )
> hist(iris$Sepal.Length, freq= FALSE, add=TRUE)
```

density.default(x = iris\$Sepal.Length)

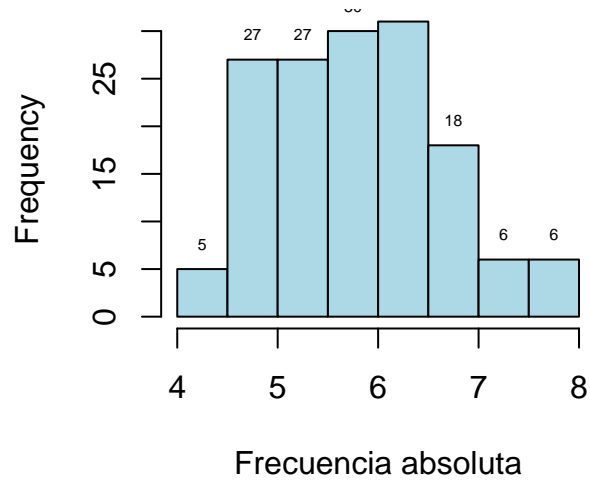


N = 150 Bandwidth = 0.2736

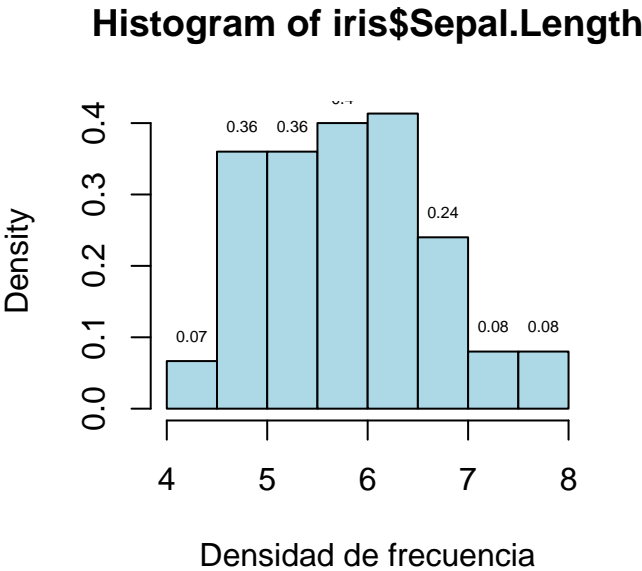
Ejercicio 2.2. Observando los histogramas siguientes, en el eje de abscisas aparece el valor de la altura de cada barra, que porcentaje de flores tienen una longitud de Sepalo menor que 6

```
> aa <- hist(iris$Sepal.Length, plot=FALSE)
> # con Freq. Absolutas
> hist(iris$Sepal.Length, col = "lightblue", , xlab="Frecuencia absoluta")
> text( labels=aa$counts, x= aa$mids, y=aa$counts, cex=0.5, pos=3)
```

Histogram of iris\$Sepal.Length



```
> # con densidad de frecuencias  
> hist(iris$Sepal.Length, freq=FALSE, col = "lightblue", xlab="Densidad de frecuencia",  
> text( labels=round(aa$density,2), x= aa$mids, y=aa$density, cex=0.5, pos=3)
```



Ejercicio 2.3. Dado los valores recogidos en la tabla siguiente, calcula el área del recinto de cada uno de los gráficos anteriores.

Punto medio	4.25	4.75	5.25	5.75	6.25	6.75	7.25	7.75
n_i	5	27	27	30	31	18	6	6
d_i	0.0667	0.3600	0.3600	0.4000	0.4133	0.2400	0.0800	0.0800

Ejercicio 2.4. Completa las siguientes tablas de frecuencias:

Cylinders	n_i	f_i	N_i	F_i
3	3	—	—	—
4	49	—	—	—
5	2	—	—	—
6	31	—	—	—
8	7	—	—	—

FuelCapacity	n_i	f_i	N_i	F_i
[9.18, 12.8)		0.12		
[12.8, 16.3)		0.38		
[16.3, 19.9)		0.31		
[19.9, 23.4)		0.18		
[23.4, 27]		0.01		

HighwayMPG	n_i	f_i	N_i	F_i
[20, 26)			22	
[26, 32)			71	
[32, 38)			88	
[38, 44)			91	
[44, 50]			93	

Horsepower	n_i	f_i	N_i	F_i
[54.8, 104)				0.26
[104, 153)				0.59
[153, 202)				0.88
[202, 251)				0.95
[251, 300]				1.00

Conjunto de datos *cars93* del plugin. N=93

2.3 Principales características numéricas

La información suministrada por una tabla de distribución de frecuencias se puede resumir en un conjunto de medidas que la caracterizan y que se pueden clasificar en:

1. Medidas de posición: proporcionan valores que determinan posiciones dentro del conjunto de los datos. Las dividimos en medidas de tendencia central y de tendencia no central.
2. Medidas de dispersión: indican la desviación de los datos respecto de ciertas medidas de posición.
3. Medidas de forma: relacionadas con la representación gráfica de la distribución.

2.4 Medidas de posición de tendencia central

Media aritmética (\bar{X}): cociente entre la suma de todos los valores observados de la variable y el número total de observaciones:

$$\bar{X} = \frac{\sum_{i=1}^k X_i n_i}{N}$$

Para datos agrupados se toman como x_i las marcas de clase (representante de todos los datos del intervalo).

Propiedades

1. $\sum_{i=1}^k (x_i - \bar{x})n_i = 0$.
2. La media de las desviaciones al cuadrado de los valores de la variable respecto a una constante c cualquiera, se hace mínima cuando esa constante c es igual a la media aritmética. Es decir:

$$\min_c \sum_{i=1}^k (x_i - c)^2 n_i = \sum_{i=1}^k (x_i - \bar{x})^2 n_i \text{ (Teorema de König)}$$

3. Linealidad de la media: dados $a, b \in \mathbb{R}$ e $Y = a + bX$ se verifica:

$$\bar{y} = a + b\bar{x}$$

4. Media en subpoblaciones: Si el total de datos se estratifica en L grupos distintos, la media aritmética del total es una media aritmética de las distintas medias de los estratos ponderadas por el número de observaciones que tienen los mismos:

$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \dots + \bar{x}_L N_L}{N_1 + N_2 + \dots + N_L}$$

5. $\min(x_i) \leq \bar{X} \leq \max(x_i)$

NOTA: A veces se introducen unos coeficientes de ponderación o pesos denominados w_i que son distintos de n_i con lo que la +media ponderada+ es:

$$\bar{x}_w = \frac{\sum_{i=1}^k w_i x_i}{\sum_i w_i}$$

Ejemplo 2.2. Calcula la media aritmética de los 10 primeros números naturales. $X = 10$ primeros números enteros. $x_1 = 1, x_2 = 2, \dots, x_{10} = 10$, frecuencias $n_i = 1, i = 1, 2, \dots, 10$, por lo tanto $\bar{x} = \frac{1+2+\dots+10}{10} = \frac{55}{10} = 5.5$

Ejemplo 2.3. Supón ahora que la ponderación de cada valor es inversamente proporcional a su valor. $x_1 = 1, x_2 = 2, \dots, x_{10} = 10$, con pesos $w_i = \frac{1}{i}$, $i = 1, 2, \dots, 10$, por lo tanto $\bar{x}_w = \frac{1*\frac{1}{1}+2*\frac{1}{2}+\dots+10*\frac{1}{10}}{\frac{1}{1}+\frac{1}{2}+\dots+\frac{1}{10}} = \frac{10}{2.928968} = 3.414172$

Ejemplo 2.4. Haz los cálculos anteriores con R. Genera una función para calcular media aritmética y media ponderada.

Media geométrica: la raíz N-ésima del producto de los N valores de la distribución elevados al número de veces que se repite cada uno de ellos:

$$g = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$$

Podemos expresar la media geométrica como una media aritmética teniendo en cuenta que el logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable:

$$\log(g) = \frac{\sum_{i=1}^k \log(x_i) n_i}{N}$$

El empleo más frecuente de la media geométrica es el de promediar porcentajes, tasas, números índices, etc.

Mediana (Me): dada una distribución de frecuencias $\{(x_i, n_i)\}_{i=1}^k$ con valores ordenados de menor a mayor, llamamos mediana, Me , al valor de la variable que deja a su izquierda la misma frecuencia que a su derecha. Es decir, serían el valor o valores de la variable que son mayores o iguales que la mitad de los datos y menores o iguales a la otra mitad de los datos.

Cálculo de la mediana:

1. Distribuciones no agrupadas: Se observa cuál es la primera frecuencia absoluta acumulada N_i que supera o iguala a $N/2$ distinguiéndose dos casos:
 - Si $N_i > N/2$, entonces $Me = x_i$.
 - Si $N_i = N/2$, entonces $Me = \frac{x_i + x_{i+1}}{2}$.
2. Distribuciones agrupadas: el cálculo se realiza en dos etapas:
 - Detectar el *intervalo mediano*: $[L_{i-1}, L_i)$ tal que $N_i \geq \frac{N}{2}$.
 - Cálculo de la mediana mediante semejanza de triángulos:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = L_{i-1} + \frac{\frac{1}{2} - F_{i-1}}{f_i} a_i$$

siendo L_{i-1} : el extremo inferior del intervalo mediano.

N_{i-1} : la frecuencia absoluta acumulada del intervalo anterior al mediano.

n_i : la frecuencia absoluta del intervalo mediano.

a_i : la amplitud del intervalo mediano.

Ejemplo 2.5. Calcula la mediana de la siguiente tabla de frecuencias:

x_i	n_i	N_i
1	1	1
2	1	2
3	1	3
4	1	4
	$N = 4$	

Variable discreta y como $N_2 = \frac{4}{2}$, la mediana es $\frac{x_i + x_{i+1}}{2} = 2.5$

Ejemplo 2.6. Calcula la mediana de la siguiente tabla de frecuencias:

x_i	n_i	N_i
1	10	10
2	5	15
3	15	30
4	10	40
	$N = 40$	

Variable discreta y como $N_3 = 30 > \frac{40}{2} = 20$, la mediana es 3

Ejemplo 2.7. Calcula la mediana de los 10 primeros números naturales. $X = 10$ primeros números enteros. $x_1 = 1, x_2 = 2, \dots, x_{10} = 10$, frecuencias $n_i = 1$, $i = 1, 2, \dots, 10$, por lo tanto la variable es discreta (no agrupada). El vector de frecuencias acumuladas es $N_1 = 1, N_2 = 2, \dots, N_{10} = 10$, el primer valor que es mayor o igual que $\frac{10}{2} = 5$ es $x_5 = 5$, como $N_5 = \frac{10}{2}$, entonces la mediana es $M_e = \frac{x_5 + x_6}{2} = 5.5$. Coincide con la media.

Ejemplo 2.8. Supón ahora que la ponderación de cada valor es inversamente proporcional a su valor. $x_1 = 1, x_2 = 2, \dots, x_{10} = 10$, con pesos $w_i = \frac{1}{i}$, $i = 1, 2, \dots, 10$, El vector de frecuencias acumuladas es $(N_1 = 1, N_2 = 1.5, \dots, N_{10} = 2.928968) = ((1, 1.5, 1.833, 2.0833, 2.2833, 2.45, 2.593, 2.7178, 2.828, 2.929)$, el primer valor que es mayor o igual que $\frac{2.928968}{2} = 1.464484$ es $x_2 = 1.5$, como es estrictamente mayor, entonces la mediana es $M_e = x_2 = 2$

Ejemplo 2.9. Calcula la mediana de la tabla agrupada del ejercicio de la pág. 4. El vector de frecuencias acumuladas es $(N_1 = 10, N_2 = 42, N_3 = 92, N_4 = 100)$, el primer índice que es mayor o igual que $\frac{100}{2} = 50$ es $i = 3$, el intervalo es $(28, 32]$, como es estrictamente mayor, entonces la mediana es

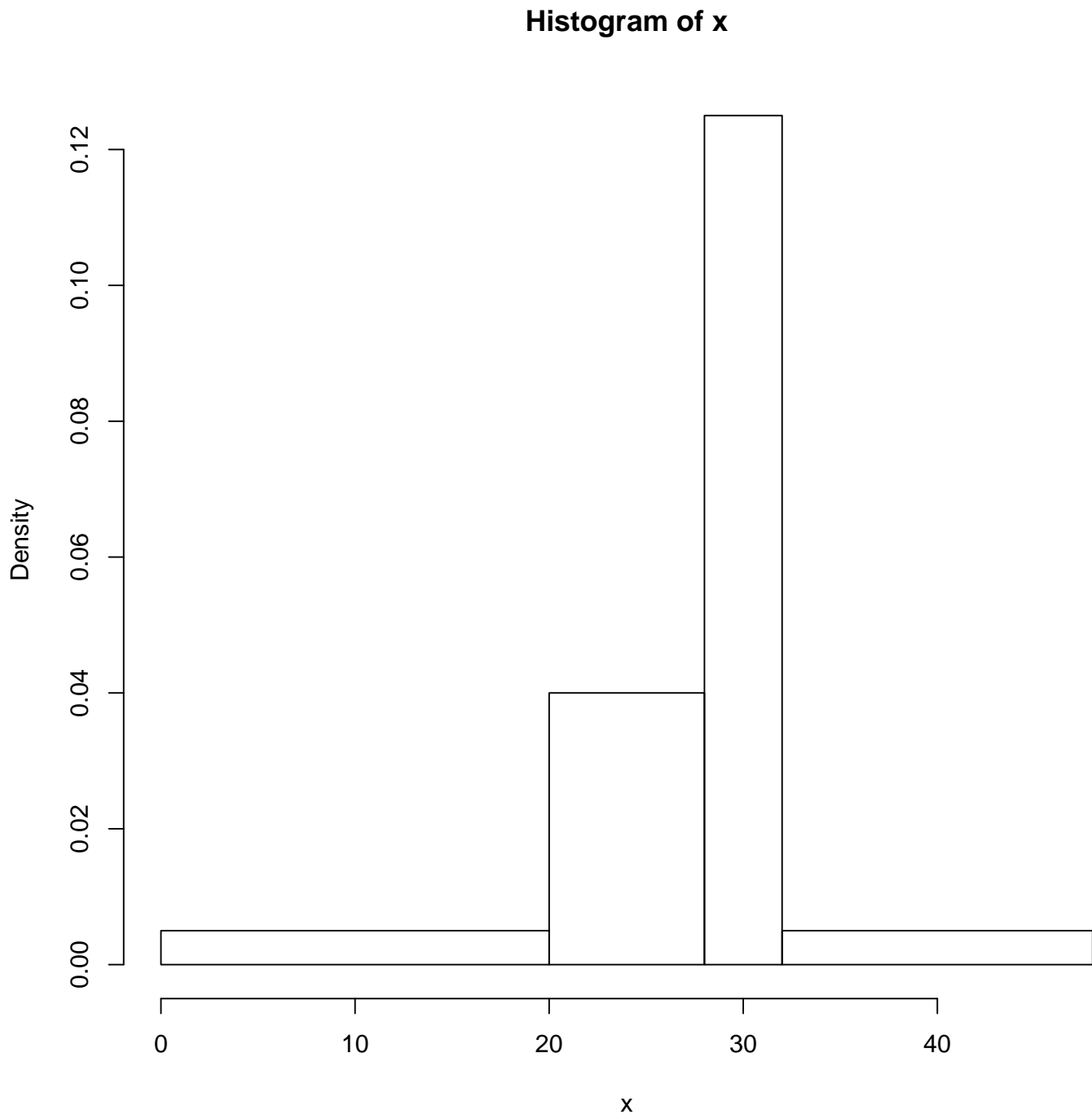
$$M_e = 28 + \frac{50 - 42}{50} * 4 = 28 + \frac{32}{50} = 28 + 0.64 = 28.64$$

Usando el paquete de R *actuar*

```
> ### Calculo con R
> #install.packages("actuar")
> require(actuar)      # instalar el paquete actuar previamente!
> Li <- c(0, 20, 28, 32, 48)      # limites de los intervalos
> ni <- c(10,32,50,8)
> x <- grouped.data(Group = Li, Frequency = ni)
> # x
> quantile(x,probs=0.5)      # función quantile.grouped.data()
```

```
50%
28.64
```

```
> # codigo actuar::quantile.grouped.data  
> hist(x) # histograma para datos agrupados --> función hist.grouped.data()
```



```

> # codigo actuar::hist.grouped.data
>
> var.group <- function (x, ...)
+ {
+   cj <- eval(expression(cj), envir = environment(x))
+   midpoints <- cj[-length(cj)] + diff(cj)/2
+   x <- as.matrix(x[-1])
+   drop(crossprod(x, midpoints^2)/colSums(x) - (crossprod(x, midpoints)/colSums(x))^2 )
+ }
>
> # exemplo de transformación unha táboa de frecuencias
> xx <- runif(1000)
> cj <- table(cut(xx, breaks=(0:10)/10))
> x <- grouped.data( Group=(0:10)/10, Freq= as.vector(cj))

```

Moda (Mo): es el valor de la variable que más veces se repite, es decir, el más frecuente. La moda puede no ser única, puede haber una moda (variable unimodal), dos modas (bimodal), etc.

Cálculo de la moda:

- Distribuciones no agrupadas: valor de la variable de mayor frecuencia absoluta o relativa.
- Distribuciones agrupadas: el cálculo se realiza en dos etapas:
 - Detectar el *intervalo modal*: $[L_{i-1}, L_i)$ con mayor densidad de datos d_i .
 - Cálculo de la moda:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i+1} + d_{i-1}} a_i$$

siendo L_{i-1} : el extremo inferior del intervalo modal,

d_{i-1} y d_{i+1} : las densidades de frecuencia de los intervalos anterior y posterior al modal respectivamente. (En caso de no existir intervalo anterior o posterior se consideran igual a 0),

a_i : la amplitud del intervalo modal.

Ejemplo 2.10. Calcula la moda en el ejercicio de la tabla agrupada del Ejercicio 2.1 Dado el vector de densidades ($d_1 = 0.5, d_2 = 4, d_3 = 12.5, d_4 = 0.5$), el intervalo (modal) con mayor densidad es $(28, 32]$. El valor de la moda es por lo tanto

$$Mo = 28 + \frac{0.5}{4 + 0.5} * 4 = 28 + \frac{2}{4.5} = 28 + 0.444 = 28.444$$

2.5 Medidas de posición no centrales: los Cuantiles

Se define *cuantil de orden p* con $0 < p < 1$ (x_p), como el valor que deja a lo sumo pN observaciones a su izquierda y $(1 - p)N$ observaciones a su derecha. Destacamos en particular los cuantiles siguientes:

- Los *cuartiles* (Q_1, Q_2 y Q_3): dividen a la distribución en cuatro partes iguales, dentro de cada cual están incluidos el 25% de los valores de la distribución.
- Los *deciles* (D_1, D_2, \dots, D_9): dividen a la distribución en diez partes iguales; dentro de cada una están incluidos el 10% de los valores.
- Los *percentiles* (P_1, P_2, \dots, P_{99}): dividen a la distribución en cien partes iguales; dentro de cada una está incluido el 1% de los valores.

Cálculo del cuantil de orden p

1. Distribuciones no agrupadas: Se observa cuál es la primera frecuencia absoluta acumulada N_i que supera o iguala a pN distinguiéndose dos casos:
 - Si $N_i > pN$, entonces $x_p = x_i$.
 - Si $N_i = pN$, entonces $x_p = \frac{x_i + x_{i+1}}{2}$.
2. Distribuciones agrupadas: Primero se detecta el intervalo que contiene al cuantil: el primer $[L_{i-1}, L_i)$ con $N_i \geq pN$, luego se aplica la fórmula:

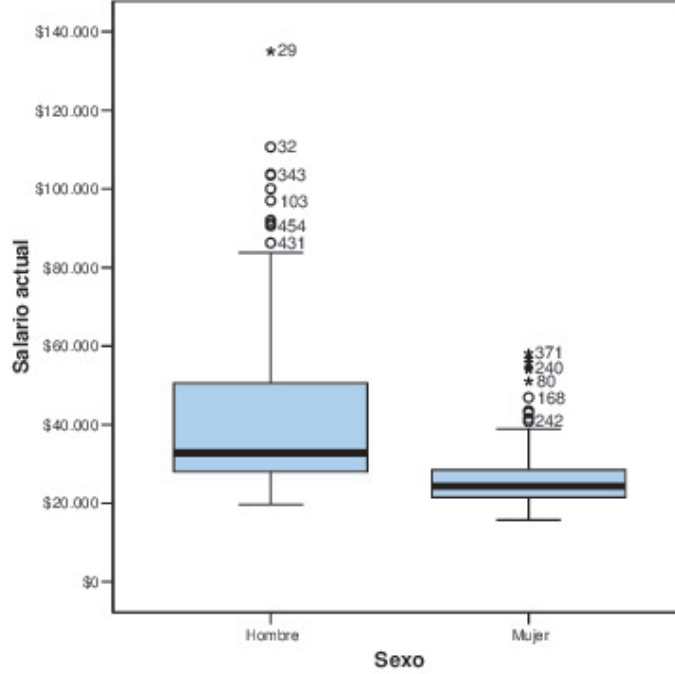
$$x_p = L_{i-1} + \frac{pN - N_{i-1}}{n_i} a_i = L_{i-1} + \frac{p - F_{i-1}}{f_i} a_i$$

Diagramas de caja

Este tipo de gráficos también se conocen como Box-Plots. Para construirlos es necesario el cálculo del primer y tercer cuartil Q_1 y Q_3 . Se definen el extremo inferior y superior LI y LS respectivamente como:

$$LI = \max \left\{ \min \{x_i\}, Q_1 - 3 \left(\frac{Q_3 - Q_1}{2} \right) \right\} \quad LS = \min \left\{ \max \{x_i\}, Q_3 + 3 \left(\frac{Q_3 - Q_1}{2} \right) \right\}$$

obteniéndose un diagrama de caja de la forma:



Ejercicio 2.5. Genera un diagrama de cajas de la variable Ancho del Pétalo en función del tipo de especie. Que conclusión en términos de los cuartiles de *Versicolor* podemos obtener?

Momentos Potenciales

Los momentos de una distribución son unos valores que la caracterizan, de tal modo que dos distribuciones son iguales si tienen todos sus momentos iguales, y son tanto más parecidas cuanto mayor sea el número de momentos iguales que tengan.

Momentos respecto al origen

El momento de orden r con respecto al origen se define como:

$$a_r = \frac{\sum_{i=1}^k x_i^r n_i}{N} = \sum_{i=1}^k x_i^r f_i$$

Los primeros momentos serán: $a_0 = 1$, $a_1 = \bar{X}$.

Momentos respecto a la media aritmética:

El momento de orden r con respecto a la media aritmética se define como:

$$m_r = \frac{\sum_{i=1}^k (x_i - \bar{x})^r n_i}{N} = \sum_{i=1}^k (x_i - \bar{x})^r f_i$$

Los primeros momentos serán: $m_0 = 1$, $m_1 = 0$, $m_2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{N} = s^2$.

Propiedad:

Utilizando el binomio de Newton, los momentos con respecto a la media se pueden expresar en función de los momentos con respecto al origen. En particular:

$$\begin{aligned} m_2 &= a_2 - a_1^2 \\ m_3 &= a_3 - 3a_2a_1 + 2a_1^3 \end{aligned}$$

Como ejercicio se propone expresar m_r los momentos respecto a la media aritmética de orden r en función de los momentos con respecto al origen.

2.6 Medidas de Dispersión

Son medidas que nos indican la desviación de los valores de la variable respecto de ciertas medidas de posición como la media aritmética o la mediana. A la mayor o menor separación de los valores respecto a otro, que se pretende sea su síntesis, se llama dispersión o variabilidad.

Medidas de Dispersión Absolutas:

Recorrido: es la diferencia entre el mayor valor y el menor valor de una distribución: $Re = x_k - x_1 = \max_{i=1 \dots k} x_i - \min_{i=1 \dots k} x_i$.

Recorrido Intercuartílico: es la diferencia existente entre el tercer cuartil y el primero: $IRQ = C_3 - C_1$.

Nos indica que en un intervalo de longitud IRQ están comprendidos el 50% de los valores centrales.

Varianza: De todas las medidas de dispersión absolutas respecto a la media aritmética, la varianza y su raíz cuadrada, la desviación típica son las más importantes.

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{N}$$

Desviación típica o estándar: Es la raíz cuadrada de la varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{N}}$$

Propiedades de la varianza:

1. La varianza y la desviación típica **NUNCA** pueden ser negativas $s^2 \geq 0$, $s \geq 0$.
2. La varianza es la medida cuadrática de dispersión óptima, ya que:

$$\min_c \frac{1}{N} \sum_{i=1}^k (x_i - c)^2 n_i = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x})^2 n_i \text{ (Teorema de König)}$$

3. La varianza es igual al momento de segundo orden respecto al origen menos el de primer orden elevado al cuadrado.

$$s^2 = a_2 - a_1^2 = \left(\sum_{i=1}^k x_i^2 f_i \right) - \left(\sum_{i=1}^k x_i f_i \right)^2$$

4. Si en la distribución de frecuencias sumamos a todos los valores de la variable una constante, la varianza no varía (un cambio de origen no afecta a la varianza).

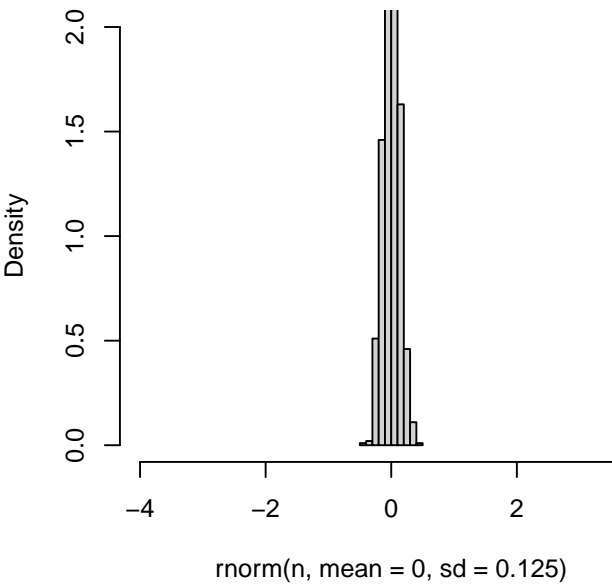
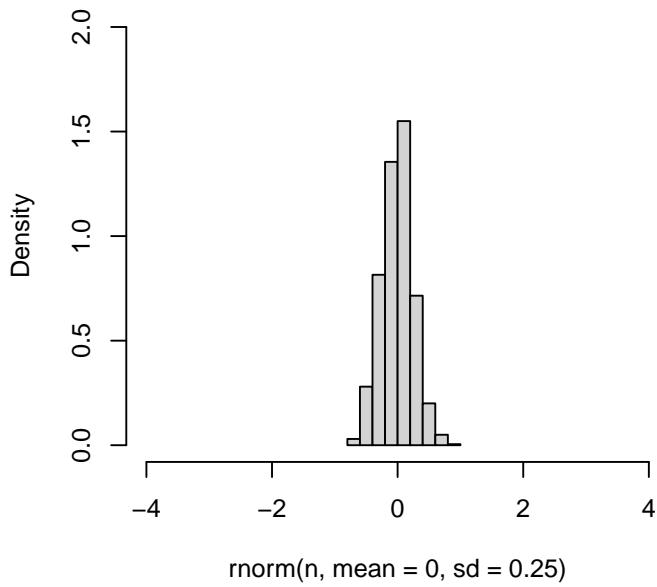
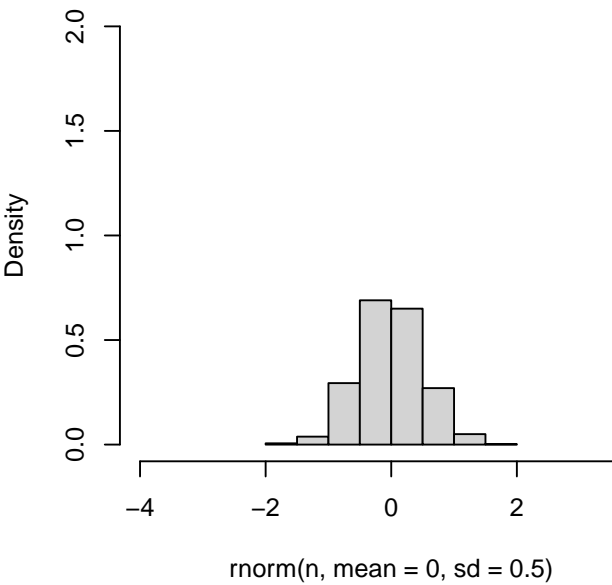
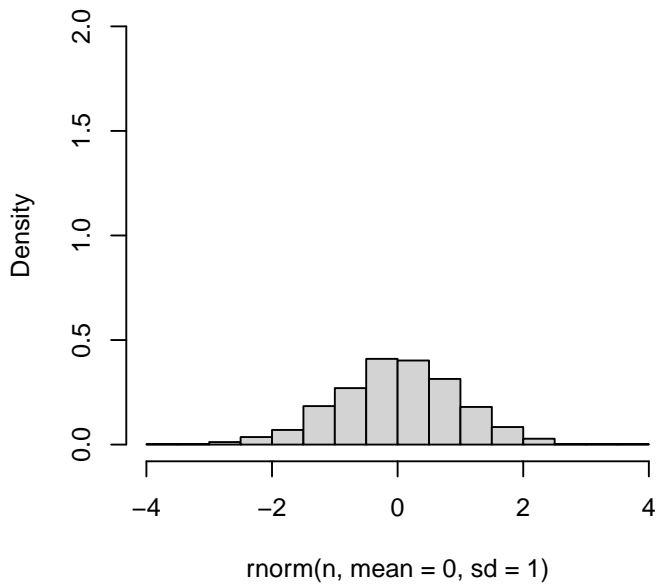
$$y_i = x_i + a ; s_Y^2 = s_X^2$$

5. Al multiplicar todos los valores de una distribución de frecuencias por una constante, la varianza queda multiplicada por el cuadrado de dicha constante.

$$y_i = bx_i ; s_Y^2 = b^2 s_X^2$$

Interpretación

```
> ### Código de R
> n <- 1000
> par( mfrow=c(2,2))
> hist( rnorm( n, mean=0, sd= 1), freq=FALSE , ylim=c(0,2), xlim=c(-4,4), main="")
> hist( rnorm( n, mean=0, sd= 0.5), freq=FALSE , ylim=c(0,2), xlim=c(-4,4), main="")
> hist( rnorm( n, mean=0, sd= 0.25), freq=FALSE , ylim=c(0,2), xlim=c(-4,4), main="")
> hist( rnorm( n, mean=0, sd= 0.125), freq=FALSE , ylim=c(0,2), xlim=c(-4,4), main="")
```



```
> par( mfrow=c(1,1))
```

2.7 Medidas de Dispersión relativas

Si queremos comparar los promedios de dos distribuciones para saber cuál de los dos es más representativo debemos utilizar medidas adimensionales, es decir, que no vengan afectadas por las unidades de medida. Estas medidas de dispersión, llamadas relativas, siempre se concretan en forma de cociente. Las más utilizadas son:

Recorrido relativo: Se define como el cociente entre el recorrido y la media aritmética:

$$R_r = \frac{Re}{\bar{x}}$$

Coefficiente de variación de Pearson: se define como la relación por cociente entre la desviación típica y la media aritmética. Si $\bar{x} \neq 0$

$$v = \frac{s}{\bar{x}}$$

Cuanto mayor sea v , menor es la representatividad de la media aritmética.

Ejercicio 2.6. Dado el conjunto de datos *iris*, responde:

1. Agrupa la variable *Sepal.Length* en 8 intervalos de igual longitud. Usa el menú *Estadística Básica/Datos/Modificar el conjunto de datos activo/Segmentar variable numérica . . .*. Obtén su distribución de frecuencias completa.
2. Agrupa nuevamente pero en 8 intervalos con extremos 4 y 8. Usa el menú *Estadística Básica/Datos/Modificar el conjunto de datos activo/Recodificar . . .*. Obtén su distribución de frecuencias completa. Genera una nueva columna con la marca de clase
3. Compara la media da v. *Sepal.Length* agrupada y sin agrupar.
4. Comprueba la propiedad 5 de la media para la v. *Sepal.Length*, i.e., $\min(xi) \leq \text{media} \leq \max(xi)$
5. Recarga el conjunto de datos original y construye el histograma de la v. *Sepal.Length*.
6. Representa y compara en un gráfico la v. *Sepal.Length* en función del tipo de flor. Visualizando el gráfico, que se puede decir del 75% de las longitudes del Sepalo de la subespecie *setosa* con respecto a la subespecie *versicolor* y *virginica*. Y del 50% de la longitud del sepalo de la subespecie *virginica* con respecto a las otras.
7. Haz un resumen numérico de la v. *Sepal.Length* en función del tipo de subespecie.

8. Que proporción de valores de la v. *Sepal.Length* están dentro del intervalo $\bar{x} - 1.96 * sd_{Sepal.Length}$, $\bar{x} + 1.96 * sd_{Sepal.Length}$
9. Edita el conxunto de datos y pon el 1º valor de la v. *Sepal.Length* a 1000. Recalcula la media nuevamente. ¿Sigue siendo ese valor representativo de todas las longitudes del sepalo?.
10. ¿Qué valor representa ahora mejor la posición central?. Que ocurre si se pone el 2º valor= 1000?, y si se hace con el 3º valor también, e con el 4º, ... Cuantos valores se tienen que modificar para que se vea alterada esta nueva medida de posición central?.
11. Añade un valor para que la media de la variable *Petal.Length* sea 4.

```
> head(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

```
> x <- mean(iris$Petal.Length) # A media dos (150+1) datos é: (sum_xi + x0)/151 =4, de aquí desp
> x0 <- 150*(4*151/150 - x); x0
```

```
[1] 40.3
```

```
> # Comprobamos que o valor x0 é o correcto
> mean( c(iris$Petal.Length,x0))
```

```
[1] 4
```

12. Modifica un valor para que la media sea igual a la mediana da v. *Petal.Length*

```
> x <- mean(iris$Petal.Length)
> y <- median(iris$Petal.Length)
> xx <- iris$Petal.Length
> xx[length(xx)] <- xx[length(xx)] + 150*(y-x)
> # Comprobamos que o valor x0 é o correcto
> mean(xx); y
```

```
[1] 4.35
```

```
[1] 4.35
```

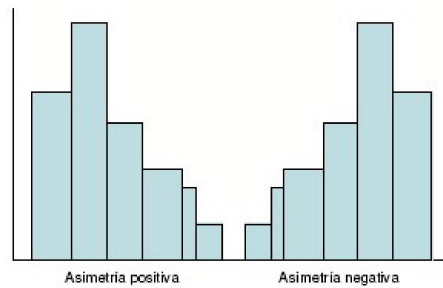
Solución: [iris-resumen-numerico.html](#)

2.8 Medidas de Forma

Están relacionadas con la representación gráfica de la distribución. Pueden ser:

Medidas de simetría:

- Asimetría positiva: Si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en derecho hay frecuencias más pequeñas (cola).
- Asimetría negativa: Cuando la cola está en el lado izquierdo.



Coefficiente de asimetría de Fisher: $g_1 = \frac{m_3}{s^3}$

Para cada coeficiente, existen tres posibilidades: - $g_1 > 0$: La asimetría es positiva o por la derecha.

- $g_1 = 0$: La distribución es simétrica.
- $g_1 < 0$: La asimetría es negativa o por la izquierda.

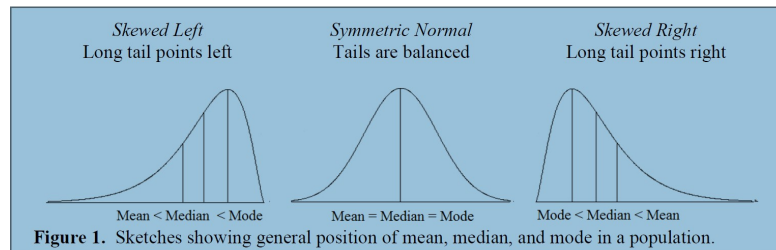


Figura: Extraído de: Doane, D.P., Seward, L.E. (2011). Measuring Skewness: A Forgotten Statistic?. *Journal of Statistics Education*, Volume 19, Number 2.

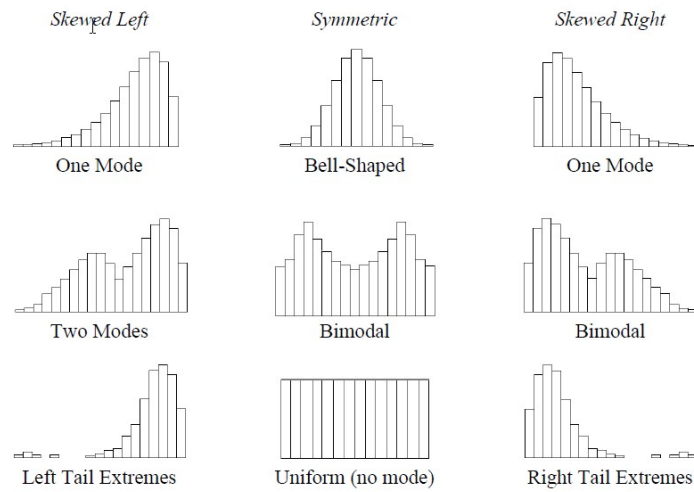
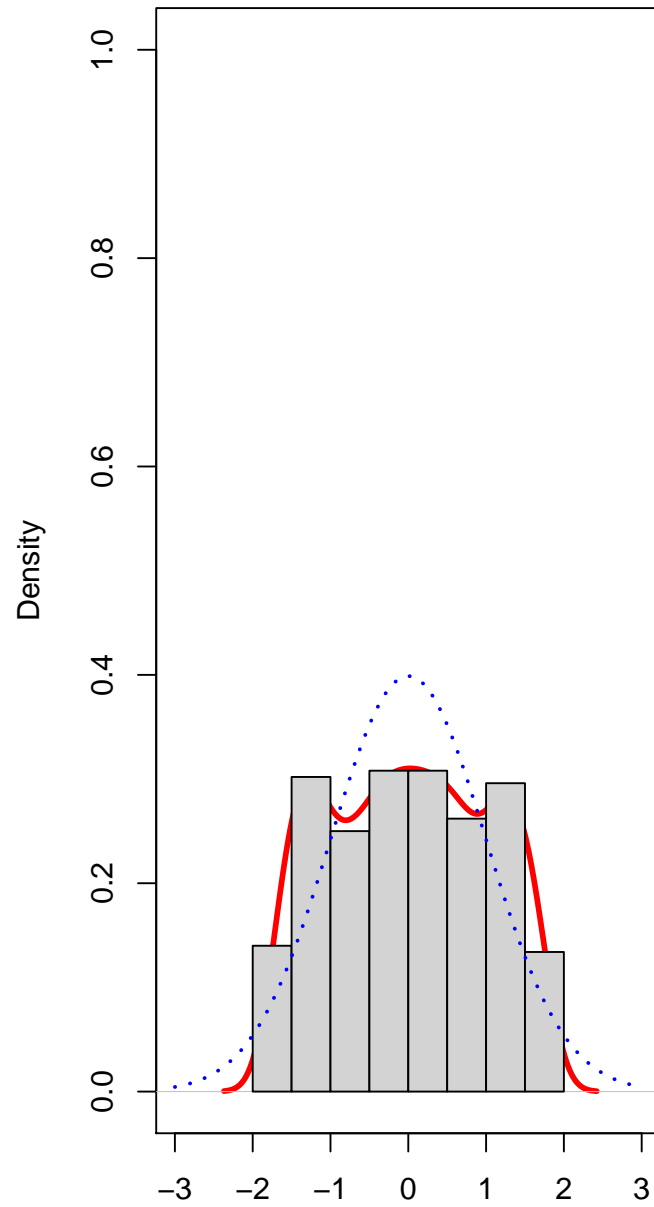
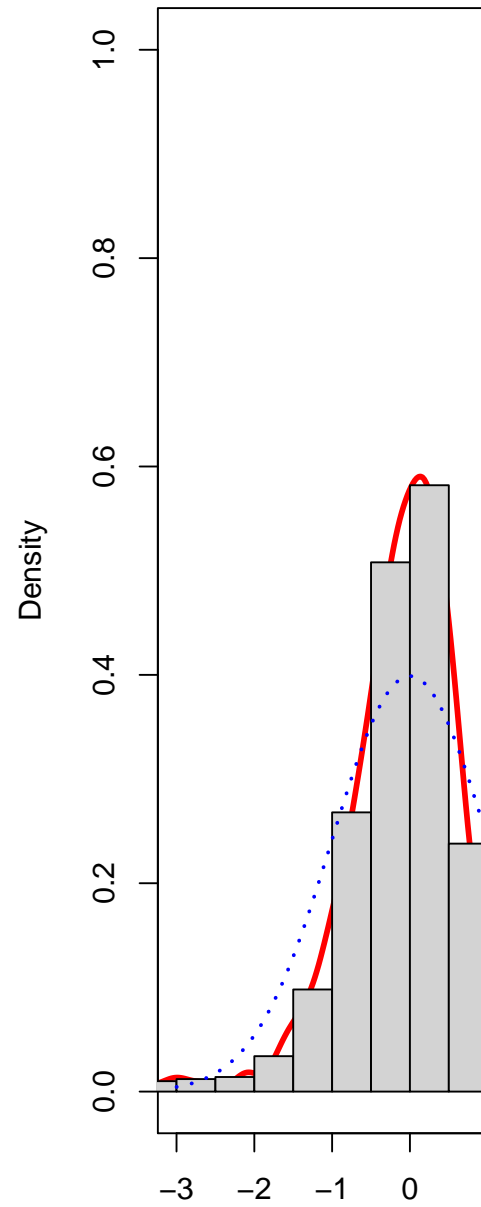


Figure 2. Illustrative prototype histograms.

Medidas de apuntamiento o curtosis

- Coeficiente de curtosis de Fisher: $g_2 = \frac{m_4}{s^4} - 3$
 - $g_2 > 0$: Distribución leptocúrtica: más apuntamiento que la distribución normal.
 - $g_2 = 0$: Distribución mesocúrtica: apuntamiento similar a la distribución normal.
 - $g_2 < 0$: Distribución platicúrtica: menos apuntamiento que la distribución normal.

Coef. Curtosis negativa**Coef. Curtosis p**

Tipificación

Una variable estadística se dice tipificada o estandarizada si su media es cero y su varianza o su desviación típica es uno. Dada una variable X con media μ , y varianza σ^2 , la nueva variable $Z = \frac{X - \mu}{\sigma}$, es su tipificada.

Ejercicio 2.7. Tipifica la variable *Sepal.Length* del conjunto de datos *iris* y comprueba que efectivamente la v. tipificada tiene media 0 y varianza 1. Menú *Estadística Básica/Datos/Modificar variables del conjunto de datos activo/Tipificar*.

```
> data(iris)
> iris <- local({
+   .Z <- scale(iris[,c("Sepal.Length")])
+   within(iris, {
+     Z.Sepal.Length <- .Z[,1]
+   })
+ })
>
> RcmdrMisc::numSummary(iris[, "Z.Sepal.Length"], statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,0.25,0.5,0.75,1))
```

	mean	sd	IQR	0%	25%	50%	75%	100%
	-4.480675e-16	1	1.569923	-1.86378	-0.8976739	-0.05233076	0.672249	2.483699
n								
	150							

Chapter 3

Ejercicios resueltos de Estadística descriptiva unidimensional con R y R-Commander

Tabla de frecuencias para una v. cualitativa ordinal

El archivo de datos *cars93* contiene datos sobre distintos modelos de coche (ver la ayuda).

```
data(cars93, package="RcmdrPlugin.TeachStat")
```

Vamos a crear una tabla de frecuencias da variable Airbags. Dicha variable es cualitativa y tiene orden, por tanto debemos ir a *Estadística Básica/Estadística Descriptiva/Distribución de Frecuencias de variables cualitativas*. En R-Commander, sólo hay que seleccionar la variable como nominal u ordinal.

```
calcular_frecuencia(df.nominal=cars93["Airbags"], ordenado.frec=FALSE, df.ordinal=NULL, cuantil.p
```

```
##
## -----
##
## Variables nominales:
##
## Variable: Airbags
##           ni    fi
## none           34 0.366
## driver          43 0.462
## driver&passenger 16 0.172
```

```
## N= 93
```

Si la variable es de tipo cuantitativo el menú que debemos utilizar para calcular su tabla de frecuencias es *Estadística Básica/Estadística Descriptiva/Resúmenes Numéricos Variables Discretas*.

Tabla de frecuencias para una v. cuantitativa continua agrupada

Con el archivo de datos *cars93*, vamos a crear una tabla de frecuencias de la variable *MaxPrice*. Es una variable cuantitativa de carácter continuo, por tanto la tabla de frecuencias con datos sin agrupar sería intratable. Es necesario agrupar los datos. Para ello, en el menú *Estadística Básica/Estadística Descriptiva/Resúmenes Numéricos Variables Discretas*, debemos seleccionar en el apartado cortes bien el número de intervalos o los extremos de los intervalos que nos interesa considerar. Por ejemplo si queremos cortar en 4 intervalos

```
calcularResumenVariablesDiscretas(data=cars93["MaxPrice"], statistics =c("mean", "sd",
  "IQR", "quantiles"), quantiles = c(0,0.25,0.5,0.75,1), groups=NULL, tablaFrecuencia=
  cortes=4)
```

```
##
## -----
##
## Resumenes numéricos:
##      mean      sd  IQR  0%  25%  50%  75% 100%  n
##  21.89892 11.03046 10.6  7.9 14.7 19.6 25.3   80 93
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: MaxPrice
##      ni      fi  Ni      Fi
## [7.83,25.9) 70 0.7527 70 0.753
## [25.9,43.9) 20 0.2151 90 0.968
## [43.9,62)   2 0.0215 92 0.989
## [62,80.1]   1 0.0108 93 1.000
## N= 93
```

Si deseamos especificar los intervalos, añadimos los puntos de cortes separados por comas:

```
calcularResumenVariablesDiscretas(data=cars93["MaxPrice"], statistics =c("mean", "sd",
  "IQR", "quantiles"), quantiles = c(0,0.25,0.5,0.75,1), groups=NULL, tablaFrecuencia=
  cortes=c(5,25,45,65,85))
```

```
##
## -----
##
```

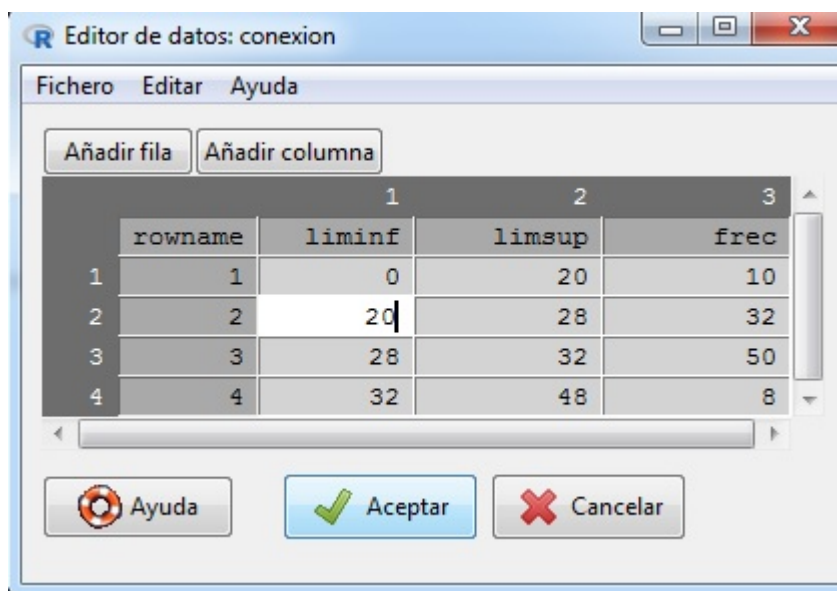
```
## Resúmenes numéricos:
##      mean      sd  IQR  0%  25%  50%  75% 100%  n
##  21.89892 11.03046 10.6  7.9 14.7 19.6 25.3   80 93
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: MaxPrice
##      ni      fi Ni      Fi
## [5,25)  69 0.7419 69 0.742
## [25,45)  22 0.2366 91 0.978
## [45,65)   1 0.0108 92 0.989
## [65,85]   1 0.0108 93 1.000
## N= 93
```

Tabla de frecuencias para una v. cuantitativa continua ya agrupada

La distribución del tiempo de conexión a la red internet en un edificio viene agrupado en intervalos por la siguiente tabla:

Tiempo(minutos)	Marca de clase	nº de Usuarios
(0-20]	10	10
(20-28]	24	32
(28-32]	30	50
(32-48]	40	8

Para ello tenemos que introducir los datos en un archivo nuevo de datos, (*Estadística Básica/Datos/Nuevo Conjunto de datos*) de forma que en una columna esté el límite inferior de cada intervalo, en otra columna el límite superior y en una tercera columna la frecuencia de cada intervalo (en nuestro caso Número de usuarios).



Para realizar la tabla de frecuencias, iremos a *Estadística Básica/Estadística Descriptiva/Resúmenes Numéricos Datos Tabulados*, y marcamos *Tabla de frecuencias*.

Alternativamente podemos cargar el conjunto de datos en el menú *Datos/Cargar conjunto de datos*.

```
load("data/conexion.RData")
calcularResumenDatosTabulados(l_inf=conexion$liminf, l_sup=conexion$limsup, ni=conexion$ni,
  statistics = c("mean", "sd", "IQR", "quantiles"), quantiles = c(0,0.25,0.5,0.75,1),
  tablaFrecuencia=TRUE)
```

```
##
## -----
##
## Resumen numérico ponderado por: ni
##
## -----
##
## Resúmenes numéricos:
## mean      sd      IQR 0%   25%   50%   75% 100%  n
## 26.88 7.055891 11.3975 0 23.75 28.64 30.64 48 100
##
## -----
##
## Distribución de frecuencia para la variable tabulada:
## Li_1 Li xi ni fi Ni Fi ai hi
## 0 20 10 10 0.10 10 0.10 20 0.005
```

```
##    20 28 24 32 0.32  42 0.42   8 0.040
##    28 32 30 50 0.50  92 0.92   4 0.125
##    32 48 40   8 0.08 100 1.00  16 0.005
```

En la tabla podemos ver la marca de clase, las frecuencias absoluta, relativa, absoluta acumulada, relativa acumulada, la amplitud de cada intervalo y la densidad.

Resumen numérico de una v. continua.

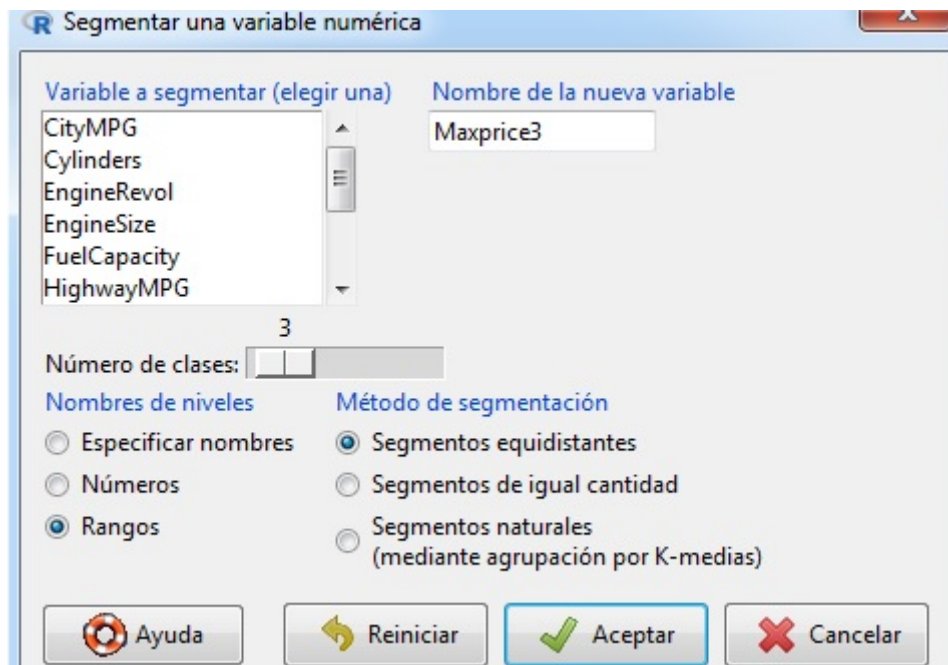
Para la v. *MaxPrice*.

- Calcula los estadísticos descriptivos más importantes de dicha variable (*Estadística Básica/Estadística Descriptiva/Resúmenes Numéricos*).

```
numSummary(cars93["MaxPrice"], statistics=c("mean", "sd", "IQR", "quantiles"),
  quantiles=c(0,0.25,0.5,0.75,1))
```

```
##      mean      sd  IQR  0%  25%  50%  75% 100%  n
## 21.89892 11.03046 10.6  7.9 14.7 19.6 25.3   80 93
```

- Segmenta la variable anterior en 3 intervalos de la misma longitud y calcula la tabla de frecuencias de dicha variable (*Estadística Básica/Datos/Modificar variables del conjunto de datos activo/Segmentar variable numérica*).



- Calcula la tabla de frecuencias de la nueva variable categorizada

```
cars93$Maxprice3 <- with(cars93, bin.var(MaxPrice, bins=3, method='intervals', labels=
calcular_frecuencia(df.nominal=cars93["Maxprice3"], ordenado.frec=FALSE, df.ordinal=NU
cuantil.p=0.5, iprint = TRUE)
```

```
##
## -----
##
## Variables nominales:
##
## Variable: Maxprice3
##          ni      fi
## (7.83,31.9] 77 0.8280
## (31.9,56]   15 0.1613
## (56,80.1]   1 0.0108
## N= 93
```

- Calcula los estadísticos descriptivos básicos de la nueva variable agrupada, y compara los resultados con los de la variable sin agrupar calculados previamente. Los datos agrupados los tenemos que introducir usando el menú *Estadística Básica/Datos/Nuevo conjunto de datos* o bien cargando el archivo *newmaxprice.RData* con el menú *Estadística Básica/Datos/Cargar conjunto de datos*

```
load("data/newmaxprice.RData")
```

```
calcularResumenDatosTabulados(l_inf=newmaxprice$liminf, l_sup=newmaxprice$limsup,
ni=newmaxprice$frec, statistics = c("mean", "sd", "se(mean)", "IQR", "mode", "quantil
"skewness", "kurtosis"), quantiles = c(0,0.25,0.5,0.75,1), tablaFrecuencia=TRUE)
```

```
##
## -----
##
## Resumen numérico ponderado por: ni
##
## -----
##
## Resúmenes numéricos:
##      mean      sd se(mean) IQR      cv skewness kurtosis  0%      25%      50%
## 24.2678 9.957077 1.032501  0 0.4103 2.088321 3.513724 7.83 15.09789 22.36578
##      75% 100% mode  n
## 29.63367 80.1 31.9 93
##
## -----
##
## Distribución de frecuencia para la variable tabulada:
##  Li_1  Li      xi ni      fi Ni      Fi      ai      hi
##  7.83 31.9 19.865 77 0.82795699 77 0.8279570 24.07 0.0343978807
```



```
## 31.90 56.0 43.950 15 0.16129032 92 0.9892473 24.10 0.0066925445
## 56.00 80.1 68.050 1 0.01075269 93 1.0000000 24.10 0.0004461696
```

Datos de Chile

El archivo de datos *Chile* del paquete *car* contiene información sobre una encuesta nacional realizada en 1988 en Chile (ver ayuda), concretamente las variables contenidas en el data.frame Chile son:

- *region*: variable que toma valores *C*, *M*, *N*, *S*, *SA* indicando las regiones Central, area Metropolitana de Santiago, Norte, Sur y Ciudad de Santiago, respectivamente,
- *poblacion*: tamaño de la población de cada región,
- *sex*: dos valores *F* female y *M* male,
- *age*: edad,
- *educacion*: *P* de Primaria, *S* de Secundaria y *PS* Post-Secundaria,
- *income*: ingresos mensuales en Pesos,
- *vote*: *A* indicando abstención, *N* votará No (en contra de Pinochet), *U* indeciso, *Y* vota a favor (de Pinochet)

1. ¿Qué tipo de variables estadísticas son?.

```
library(car)
# help(Chile)
```

- *region*: variable cualitativa nominal
- *poblacion*: variable cuantitativa
- *sex*: variable cualitativa nominal
- *age*: variable cuantitativa continua
- *educacion*: variable cualitativa ordinal
- *income*: variable cuantitativa
- *vote*: variable cualitativa nominal

2. Describe completamente la variable *vote*. Da una representación gráfica adecuada. Interpreta los resultados.

```
library(colorspace, pos=17)
```

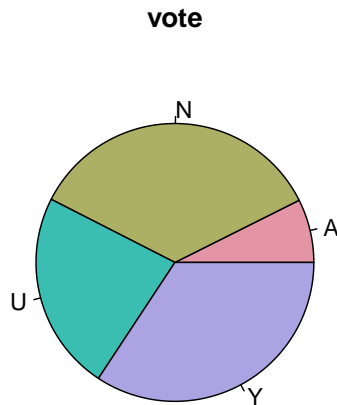
```
## Warning: package 'colorspace' was built under R version 4.0.3
```

```
calcular_frecuencia(df.nominal=Chile["vote"], ordenado.frec=FALSE, df.ordinal=NULL,
  cuantil.p=0.5, iprint = TRUE)
```

```
##
```

```
## -----
##
## Variables nominales:
##
## Variable: vote
##      ni      fi
## A 187 0.0739
## N 889 0.3511
## U 588 0.2322
## Y 868 0.3428
## N= 2532

with(Chile, pie(table(vote), labels=levels(vote), xlab="", ylab="", main="vote",
col=rainbow_hcl(length(levels(vote)))))
```



3. Agrupar la variable *income* en los subintervalos $((0, 10000], (10000, 20000], (20000, 50000], (50000, 250000])$. Da su distribución de frecuencias completa. Dar la representación gráfica más adecuada manteniendo esos intervalos.

Para recodificar una variable, es preciso ir al menú de (*Estadística Básica/Datos/Modificar variables del conjunto de datos activo/Recodificar Variable*). Se selecciona la variable que queremos recodificar y se asigna un nombre a la nueva variable. En el cuadro de *Introducir directrices de codificación*, se establecen los cambios. Si el nuevo valor es categórico hay que delimitarlo por comillas.

El archivo está disponible como *chilerecod.RData*

Notas Examen.

- *id*: indicadora del número de examen.
- *DNI*: nº de identificación nacional.
- *Asistencia*: variable indicadora de asistencia (0/1).
- *Ex1*: nota del primer ejercicio.
- *Ex2*: nota del segundo ejercicio.
- *Ex3*: nota del tercer ejercicio.

- *Nota*: nota final del examen.

1. Importa el archivo de datos (*Estadística Básica/Datos/Importar Datos/Desde archivos de texto...*). El separador de campos es el punto y coma.

```
##          id          DNI          Asistencia          Ex1
## Min.   : 1.00   Min.   : 456997   Min.   :0.0000   Min.   :0.000
## 1st Qu.: 42.75   1st Qu.:22488458   1st Qu.:1.0000   1st Qu.:0.000
## Median : 84.50   Median :46368913   Median :1.0000   Median :1.500
## Mean   : 84.50   Mean   :46891636   Mean   :0.7619   Mean   :1.518
## 3rd Qu.:126.25   3rd Qu.:69305889   3rd Qu.:1.0000   3rd Qu.:3.000
## Max.   :168.00   Max.   :99501779   Max.   :1.0000   Max.   :4.000
##          Ex2          Ex3          Nota
## Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
## 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000
## Median :0.0000   Median :0.5000   Median : 2.625
## Mean   :0.6161   Mean   :0.8438   Mean   : 2.978
## 3rd Qu.:0.2500   3rd Qu.:1.5000   3rd Qu.: 5.000
## Max.   :3.0000   Max.   :3.0000   Max.   :10.000
```

Se observa que todas las variables del archivo están consideradas como numéricas.

2. Clasificar estadísticamente las variables del archivo.

- *DNI*: Variable cualitativa nominal.
- *Asistencia*: Variable cualitativa nominal.
- *Ex1*: Variable cuantitativa continua.
- *Ex2*: Variable cuantitativa continua.
- *Ex3*: Variable cuantitativa continua.
- *Nota*: Variable cuantitativa continua.

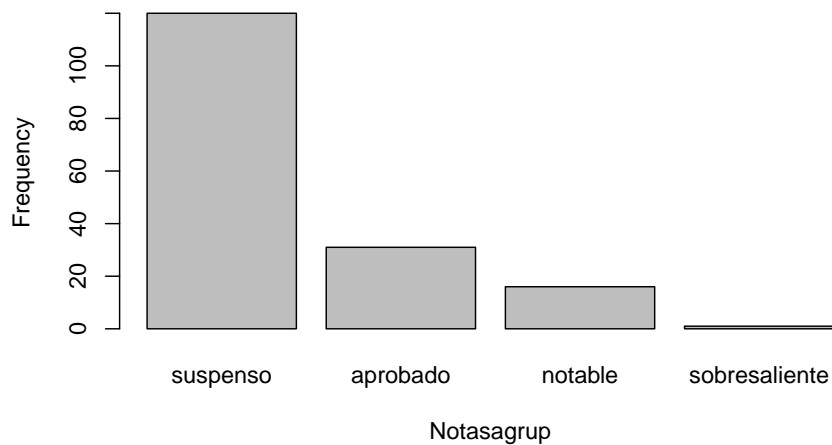
3. Agrupa la variable *Nota* en las categorías Suspenso, Aprobado, Notable y Sobresaliente. Obten la distribución de frecuencias completa de la variable agrupada. Da una representación gráfica de la misma.

NOTA: Primero recodificamos y después reasignamos el orden dado por defecto (lexicográfico). Menús *Recodificar* y *Estadística Básica/Datos/Modificar .../Reordenar niveles de factor*.

```
notas$Notasagrup<- recode(notas$Nota,"10:4.999='suspenso';5:6.999='aprobado';7:8.999='
notas$Notasagrup=as.factor(notas$Notasagrup)
notas$Notasagrup <- with(notas, factor(Notasagrup, levels=c('suspenso',
'aprobado','notable','sobresaliente'), ordered=TRUE))
calcular_frecuencia(df.nominal=NULL, ordenado.frec=FALSE, df.ordinal=notas["Notasagrup",])

##
## -----
##
## Variables ordinales:
```

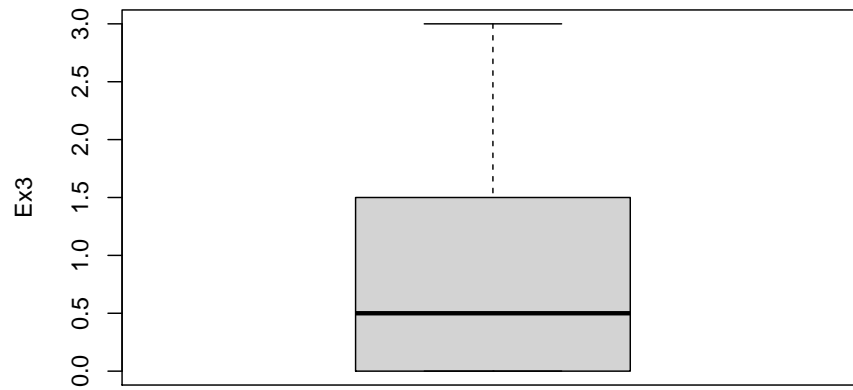
```
##
## Variable: Notasagrup
##      ni      fi  Ni   Fi
## suspenso    120 0.71429 120 0.714
## aprobado    31 0.18452 151 0.899
## notable     16 0.09524 167 0.994
## sobresaliente 1 0.00595 168 1.000
## N= 168
##
## Cuantil: 0.5
##      Variable      Fi
## Notasagrup suspenso 0.7142857
with(notas, Barplot(notas$Notasagrup, xlab="Notasagrup", ylab="Frequency"))
```



4. Resume numéricamente la variable *Ex3*. Genera el boxplot y obtén el valor numérico central de la caja.

```
## Warning in cv(X): not all values are positive
```

```
##      mean      sd IQR      cv skewness kurtosis 0% 25% 50% 75% 100%  n
## 0.84375 0.9060209 1.5 1.073803 0.5567103 -1.002481 0 0 0.5 1.5 3 168
```



El valor numérico central de la caja es el valor de la mediana o cuantil 0.5.

- Resume numéricamente la variable Nota agrupada.

```
##
## -----
##
## Variables ordinales:
##
## Variable: Notasagrup
##      ni      fi  Ni   Fi
## suspenso    120 0.71429 120 0.714
## aprobado    31 0.18452 151 0.899
## notable     16 0.09524 167 0.994
## sobresaliente 1 0.00595 168 1.000
## N= 168
##
## Cuantil: 0.5
##      Variable      Fi
## Notasagrup suspenso 0.7142857
```

Se observa que más del 71% de los alumnos han suspendido.

```
##
## -----
##
## Resumen numérico ponderado por: ni
```

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd      IQR 0%    25%   50%    75% 100%   n
## 3.704911 1.989337 0.6764516 0 1.7465 3.493 5.377097 10 168
##
## -----
##
## Distribución de frecuencia para la variable tabulada:
## Li_1  Li   xi  ni      fi  Ni      Fi  ai      hi
## 0.00  4.99 2.495 120 0.714285714 120 0.7142857 4.99 0.143143430
## 4.99  6.99 5.990 31 0.184523810 151 0.8988095 2.00 0.092261905
## 6.99  8.99 7.990 16 0.095238095 167 0.9940476 2.00 0.047619048
## 8.99 10.00 9.495  1 0.005952381 168 1.0000000 1.01 0.005893446
```

5. En media y para cada grupo de la v. *Nota*, ¿qué ejercicio (de los 3 posibles) fue el mejor?

```
##              Ex1      Ex2      Ex3 Ex1:n Ex2:n Ex3:n
## suspenso      0.8541667 0.1791667 0.5604167 120 120 120
## aprobado      2.9919355 1.1129032 1.5564516 31 31 31
## notable        3.4843750 2.7812500 1.4531250 16 16 16
## sobresaliente 4.0000000 3.0000000 3.0000000 1 1 1
```

Por ejemplo, para el grupo de los aprobados el ejercicio que mejor resultado obtenido (en media) por los alumnos fue el primero

- Si el profesor quiere aprobar el 50 % de los alumnos presentados, ¿cuál es la nota de corte?

```
##      0%    25%    50%    75%   100%
## 0.000 0.000 2.625 5.000 10.000
```

La nota de corte debe ser la mediana: 2.625

Ejemplo de estadísticos de forma.

- Con el conjunto de datos *iris*, obtén el coeficiente de simetría de Fisher y el coef. de apuntamiento de la v. *Sepal.Length*. Marca en la pestaña de *Estadísticos* las opciones indicadas en el menú de *Estadística Básica/Estadística descriptiva/Resúmenes numéricos...*

```
data(iris)
numSummary(iris[, "Sepal.Length"], statistics=c("mean", "sd", "quantiles", "skewness",
"skurtosis"), quantiles=c(0,0.25,0.5,0.75,1), type="2")
```

```
##      mean      sd skewness kurtosis 0% 25% 50% 75% 100%   n
## 5.843333 0.8280661 0.314911 -0.552064 4.3 5.1 5.8 6.4 7.9 150
```

- Añade un valor para que el coeficiente de simetría sea negativo. Nuevamente, añade otro valor hasta que el valor del coeficientes sea próximo a 1.

Chapter 4

Ejercicios propuestos de Estadística descriptiva unidimensional para realizar con R-Commander

A continuación aparecen diversos ejercicios resueltos sin el código que los genera.

Ejercicio 4.1. 1. Abre el archivo *cars93* del plugin y calcula la tabla de frecuencias de la variable *Cylinders*

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd IQR 0% 25% 50% 75% 100%  n NA
##  4.967391 1.304692   2  3   4   4   6   8 92  1
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: Cylinders
##   ni    fi Ni    Fi
##  3   3 0.0326  3 0.0326
##  4  49 0.5326 52 0.5652
##  5   2 0.0217 54 0.5870
##  6  31 0.3370 85 0.9239
##  8   7 0.0761 92 1.0000
```

```
## N= 92
```

2. Calcular la tabla de frecuencias de la variable *Cylinders* para los coches que tengan *Airbag* tanto para conductor como pasajeros. Para ello es necesario filtrar el archivo en *Estadística Básica/Datos/Conjunto de datos Activo/Filtrar el Conjunto de datos activo* utilizando la expresión *Airbags=="driver&passenger"*. Es recomendable cambiar el nombre al nuevo conjunto de datos filtrado.

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd  IQR 0%   25% 50% 75% 100%  n
##   5.8125 1.167262 0.25  4  5.75  6   6   8 16
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: Cylinders
##      ni      fi Ni      Fi
##  4   3 0.1875  3 0.188
##  5   1 0.0625  4 0.250
##  6  10 0.6250 14 0.875
##  8   2 0.1250 16 1.000
## N= 16
```

3. Calcular la tabla de frecuencias de la variable *MaxPrice* para los coches que tengan más de dos cilindros. La tabla de frecuencia debe de tener los siguientes intervalos:[0,25)[25,40) y [40,100].

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd  IQR 0%   25% 50% 75% 100%  n
##  21.7837 11.03447 10.425 7.9 14.575 19.55 25   80 92
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: MaxPrice
##      ni      fi Ni      Fi
## [0,25)  69 0.7500 69 0.750
## [25,40)  18 0.1957 87 0.946
## [40,100]  5 0.0543 92 1.000
```

N= 92

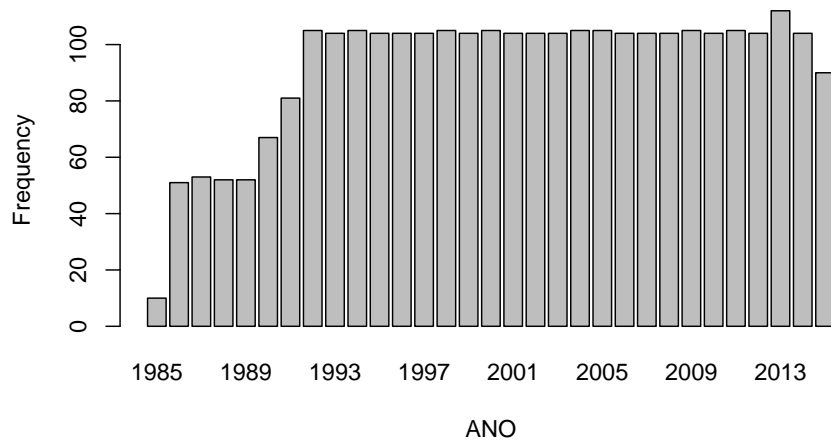
Ejercicio 4.2. El archivo de datos *hist.txt*. Contiene información sobre los resultados del juego semanal de Primitiva. Concretamente los resultados ordenados

1. Que tipo de variables estadísticas están en el archivo.
 - ANO -> numérica - discreta (también puede ser considerada v. cualitativa - ordinal)
 - FECHA -> neste formato: cualitativa ordinal
 - SEMANA -> considerámola ordinal
 - DIA -> nominal
 - JUEGO -> nominal
 - N1,...,N6, C, R -> v. numéricas discretas
 - JOKER -> v. discreta
2. Da la distribución de frecuencias completa de la variable *ANO*. Representala Gráficamente.

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd IQR   0%  25%  50%  75% 100%    n
## 2001.569 8.144567  14 1985 1995 2002 2009 2015 2864
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: ANO
##      ni      fi   Ni      Fi
## 1985  10 0.00349   10 0.00349
## 1986  51 0.01781   61 0.02130
## 1987  53 0.01851  114 0.03980
## 1988  52 0.01816  166 0.05796
## 1989  52 0.01816  218 0.07612
## 1990  67 0.02339  285 0.09951
## 1991  81 0.02828  366 0.12779
## 1992 105 0.03666  471 0.16446
## 1993 104 0.03631  575 0.20077
## 1994 105 0.03666  680 0.23743
## 1995 104 0.03631  784 0.27374
## 1996 104 0.03631  888 0.31006
## 1997 104 0.03631  992 0.34637
## 1998 105 0.03666 1097 0.38303
## 1999 104 0.03631 1201 0.41934
## 2000 105 0.03666 1306 0.45601
```

```
## 2001 104 0.03631 1410 0.49232
## 2002 104 0.03631 1514 0.52863
## 2003 104 0.03631 1618 0.56494
## 2004 105 0.03666 1723 0.60161
## 2005 105 0.03666 1828 0.63827
## 2006 104 0.03631 1932 0.67458
## 2007 104 0.03631 2036 0.71089
## 2008 104 0.03631 2140 0.74721
## 2009 105 0.03666 2245 0.78387
## 2010 104 0.03631 2349 0.82018
## 2011 105 0.03666 2454 0.85684
## 2012 104 0.03631 2558 0.89316
## 2013 112 0.03911 2670 0.93226
## 2014 104 0.03631 2774 0.96858
## 2015 90 0.03142 2864 1.00000
## N= 2864
```

Para calcular un gráfico de barras de la variable ANO, tenemos que transformarla a factor en el menú de *Datos/Modificar variables del conjunto de datos activo/Convertir variable Numérica en Factor*

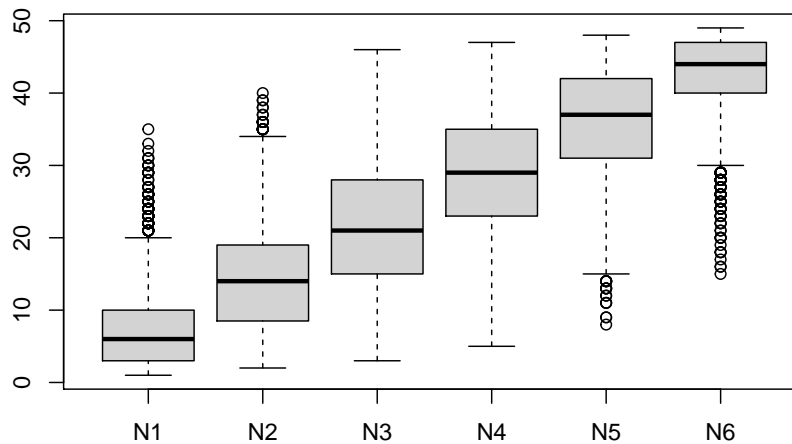


3. Describe numéricamente las variables $N1$, $N2$, \dots , $N6$ (ten en cuenta que las variables están ordenadas).

```
##          mean      sd   IQR  skewness  kurtosis 0%   25% 50% 75% 100%   n
## N1  7.169344 5.826424   7.00  1.3145040  1.68426466  1   3.00   6  10   35 2864
```

```
## N2 14.437849 7.554643 10.25 0.5805783 -0.21978376 2 8.75 14 19 40 2864
## N3 21.582751 8.314512 13.00 0.1677392 -0.57611135 3 15.00 21 28 46 2864
## N4 28.765363 8.264750 12.00 -0.2387364 -0.51575319 5 23.00 29 35 47 2864
## N5 35.816341 7.466572 11.00 -0.6461991 0.02712424 8 31.00 37 42 48 2864
## N6 42.899791 5.666116 7.00 -1.3327754 1.92385771 15 40.00 44 47 49 2864
```

Para realizar un diagrama de caja comparando las 6 variables, R commander por defecto sólo permite mostrar uno de cada vez (usa la función `Boxplot`), por tanto es necesario usar la función `boxplot`. Mira la ayuda de la función, escribe el código en la ventana de sintaxis y ejecuta.



4. ¿Qué porcentaje de veces salió un número menor que 5? ¿Y si consideramos solamente los jueves?

```
## [1] 0.08228585
```

Si consideramos sólo los jueves

```
## [1] 0.08285593
```

5. Basándonos en el archivo que tenemos, si cubrimos un boleto, ¿que números serían?

```
## N
## 38 39 47 3 30 23 45 41 29 5 6 37 22 36 1 48 34 15 14 10
## 387 386 381 375 374 373 371 368 361 360 360 360 359 359 358 358 355 354 353 352
## 11 7 9 35 44 46 17 31 40 42 12 16 21 4 27 2 25 26 13 32
## 352 350 350 350 347 347 346 344 344 344 343 343 343 341 341 340 340 340 339 339
```

```
## 33 19 43 28 18 24 49 8 20
## 339 338 336 334 332 327 321
```

6. ¿Cuál es la distribución de frecuencias completa de la variable estadística *número de bola en el sorteo*? Para hacer la tabla con los menús de R-Commander es necesario pasarla a un dataframe

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd IQR 0% 25% 50% 75% 100%      n
## 25.11191 14.19471 25  1  13  25  38  49 17184
##
## -----
##
## Distribución de frecuencias para variables discretas:
##
## Variable: N
##      ni      fi      Ni      Fi
## 1  358 0.0208    358 0.0208
## 2  340 0.0198    698 0.0406
## 3  375 0.0218   1073 0.0624
## 4  341 0.0198   1414 0.0823
## 5  360 0.0209   1774 0.1032
## 6  360 0.0209   2134 0.1242
## 7  350 0.0204   2484 0.1446
## 8  327 0.0190   2811 0.1636
## 9  350 0.0204   3161 0.1840
## 10 352 0.0205   3513 0.2044
## 11 352 0.0205   3865 0.2249
## 12 343 0.0200   4208 0.2449
## 13 339 0.0197   4547 0.2646
## 14 353 0.0205   4900 0.2851
## 15 354 0.0206   5254 0.3057
## 16 343 0.0200   5597 0.3257
## 17 346 0.0201   5943 0.3458
## 18 334 0.0194   6277 0.3653
## 19 338 0.0197   6615 0.3850
## 20 321 0.0187   6936 0.4036
## 21 343 0.0200   7279 0.4236
## 22 359 0.0209   7638 0.4445
## 23 373 0.0217   8011 0.4662
## 24 332 0.0193   8343 0.4855
## 25 340 0.0198   8683 0.5053
## 26 340 0.0198   9023 0.5251
## 27 341 0.0198   9364 0.5449
```

```
## 28 336 0.0196 9700 0.5645
## 29 361 0.0210 10061 0.5855
## 30 374 0.0218 10435 0.6073
## 31 344 0.0200 10779 0.6273
## 32 339 0.0197 11118 0.6470
## 33 339 0.0197 11457 0.6667
## 34 355 0.0207 11812 0.6874
## 35 350 0.0204 12162 0.7078
## 36 359 0.0209 12521 0.7286
## 37 360 0.0209 12881 0.7496
## 38 387 0.0225 13268 0.7721
## 39 386 0.0225 13654 0.7946
## 40 344 0.0200 13998 0.8146
## 41 368 0.0214 14366 0.8360
## 42 344 0.0200 14710 0.8560
## 43 338 0.0197 15048 0.8757
## 44 347 0.0202 15395 0.8959
## 45 371 0.0216 15766 0.9175
## 46 347 0.0202 16113 0.9377
## 47 381 0.0222 16494 0.9598
## 48 358 0.0208 16852 0.9807
## 49 332 0.0193 17184 1.0000
## N= 17184
```

7. Haz dos archivos, uno con los datos del jueves y otro con los del sábado (*Estadística Básica/Datos/Conjunto de datos Activo/Filtrar el Conjunto de datos activo*). Compara numéricamente las variables $N1$, . . . , $N6$ para el jueves (*Estadística Básica/Estadística Descriptiva/ Resúmenes Numéricos Variables Discretas*).

```
##
## -----
##
## Resúmenes numéricos:
##      mean      sd IQR      cv  skewness  kurtosis  0% 25% 50% 75% 100%
## N1  7.304265 5.980891   8 0.8188218  1.2992647  1.5231002  1   3   5  11  33
## N2 14.500318 7.637352  12 0.5267023  0.5530669 -0.2962520  2   8  14  20  40
## N3 21.690006 8.469880  13 0.3904969  0.2072778 -0.6149106  4  15  21  28  46
## N4 28.832591 8.309344  12 0.2881928 -0.2469576 -0.4934457  5  23  29  35  47
## N5 35.950987 7.334207  10 0.2040057 -0.6667199  0.1598085  9  31  37  41  48
## N6 43.042648 5.426882   7 0.1260815 -1.2503957  1.5419675 15  40  45  47  49
##      n
## N1 1571
## N2 1571
## N3 1571
## N4 1571
## N5 1571
```

N6 1571

Ejercicio 4.3. De una encuesta realizada a 100 clientes de una empresa se han obtenido los siguientes datos sobre sus ingresos por nómina en euros:

Ingresos	Número de nóminas
(600-1000]	15
(1000-1200]	29
(1200-2000]	56

Calcular:

- Ingreso medio de las familias.
- Una entidad bancaria pretende que el 40% central de las nóminas sean domiciliadas en su oficina.

Halla los ingresos de las nóminas mínima y máxima que pretenden ser captadas.

- ¿Qué nómina debe tener un cliente para que sea superior al de la mitad de los encuestados? - ¿Se puede suponer que la distribución de nóminas es simétrica?
- ¿Cuál es la cuantía de nómina más frecuente entre los clientes encuestados?
- Si el coeficiente de variación de Pearson de la nómina los empleados de la empresa es de 1.1, ¿qué valor medio (empleados versus clientes) es el más representativo de su muestra?

##

##

Resumen numérico ponderado por: ni

##

##

Resúmenes numéricos:

##	mean	sd	IQR	cv	skewness	kurtosis	0%	25%	50%
##	1335	313.4884	473.3374	0.2348228	-0.5294634	-1.350087	600	1068.966	1285.714
##	75%	100%	30%	70%	mode	n			
##	1642.857	2000	1103.448	1571.429	1130.233	100			

##

##

Distribución de frecuencia para la variable tabulada:

##	Li_1	Li	xi	ni	fi	Ni	Fi	ai	hi
##	600	1000	800	15	0.15	15	0.15	400	0.000375
##	1000	1200	1100	29	0.29	44	0.44	200	0.001450
##	1200	2000	1600	56	0.56	100	1.00	800	0.000700

Chapter 5

Estadística Descriptiva Bidimensional

```
require(RcmdrMisc)

## Warning: package 'RcmdrMisc' was built under R version 4.0.3
## Warning: package 'car' was built under R version 4.0.3
## Warning: package 'carData' was built under R version 4.0.3
## Warning: package 'sandwich' was built under R version 4.0.3
require(RcmdrPlugin.TeachStat, warn.conflicts=FALSE, quietly=TRUE)

## Warning: package 'RcmdrPlugin.TeachStat' was built under R version 4.0.3
```

5.1 Introducción

Cuando consideramos dos características de una población o muestra para estudiarlas *conjuntamente*, podemos encontrarnos con los siguientes supuestos:

- Variable bidimensional: cuando los dos caracteres son variables estadísticas. Por ejemplo: la Altura y el Peso.
- Atributo bidimensional: cuando los dos caracteres son atributos. Ejemplo: Estado Civil y Religión de una muestra de ciudadanos.
- Distribución bidimensional mixta: si las características son una variable y un atributo. Por ejemplo: Edad y Sexo de una muestra de alumnos.

Una variable bidimensional se suele denotar por: (X, Y) y se denomina *variable conjunta de X e Y* . Se trata de un par ordenado, donde X , Y son las dos variables.

Los valores vienen expresados por (x_i, y_j) con $x_i \in X$, $y_j \in Y$ para todo $i = 1, \dots, h$ y $j = 1, \dots, k$.

5.2 Tablas de frecuencias de doble entrada

- Frecuencia absoluta bidimensional del par (x_i, y_j) : n_{ij} es el número de veces que se presenta conjuntamente el par de valores (x_i, y_j) .
- Frecuencia relativa bidimensional del par (x_i, y_j) : $f_{ij} = \frac{n_{ij}}{N}$.

Se llama *distribución de frecuencias absolutas de (X, Y)* , al conjunto de pares de valores (x_i, y_j) , asociado a las frecuencias absolutas n_{ij} de dichos pares: (x_i, y_i, n_{ij}) .

Análogamente la *distribución de frecuencias relativas de (X, Y)* , al conjunto de pares de valores (x_i, y_j) , asociado a las frecuencias relativas f_{ij} de dichos pares: (x_i, y_i, f_{ij}) .

Tablas de correlación y contingencia: Generalmente hay que manejar un número elevado de valores o modalidades en variables bidimensionales o atributos bidimensionales de frecuencia. Para facilitar este trabajo de organizar y resumir los datos, (valores o modalidades), contamos con *tablas de Correlación* para variables bidimensionales de frecuencias y *tablas de Contingencia* para atributos bidimensionales. Dichas tablas son cuadros de doble entrada, como el siguiente:

X/Y	y_1	\dots	y_j	\dots	y_k	Total
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1k}	$n_{1.}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_i	n_{i1}	\dots	n_{ij}	\dots	n_{ik}	$n_{i.}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_h	n_{h1}	\dots	n_{hj}	\dots	n_{hk}	$n_{h.}$
Total	$n_{.1}$	\dots	$n_{.j}$	\dots	$n_{.k}$	N

En la última columna y en la última fila se escriben los totales por columna y fila respectivamente. Por tanto $n_{i.}$ significa el número total de veces que se ha presentado el valor $X = x_i$ con independencia de los valores que tome la variable Y . Análogamente, $n_{.j}$ significa el número total de veces que se presentó el valor $Y = y_j$ con independencia de los valores que toma X .

Se debe cumplir siempre:

- $\sum_{i=1}^h \sum_{j=1}^k n_{ij} = N$ y $\sum_{i=1}^h \sum_{j=1}^k f_{ij} = 1$.
- $n_{i.} = \sum_{j=1}^k n_{ij}$ y $n_{.j} = \sum_{i=1}^h n_{ij}$

$$\bullet \quad f_{i.} = \frac{n_{i.}}{N} = \sum_{j=1}^k \frac{n_{ij}}{N} = \sum_{j=1}^k f_{ij} \text{ y } f_{.j} = \frac{n_{.j}}{N} = \sum_{i=1}^h \frac{n_{ij}}{N} = \sum_{i=1}^h f_{ij}$$

Los distintos valores x_i pueden aparecer agrupados en intervalos del tipo $[L_{i-1}, L_i)$ y los valores y_j en intervalos $[L_{j-1}, L_j)$ como ya ocurría en las distribuciones unidimensionales.

Por otra parte, en la *tabla de contingencia*, podemos aplicar lo dicho para la *tabla de correlación* en lo referente a fórmulas con frecuencias absolutas y frecuencias relativas.

5.3 Distribuciones marginales

La distribución marginal de la variable X , en una distribución de una variable bidimensional de frecuencias (X, Y) , viene definida por los valores que toma dicha variable y las frecuencias de los mismos, independientemente de los valores que tome la otra variable Y .

Así las distribuciones marginales absolutas y relativas de la variable X , son :

$$\{(x_i, n_{i.}) \mid i = 1, \dots, h\} \text{ y } \{(x_i, f_{i.}) \mid i = 1, \dots, h\}$$

donde

$$n_{i.} = \sum_{j=1}^k n_{ij} \text{ y } f_{i.} = \sum_{j=1}^k f_{ij} = \frac{n_{i.}}{N}$$

Verificandose que $\sum_{i=1}^h n_{i.} = N$ y $\sum_{i=1}^h f_{i.} = 1$.

A través de estas *distribuciones marginales*, que son *distribuciones unidimensionales de frecuencias*, podemos determinar las medidas habituales estudiadas anteriormente, como la media aritmética, moda, mediana, etc. Por ejemplo la media de la variable X se calcularía como sigue:

$$\bar{X} = \sum_{i=1}^h \frac{x_i n_{i.}}{N}$$

Las distribuciones marginales de Y son:

$$\{(y_j, n_{.j}) \mid j = 1, \dots, k\} \text{ y } \{(y_j, f_{.j}) \mid j = 1, \dots, k\}$$

donde

$$n_{.j} = \sum_{i=1}^h n_{ij} \text{ y } f_{.j} = \sum_{i=1}^h f_{ij} = \frac{n_{.j}}{N}$$

Verificándose que $\sum_{j=1}^k n_{.j} = N$ y $\sum_{j=1}^k f_{.j} = 1$.

5.4 Distribuciones condicionadas

En las distribuciones bidimensionales (X, Y) , nos puede interesar estudiar una variable condicionada a que la otra variable tome un determinado valor o valores (a veces intervalo). Tenemos así la *distribución de la variable Y dado un valor $X = x_i$* , simbolizada por la expresión: $Y/X = x_i$.

También podríamos escribir $X/Y = y_j$ cuando la distribución que se quiere determinar es la de la variable X , condicionada por un valor de Y igual a y_j . Las tablas son las siguientes:

$X/Y = y_j$	n_i/y_j
x_1	n_{1j}
\vdots	\vdots
x_h	n_{hj}
	$n_{.j}$

ddddasdfsdfasdf

$Y/X = x_i$	n_j/x_i
y_1	n_{i1}
\vdots	\vdots
y_k	n_{ik}
	$n_{i.}$

Para calcular las frecuencias relativas de la variable X condicionada por un valor de Y cualquiera, tenemos la siguiente expresión:

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{n_{ij}/N}{n_{.j}/N} = \frac{f_{ij}}{f_{.j}}$$

Esta igualdad nos lleva a afirmar que : *la frecuencia relativa CONDICIONADA es igual al cociente entre la frecuencia relativa CONJUNTA y la frecuencia relativa MARGINAL, con respecto de quien se considera la condición* }.

Una vez calculadas las distribuciones condicionadas, como éstas son unidimensionales, podemos calcular características de las mismas como media, mediana, moda, etc.

El cálculo de la distribución de Y condicionada a X utilizando valores en intervalos, es idéntico y sigue las reglas seguidas en las distribuciones unidimensionales.

Ejercicio 5.1. Introduce la siguiente tabla y obtén las distribución marginales.

	Fumador	No Fumador
Hombre	10	20
Mujer	30	40

- Tabla de frecuencias relativas:

```
.Table <- matrix(c(10,20,30,40), 2, 2, byrow=TRUE)
dimnames(.Table) <- list("Sexo"=c("H", "M"), "Fumador"=c("F", "N-F"))
.Table # Counts
```

```
##      Fumador
## Sexo  F N-F
##   H 10  20
##   M 30  40
```

```
totPercents(.Table) # Percentage of Total
```

```
##      F N-F Total
## H    10  20    30
## M    30  40    70
## Total 40  60   100
```

```
remove(.Table)
```

- Tablas de marginales:

```
.Table <- matrix(c(10,20,30,40), 2, 2, byrow=TRUE)
dimnames(.Table) <- list("Sexo"=c("H", "M"), "Fumador"=c("F", "N-F"))
.Table # Counts
```

```
##      Fumador
## Sexo  F N-F
##   H 10  20
##   M 30  40
```

```
rowPercents(.Table) # Row Percentages
```

```
##      Fumador
## Sexo  F  N-F Total Count
##   H 33.3 66.7   100    30
##   M 42.9 57.1   100    70
```

```
remove(.Table)
```

```
# Ahora por columnas
```

```
.Table <- matrix(c(10,20,30,40), 2, 2, byrow=TRUE)
```

```
dimnames(.Table) <- list("Sexo"=c("H", "M"), "Fumador"=c("F", "N-F"))
.Table # Counts
```

```
##      Fumador
## Sexo  F N-F
##   H  10  20
##   M  30  40
```

```
colPercents(.Table) # Column Percentages
```

```
##      Fumador
## Sexo      F   N-F
##   H      25  33.3
##   M      75  66.7
## Total 100 100.0
## Count  40  60.0
```

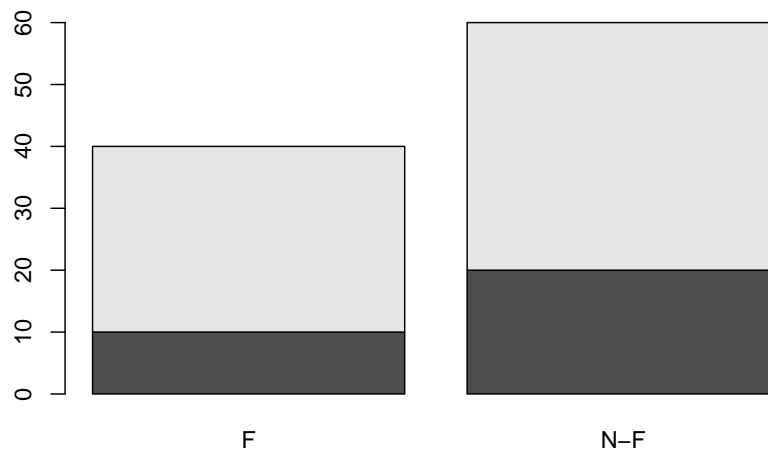
```
remove(.Table)
```

Representación gráfica:

```
.Table <- matrix(c(10,20,30,40), 2, 2, byrow=TRUE)
dimnames(.Table) <- list("Sexo"=c("H", "M"), "Fumador"=c("F", "N-F"))
```

```
barplot(.Table) # representación de frecuencias absolutas
```

```
barplot(100*.Table/sum(.Table)) # representación de frecuencias relativas en tanto por
```



para ver más formas de representaciones de variables bidimensionales ver los ejemplos
de la ayuda de barplot

5.5 Momentos de una distribución bidimensional numérica

Los momentos de orden (r, s) de una variable bidimensional (X, Y) se definen como:

Momentos centrados en el origen de orden (r, s) :

$$a_{rs} = \sum_{i=1}^h \sum_{j=1}^k x_i^r y_j^s f_{ij}$$

Casos particulares: $a_{00} = 1$, $a_{10} = \bar{x}$, $a_{01} = \bar{y}$.

Momentos centrados en las medias de orden (r, s) :

$$m_{rs} = \sum_{i=1}^h \sum_{j=1}^k (x_i - \bar{x})^r (y_j - \bar{y})^s f_{ij}$$

De esta expresión se deduce muy fácilmente que $m_{10} = 0$, $m_{01} = 0$, $m_{20} = s_X^2$ (varianza de X) y $m_{02} = s_Y^2$ (varianza de Y).

El momento con respecto a la media más interesante de una distribución bidimensional es la *Covarianza*. Este momento es el que resulta de tomar $r = s = 1$.

$$s_{XY} = m_{11} = \sum_{i=1}^h \sum_{j=1}^k (x_i - \bar{x})(y_j - \bar{y}) f_{ij}$$

Su *signo* nos indica el sentido de la variación *conjunta* de ambas variables. Por ello, si la covarianza es positiva, las dos variables, varían en el mismo sentido, y si es negativa, en sentido opuesto. (El caso de la covarianza igual a 0 será estudiado más adelante).

Propiedades:

- La covarianza admite una expresión más reducida y cómoda para hacer cálculos de modo más rápido.

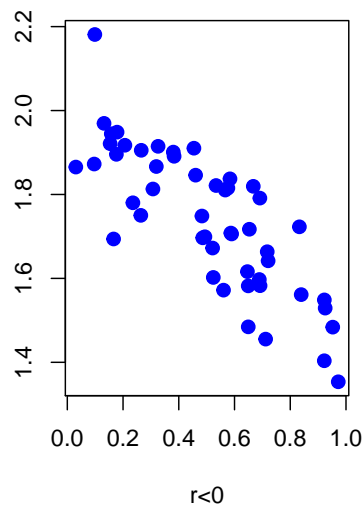
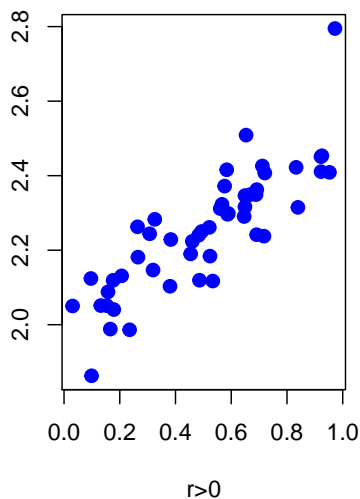
$$m_{11} = \left(\sum_{i=1}^h \sum_{j=1}^k x_i y_j f_{ij} \right) - \bar{x} \cdot \bar{y} = a_{11} - a_{10} a_{01}$$

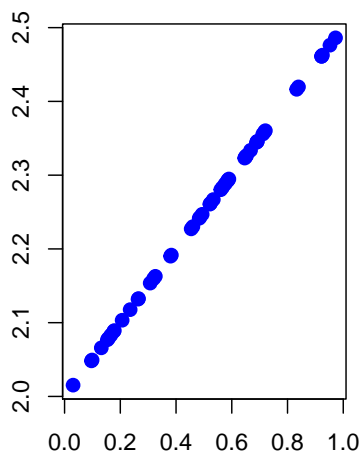
- A la covarianza no le afectan los cambios de origen (las traslaciones), aunque sí depende de los cambios de escala. Sea $x'_i = a + bx_i$ e $y'_j = c + dy_j$ entonces $s_{X',Y'} = bds_{XY}$.
- $s_{XX} = s_X^2$

Coefficiente de Correlación de Pearson: Se define como coeficiente de correlación lineal entre dos variables X e Y :

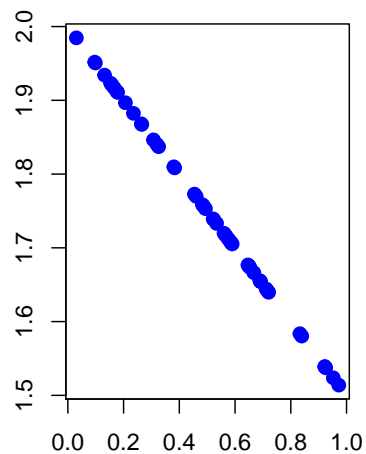
$$r_{XY} = r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$$

Es una *medida adimensional*, es decir, no depende de las unidades de medida de las variables X e Y . Mide el grado de dependencia lineal entre las dos variables. Toma valores en el intervalo $[-1, 1]$. Un valor de r cercano o igual a 0 implica poca o ninguna relación lineal entre X e Y (si $r = 0$ se dice que ambas variables están incorreladas), mientras que cuanto más se acerque a 1 o a -1, más fuerte será la relación lineal entre X e Y , directa o inversa respectivamente.

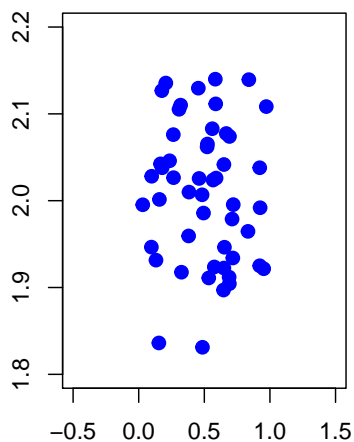




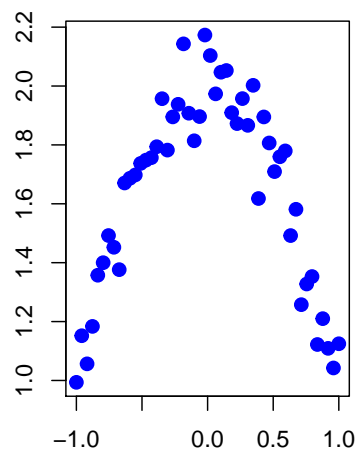
$r=1$, relación positiva perfecta.



$r=-1$ relación negativa perfecta.



r próximo a 0. Poca o ninguna relación.



r próximo a 0, pero con relación cuadrática

NOTA: Una correlación elevada no implica *causalidad*. Si se observa un valor de r próximo a 1 o a -1, no es correcto llegar a la conclusión de que un cambio en X causa un cambio en Y . La única conclusión válida es que puede existir una tendencia lineal entre X e Y

5.6 Independencia Estadística

Dos variables X, Y se dice que son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir:

$$\frac{n_{ij}}{N} = \frac{n_{i.}}{N} \cdot \frac{n_{.j}}{N} \quad \forall ij$$

En este caso las frecuencias relativas condicionadas serán:

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}n_{.j}/N}{n_{.j}} = \frac{n_{i.}}{N} = f_{i.} \quad \forall i$$

Es decir, las frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias relativas marginales, lo que nos indica que el condicionamiento, en cuanto tal, no existe: *las variables son independientes*, puesto que en las distribuciones marginales se estudia el comportamiento de una variable con independencia de los valores que pueda tomar la otra.

Hay relación entre la covarianza y la independencia estadística de las variables X, Y . *Si las variables X e Y son independientes, su covarianza es cero.* Hay que tener en cuenta que el recíproco no siempre es cierto, es decir, el hecho de que la covarianza sea nula no implica necesariamente que las variables sean independientes.

Dada una variable estadística bidimensional, sus características numéricas más importantes pueden expresarse en forma de:

Vector de medias	Matriz de varianzas y covarianzas
$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$	$\Sigma = \begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix}$

Chapter 6

Ejercicios resueltos con R de Análisis Descriptivo Bidimensional

Ejemplo 6.1. Abre el archivo de datos *CPU.txt* teniendo en cuenta que tiene los nombres de las variables en la primera fila, el separador de variables son las comas y el carácter decimal está definido por puntos.

Este archivo contiene variables medidas en una serie de ordenadores (véase *informacioncpu.tx* para más información).

Realiza un resumen numérico e indica qué variables son numéricas y cuales de tipo factor.

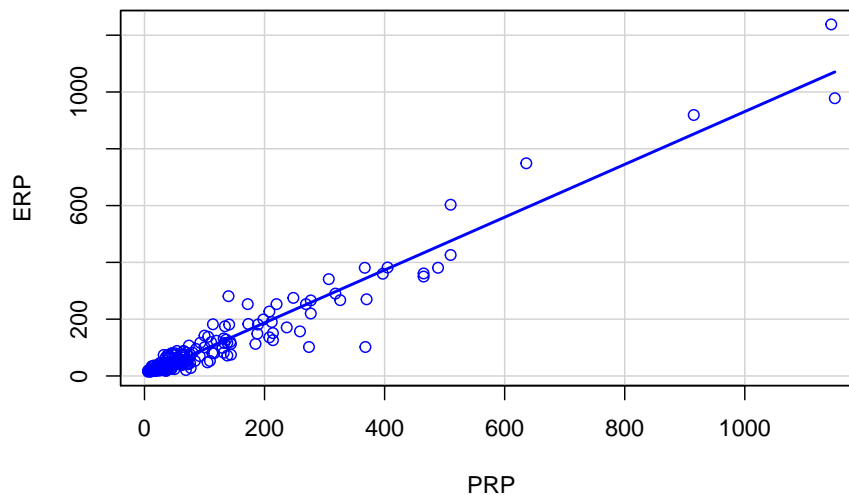
```
##      vendor              Model              MYCT              MMIN
## Length:209          Length:209          Min.   : 17.0   Min.   : 64
## Class :character    Class :character    1st Qu.: 50.0   1st Qu.: 768
## Mode  :character    Mode  :character    Median : 110.0  Median : 2000
##                                     Mean   : 203.8   Mean   : 2868
##                                     3rd Qu.: 225.0  3rd Qu.: 4000
##                                     Max.    :1500.0  Max.    :32000
##      MMAX              CACH              CHMIN              CHMAX
## Min.   : 64   Min.   : 0.00   Min.   : 0.000   Min.   : 0.00
## 1st Qu.: 4000 1st Qu.: 0.00   1st Qu.: 1.000   1st Qu.: 5.00
## Median : 8000 Median : 8.00   Median : 2.000   Median : 8.00
## Mean   :11796 Mean   : 25.21   Mean   : 4.699   Mean   : 18.27
## 3rd Qu.:16000 3rd Qu.: 32.00   3rd Qu.: 6.000   3rd Qu.: 24.00
## Max.    :64000 Max.    :256.00   Max.    :52.000   Max.    :176.00
##      PRP              ERP              ERP_agrup              CACH_agrup
## Min.   : 6.0   Min.   : 15.00   Length:209      Length:209
```

```
## 1st Qu.: 27.0 1st Qu.: 28.00 Class :character Class :character
## Median : 50.0 Median : 45.00 Mode :character Mode :character
## Mean : 105.6 Mean : 99.33
## 3rd Qu.: 113.0 3rd Qu.: 101.00
## Max. :1150.0 Max. :1238.00
```

- a) La variable *ERP* indica la eficiencia relativa estimada de cada ordenador. Dibuja un gráfico de dispersión de la Eficiencia (*ERP*) frente a published relative performance (*PRP*). Utiliza el menú de *Gráficos/Diagrama de Dispersión*. Observa si Eficiencia aumenta en función de la memoria caché o no.

```
## Warning in plot.window(...): "reg.line" is not a graphical parameter
## Warning in plot.window(...): "spread" is not a graphical parameter
## Warning in plot.window(...): "span" is not a graphical parameter
## Warning in plot.window(...): "levels" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "reg.line" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "spread" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "span" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "levels" is not a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "reg.line" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "spread" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "span" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "levels" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "reg.line" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "spread" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "span" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "levels" is not a
## graphical parameter
## Warning in box(...): "reg.line" is not a graphical parameter
## Warning in box(...): "spread" is not a graphical parameter
```

```
## Warning in box(...): "span" is not a graphical parameter
## Warning in box(...): "levels" is not a graphical parameter
## Warning in title(...): "reg.line" is not a graphical parameter
## Warning in title(...): "spread" is not a graphical parameter
## Warning in title(...): "span" is not a graphical parameter
## Warning in title(...): "levels" is not a graphical parameter
```



2) Calcula la covarianza y el coeficiente de correlación entre las dos variables. Las funciones de R `cov` y `cor`, calculan la covarianza y el coeficiente de correlación entre un par de variables.

```
cov(CPU$ERP,CPU$PRP)
```

```
## [1] 24055.19
```

```
cor(CPU$ERP,CPU$PRP)
```

```
## [1] 0.9664717
```

En el Menú de RCommander, *Estadísticos/Resúmenes/Matriz de Correlaciones* se calcula la matriz de correlaciones de un conjunto de variables. La matriz de correlaciones contiene en cada elemento i, j la correlación existente entre la variable X_i y X_j

```
##          PRP          ERP
## PRP 1.0000000 0.9664717
```

```
## ERP 0.9664717 1.0000000
```

Obsérvese que la correlación de una variable consigo misma es siempre 1.

Ejemplo 6.2. En el mismo archivo del ejemplo anterior, existen dos variables factor *ERP_codif* y *CACH_codif* que categorizan la Eficiencia y Memoria Caché en 3 y 5 categorías, respectivamente.

- a) Calcula la tabla de frecuencias absoluta de la variable bidimensional (*CACH_agrup*, *ERP_agrup*) (*Estadística Básica/Estadística Descriptiva/Tabla de doble entrada*)

```
## Warning: package 'abind' was built under R version 4.0.3
```

```
##
## Frequency table:
##           ERP_agrup
## CACH_agrup ALTA BAJA MEDIA
## (-Inf,12]   6  70   46
## (12,32]    26   1   21
## (160,256]   2   0    0
## (32,65]    22   0    2
## (65,160]   13   0    0
```

- b) Calcula la tabla de frecuencias relativas de la variable bidimensional (en la pestaña de Estadísticos marcar en Porcentajes totales) (*CACH_agrup*, *ERP_agrup*).

```
##
## Frequency table:
##           CACH_agrup
## ERP_agrup (-Inf,12] (12,32] (160,256] (32,65] (65,160]
## ALTA           6    26         2     22      13
## BAJA           70     1         0      0       0
## MEDIA          46    21         0      2       0
##
## Total percentages:
##           (-Inf,12] (12,32] (160,256] (32,65] (65,160] Total
## ALTA           2.9   12.4         1    10.5     6.2    33
## BAJA           33.5    0.5         0     0.0     0.0    34
## MEDIA          22.0   10.0         0     1.0     0.0    33
## Total          58.4   23.0         1    11.5     6.2   100
```

Al pedir porcentajes totales, por defecto también calcula la tabla de frecuencias absoluta.

El porcentaje de ordenadores con Eficiencia Alta y Memoria caché menor o iguala 32 es $12.4 + 2.9 = 15.3$.

El porcentaje de ordenadores con eficiencia media y memoria entre 12 y 32 es de un 10%.

- c) Calcula la distribución de la memoria Caché agrupada condicionada a que la eficiencia es ALTA. En el mismo menú (*Estadística Básica/Estadística Descriptiva/Tabla de doble entrada*) seleccionar en el botón estadísticos porcentaje por filas o por columnas según el orden seleccionado para las variables.

```
##
## Frequency table:
##      CACH_agrup
## ERP_agrup (-Inf,12] (12,32] (160,256] (32,65] (65,160]
##      ALTA          6      26          2      22      13
##      BAJA         70       1          0       0       0
##      MEDIA        46      21          0       2       0
##
## Row percentages:
##      CACH_agrup
## ERP_agrup (-Inf,12] (12,32] (160,256] (32,65] (65,160] Total Count
##      ALTA          8.7    37.7         2.9    31.9    18.8    100    69
##      BAJA         98.6     1.4         0.0     0.0     0.0    100    71
##      MEDIA        66.7    30.4         0.0     2.9     0.0    100    69
```

Aquí podemos ver la distribución de CACH_agrup condicionada a cada uno de los valores de ERP_agrup. Dado el enunciado, nos interesa la fila correspondiente a ERP_agrup=ALTA. Nótese que para esa distribución el total de ordenadores es de 69 y la suma de porcentajes de esa fila es 100.

Análogamente se podrían alcular los porcentajes por columnas, sin más que seleccionar en el menu de Estadísticos *Porcentajes por columnas*:

```
##
## Frequency table:
##      CACH_agrup
## ERP_agrup (-Inf,12] (12,32] (160,256] (32,65] (65,160]
##      ALTA          6      26          2      22      13
##      BAJA         70       1          0       0       0
##      MEDIA        46      21          0       2       0
##
## Column percentages:
##      CACH_agrup
## ERP_agrup (-Inf,12] (12,32] (160,256] (32,65] (65,160]
##      ALTA          4.9    54.2        100    91.7        100
##      BAJA         57.4     2.1         0     0.0         0
##      MEDIA        37.7    43.8         0     8.3         0
##      Total       100.0   100.1        100   100.0        100
##      Count       122.0    48.0         2    24.0         13
```

Ejemplo 6.3. La cotización en Bolsa de dos empresas A,B durante la última semana son las siguientes

Empresa	Lunes	Martes	Miércoles	Jueves	Viernes
A	8	7	5	7	8
B	6	4.5	4	4.5	5

Calcula el coeficiente de correlación de Pearson entre las cotizaciones de la empresa A y de la empresa B e interpreta el resultado. Calcula también el valor de la covarianza.

En primer lugar debemos introducir los datos en un conjunto nuevo de datos, recordar que cada empresa corresponde a una variable.

```
cor(bolsa[,c("A", "B")], use="complete")
```

```
##           A           B
## A 1.0000000 0.8075729
## B 0.8075729 1.0000000
```

El valor del coeficiente de correlaciones 0.8075729 lo que indica una correlación lineal positiva, ambas empresas cotizan al alza o a la baja simultáneamente.

A continuación vamos a calcular la covarianza. Dado que el coeficiente de correlación es el cociente entre la covarianza y el producto de las dos desviaciones típicas, la covarianza podemos calcularla del siguiente modo:

```
numSummary(bolsa[,c("A", "B")], statistics=c("sd"), quantiles=c(0,0.25,0.5,0.75,1))
```

```
##           sd n
## A 1.2247449 5
## B 0.7582875 5
```

El valor de la covarianza se obtiene de:

```
0.8075729*1.2247449*0.7582875
```

```
## [1] 0.75
```

Ejemplo 6.4. Abre el archivo cars93 del paquete *RcmdrPlugin.InferenciaEstadistica*.

- a) Calcula la tabla de frecuencias relativa de la variable bidimensional (*Airbags*, *USA*). ¿Cuál es el porcentaje de coches del archivo que son Americanos y con Airbag sólo para el pasajero?

```
##
## Frequency table:
##           USA
## Airbags    nonUS US
## none              18 16
## driver              20 23
## driver&passenger    7  9
##
```



```
## Total percentages:
##           nonUS   US Total
## none           19.4 17.2 36.6
## driver          21.5 24.7 46.2
## driver&passenger 7.5 9.7 17.2
## Total          48.4 51.6 100.0
```

- b) Dentro de los coches americanos, ¿cuál es el porcentaje de coches con cambio manual (Variable *Manual*)?

```
##
## Frequency table:
##           USA
## Manual nonUS US
##   No         6 26
##   Yes        39 22
##
## Column percentages:
##           USA
## Manual nonUS   US
##   No        13.3 54.2
##   Yes        86.7 45.8
##   Total    100.0 100.0
##   Count     45.0 48.0
```

La solución es: 45.8%

- c) Hemos considerado sólo los coches con cambio *Manual*, qué porcentaje de los mismos no es americano

```
##
## Frequency table:
##           USA
## Manual nonUS US
##   No         6 26
##   Yes        39 22
##
## Row percentages:
##           USA
## Manual nonUS   US Total Count
##   No        18.8 81.2   100    32
##   Yes        63.9 36.1   100    61
```

La solución es: 63.9%

- d) Analizar si existe algún tipo de correlación entre las variables (*MinPrice* y *MaxPrice*).

```
##           MaxPrice MinPrice
## MaxPrice 1.0000000 0.9067561
```

```
## MinPrice 0.9067561 1.0000000
```

e) Calcula el valor de la covarianza entre las variables *MinPrice* y *MaxPrice*.

```
##               mean          sd  IQR  0%  25%  50%  75% 100%  n
## MaxPrice 21.89892 11.030457 10.6  7.9 14.7 19.6 25.3 80.0 93
## MinPrice 17.12581  8.746029  9.5  6.7 10.8 14.7 20.3 45.4 93
## [1] 87.47721

## Warning: package 'Rcmdr' was built under R version 4.0.3
## Warning: package 'RcmdrMisc' was built under R version 4.0.3
## Warning: package 'car' was built under R version 4.0.3
## Warning: package 'carData' was built under R version 4.0.3
## Warning: package 'sandwich' was built under R version 4.0.3
## Warning: package 'effects' was built under R version 4.0.3
Warning: package 'rgl' was built under R version 4.0.3
```