Práctica 2: Estadística descriptiva univariante

Estadística

Grado en Inteligencia Artificial (UDC)

Índice

2. Estadística descriptiva univariante

2.1. Estadística descriptiva

La estadística descriptiva proporciona herramientas que facilitan la extracción e interpretación de la información contenida en un conjunto de datos. Estas herramientas son principalmente representaciones gráficas y medidas (numéricas) que resumen características de los datos.

El conjunto de datos se corresponde con elementos de una población (típicamente una muestra), denominados **individuos** o casos, para los que se observaron determinadas características, denominadas **variables** estadísticas. Las variables pueden ser de varios tipos:

- **Cualitativas (o categóricas)**: Indican un atributo (no numérico). Los valores que toman se denominan modalidades (o categorías). En R se recomienda codificarlas como `factor` (internamente se almacenan como enteros con una etiqueta asociada)
  - **Nominales**: sus valores (modalidades) son simples etiquetas (sexo, marca, modelo…).
  - **Ordinales**: sus valores están ordenados (grado de satisfacción, nivel de estudios, grupo de edad…).
- **Cuantitativas (o numéricas)**: Toman valores numéricos (número de hijos, número de fallos, peso, precio, tiempo de procesamiento).
  - **Discretas**: toman un número finito (o infinito numerable) de valores distintos (típicamente en el conjunto de números naturales).
  - **Continuas**: toman cualquier valor en un intervalo de valores dado (toman valores en el conjunto de números reales).

El procedimiento de análisis empleado **depende del tipo de variables** involucradas (por ejemplo, algunas medidas solo tienen sentido si la variable es como mínimo ordinal y emplearlas con variables nominales es un error).

Al realizar un análisis descriptivo en la práctica se distingue principalmente entre dos tipos de variables, dependiendo del número de valores distintos que tomen. Las variables discretas se suelen tratar como variables

cualitativas (ordinales) si toman pocos valores distintos y como variables continuas en caso contrario.

Hay tres etapas básicas en cualquier análisis estadístico de un conjunto de datos:

1. Obtención y preparación de los datos (transformando otros si es necesario).
2. Selección y ejecución del procedimiento con las opciones adecuadas.
3. Análisis de los resultados obtenidos (que pueden sugerir repetir los pasos anteriores).

2.2. Obtención y preparación de los datos

Como ya se comentó anteriormente, el objeto de R en el que se suelen almacenar los datos es el `data.frame` (ver Sección 2.3 del libro de referencia). Esta estructura de datos es rectangular, las filas se corresponden con observaciones (casos o individuos) y las columnas con variables (características de los individuos)[1].

Hay una gran cantidad de operaciones que pueden ser de interés en la manipulación de datos:

- Operaciones con conjuntos de datos (Sección 4.1):
  - importar
  - exportar
  - combinar
  - reorganizar
  - …
- Operaciones con variables (Sección 4.2.1):
  - crear
  - recodificar (e.g. categorizar)
  - …
- Operaciones con casos (Sección 4.2.2):
  - ordenar
  - filtrar
  - …

En esta práctica se mostrarán algunas de las operaciones más básicas a medida que se van realizando los distintos análisis. Para más detalles ver el Capítulo 4 o el Apéndice B.

Normalmente emplearemos `load()` para cargar conjuntos de datos almacenados en ficheros con el formato por defecto de R (normalmente con extensión *.RData* o *.rda*). Por ejemplo:

```
load("movil.RData")
# View(movil)
str(movil)
```

```
## 'data.frame':    50 obs. of  7 variables:
##  $ sexo    : Factor w/ 2 levels "Hombre","Mujer": 1 1 2 2 1 2 2 2 2 1 ...
##  $ marca   : Factor w/ 3 levels "Apple","Huawei",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ nsatisfa: Factor w/ 3 levels "Bajo","Medio",..: 2 1 2 3 3 3 3 2 2 2 ...
##  $ ncamaras: num  4 4 4 2 1 4 4 4 4 4 ...
##  $ precio  : num  741 640 735 588 844 ...
##  $ bateria : num  30 26 29 23 33 36 26 24 29 28 ...
##  $ peso    : num  151 158 160 161 153 ...
##  - attr(*, "variable.labels")= Named chr [1:7] "Sexo" "Fabricante" "Nivel de satisfacción" "Número d
##   ..- attr(*, "names")= chr [1:7] "sexo" "marca" "nsatisfa" "ncamaras" ...
```

Con la función `str()` podemos ver qué variables están almacenadas como factores y cuáles como vectores numéricos.

[1] Además, se va a asumir que hay independencia entre casos, pero que puede haber dependencia entre variables, al ser características del mismo individuo.

Se pueden importar datos externos en casi cualquier formato a R (aunque puede requerir instalar paquetes adicionales). Lo habitual es utilizar código, aunque en RStudio se puede emplear los submenús en *File* > *Import Dataset* para importar datos (se previsualizará el resultado y al finalizar escribirá el código por nosotros).

Por ejemplo, para cargar un conjunto de datos almacenado en un fichero de texto podemos emplear la función:

```
read.table(file, header = FALSE, sep = "", dec = ".",
           stringsAsFactors = FALSE, ...)
```

También están disponibles otras funciones con valores por defecto de los parámetros adecuados para otras situaciones (las funciones principales asumen el formato anglosajón y por ejemplo el separador de decimales es por defecto `dec = "."`):

```
read.delim(file, header = TRUE, sep = "\t", dec = ".")
read.delim2(file, header = TRUE, sep = "\t", dec = ",")
read.csv2(file, header = TRUE, sep = ";", dec = ",")
```

Por ejemplo:

```
movil <- read.table("movil.txt", header = TRUE)
# movil <- read.csv2(movil, file = "movil.csv")
str(movil)
```

```
## 'data.frame':    50 obs. of  7 variables:
##  $ sexo    : chr  "Hombre" "Hombre" "Mujer" "Mujer" ...
##  $ marca   : chr  "Apple" "Apple" "Apple" "Apple" ...
##  $ nsatisfa: chr  "Medio" "Bajo" "Medio" "Alto" ...
##  $ ncamaras: int  4 4 4 2 1 4 4 4 4 4 ...
##  $ precio  : num  741 640 735 588 844 ...
##  $ bateria : int  30 26 29 23 33 36 26 24 29 28 ...
##  $ peso    : num  151 158 160 161 153 ...
```

Después de cargar los datos puede que sea necesario transformar variables. Por ejemplo, convertir variables en factores:

```
movil$sexo <- as.factor(movil$sexo)
movil$marca <- as.factor(movil$marca)
movil$nsatisfa <- factor(movil$nsatisfa,
    levels=c("Bajo", "Medio", "Alto"), ordered = TRUE)
```

También nos puede interesar crear nuevas variables a partir de las actuales (Sección 4.2.1).

Para filtrar observaciones y seleccionar variables se puede utilizar la función `subset()`.

Para manipular factores (variables cualitativas) pueden resultar de interés las herramientas en el paquete `forcats` de la colección `tidyverse`.

2.2.1. Ejercicio

El conjunto de datos `ecars` almacenado en el archivo *ecars.RData*, contiene información sobre 103 vehículos eléctricos vendidos en Europa (proporcionada por ev-database.org, y también disponible en kaggle).

a) Carga estos datos en R y clasifica las variables en cualitativas o cuantitativas. Puedes acceder a información sobre las variables y la estructura del conjunto de datos con los comandos:

```
as.data.frame(attr(ecars, "variable.labels"))
str(ecars)
```

b) Crea una nueva variable `logprecio` que contenga el precio en escala logarítmica (`log(ecars$precio)`).

c) Almacena el conjunto de datos en el fichero *ecarsb.RData*.

2.3. Estadística descriptiva univariante

En esta práctica nos centraremos en el análisis de una única variable (el primer paso suele ser analizar las variables de forma independiente y posteriormente de forma conjunta, con el objetivo final de entender la relación entre ellas).

Como ya se comentó, en la selección de los métodos de análisis descriptivo se distingue principalmente entre dos tipos de variables, las que toman un número relativamente pequeño de valores distintos (cualitativas o discretas) y las que toman muchos valores distintos (discretas o continuas).

El primer paso en cualquier análisis descriptivo suele ser la generación de gráficos. Adicionalmente se puede completar la información con valores numéricos, medidas descriptivas. No obstante, como estas medidas se emplean en la construcción de los gráficos, en las siguientes secciones se comenzará introduciéndolas.

Por ejemplo, podemos obtener estadísticos descriptivos básicos de las variables del conjunto de datos con la función `summary()`:

```
summary(movil)
```

```
##      sexo       marca     nsatisfa     ncamaras        precio
##  Hombre:24   Apple :12   Bajo :14   Min.   :1.00   Min.   :140.0
##  Mujer :26   Huawei:18   Medio:16   1st Qu.:1.00   1st Qu.:519.8
##              Xiaomi:20   Alto :20   Median :2.00   Median :609.9
##                                     Mean   :2.42   Mean   :600.1
##                                     3rd Qu.:4.00   3rd Qu.:704.7
##                                     Max.   :4.00   Max.   :941.2
##     bateria           peso
##  Min.   : 9.00   Min.   :140.1
##  1st Qu.:16.25   1st Qu.:151.2
##  Median :24.00   Median :157.4
##  Mean   :22.44   Mean   :158.2
##  3rd Qu.:28.00   3rd Qu.:161.1
##  Max.   :36.00   Max.   :204.4
```

2.4. Descripción de variables cualitativas y cuantitativas discretas

2.4.1. Tabla de frecuencias (unidimensional)

La tabla de frecuencias es una tabla donde se presentan las modalidades (o clases) observadas y sus frecuencias de aparición (los valores que toma una variable pueden repetirse).

Supongamos que tenemos una muestra $\{x_1, ..., x_n\}$ de tamaño n de la una variable X y que presenta k modalidades $\{c_1, ..., c_k\}$. La tabla de frecuencias permite resumir la información de X utilizando:

- **Frecuencia absoluta** n_i: representa el número de veces que aparece la modalidad c_i.
- **Frecuencia relativa** f_i: representa la proporción[2] de individuos con la modalidad c_i, $f_i = \frac{n_i}{n}$.

En el caso de variables ordinales o discretas:

- **Frecuencia absoluta acumulada** N_i: es el número de veces que aparece la modalidad c_i o valores anteriores, $N_i = n_1 + n_2 + ... + n_i$.

[2]En estadística se emplean proporciones (en escala de 0 a 1) en las definiciones y al realizar los cálculos, pero para comunicar los resultados lo habitual es emplear porcentajes.

- **Frecuencia relativa acumulada** F_i: es la proporción de veces que aparece la modalidad c_i o valores anteriores (la frecuencia absoluta acumulada dividida por el tamaño muestral), $F_i = f_1 + f_2 + ... + f_i = \frac{N_i}{n}$.

Las frecuencias se suelen presentar en una tabla de frecuencias, que adopta esta forma:

Modalidad	Frec. abs.	Frec. rel.	Frec. abs. acum.	Frec. rel. acum.
c_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_i	n_i	f_i	N_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Podemos obtener la tabla de frecuencias absolutas con la función `table()`. Por ejemplo:

```
frec <- table(movil$marca)
frec
```

```
##
##  Apple Huawei Xiaomi
##     12     18     20
```

A partir de la que podemos calcular las frecuencias relativas (proporciones) o los porcentajes de las categorías:

```
# Frecuencias relativas
frec/sum(frec) # prop.table(frec)
```

```
##
##  Apple Huawei Xiaomi
##   0.24   0.36   0.40
```

```
# Porcentajes
porc <- 100*frec/sum(frec)
porc
```

```
##
##  Apple Huawei Xiaomi
##     24     36     40
```

Es decir, hay 20 móviles de la marca Xiaomi, lo que supone un 40 % de las observaciones (la categoría más frecuente, denominada **moda**). La categoría menos frecuente es Apple con un 24 %.

En el caso de variables nominales no tiene sentido calcular frecuencias acumuladas pero pueden resultar de interés para variables ordinales o discretas. Podemos emplear la función `cumsum()` para calcularlas. Por ejemplo, para la variable `ncamaras` (discreta) obtendríamos:

```
frec <- table(movil$ncamaras) # Frecuencias absolutas
frec
```

```
##
##  1  2  4
## 15 17 18
```

```
cumsum(frec) # Frecuencias absolutas acumuladas
```

```
##  1  2  4
```

```
## 15 32 50
```

```
porc <- 100*frec/sum(frec)    # Porcentajes (equiv. frecuencias relativas)
porc
```

```
##
##  1  2  4
## 30 34 36
```

```
cumsum(porc)  # Porcentajes acumulados (equiv. frecuencias relativas acumuladas)
```

```
##   1   2   4
##  30  64 100
```

Es decir, un 64 % de los móviles tienen dos o menos cámaras.

Si la variable es cualitativa o discreta la tabla de frecuencias absolutas contiene toda la información contenida en la muestra. Si la variable es continua (o discreta que toma muchos valores distintos), puede categorizarse agrupando los datos numéricos en clases, con la consecuente pérdida de información (aunque puede simplificar algunos análisis). Por ejemplo podríamos agrupar la variable `precio` en intervalos con una amplitud de 300 €:

```
movil$preciocat <- cut(movil$precio,
    breaks = seq(0, 1200, len = 5), include.lowest = TRUE)
frec <- table(movil$preciocat)
frec
```

```
##
##       [0,300]     (300,600]     (600,900] (900,1.2e+03]
##             2            23            23             2
```

```
100*frec/sum(frec)
```

```
##
##       [0,300]     (300,600]     (600,900] (900,1.2e+03]
##             4            46            46             4
```

```
cumsum(100*frec/sum(frec))
```

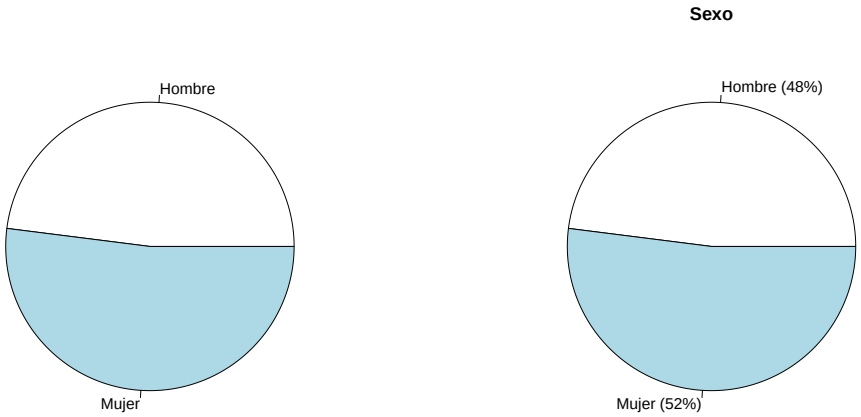
```
##       [0,300]     (300,600]     (600,900] (900,1.2e+03]
##             4            50            96           100
```

2.4.2. Representaciones gráficas

En el caso que la variable este compuesta por sólo unas pocas categorías o valores diferentes, la representación gráfica de la tabla de frecuencia permite acceder de forma más rápida y clara a la información contenida en la variable (forma y distribución, frecuencias más comunes…).

En el caso de variables nominales (con muy pocas categorías) se puede utilizar un **gráfico de sectores**: se divide un círculo en sectores de forma que la amplitud (el ángulo) de cada sector es proporcional a la frecuencia de la modalidad correspondiente. Podemos generar este gráfico en R con la función `pie()`. Por ejemplo:

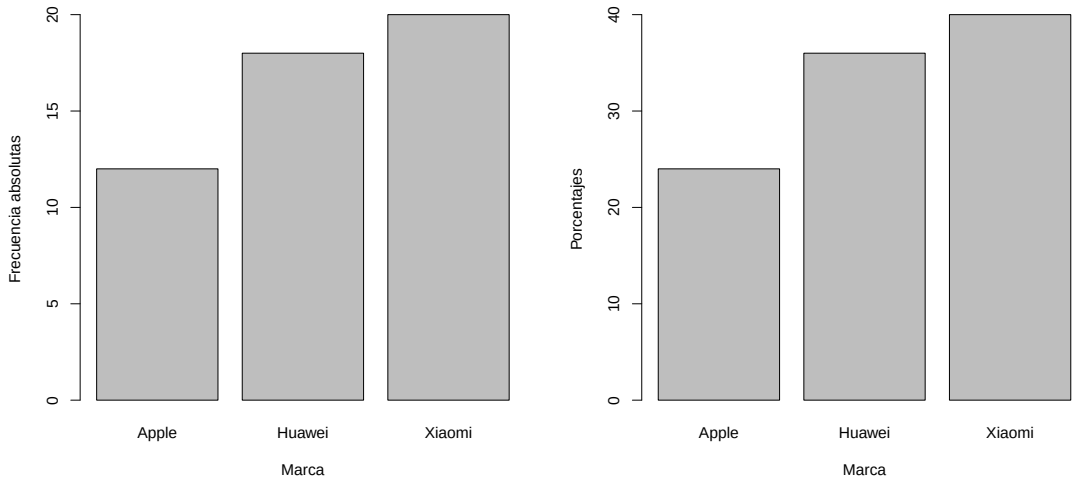
```
old.par <- par(mfrow = c(1,2))
frec <- table(movil$sexo)
pie(frec)
labels <- paste0(names(frec), " (", 100*frec/sum(frec), "%)")
pie(frec, main = "Sexo", labels = labels)
```




```
par(old.par)
```

En general se suele emplear un **gráfico de barras**: en el eje X (abscisas) se representan las modalidades c_i y en el eje Y (ordenadas) las frecuencias absolutas n_i (o relativas f_i), y se dibujan barras verticales con altura igual a la frecuencia considerada.

```
old.par <- par(mfrow = c(1,2))
frec <- table(movil$marca)
barplot(frec, ylab = "Frecuencia absolutas", xlab = "Marca")
porc <- 100*frec/sum(frec)    # Porcentajes (equiv. frecuencias relativas)
barplot(porc, ylab = "Porcentajes", xlab = "Marca")
```

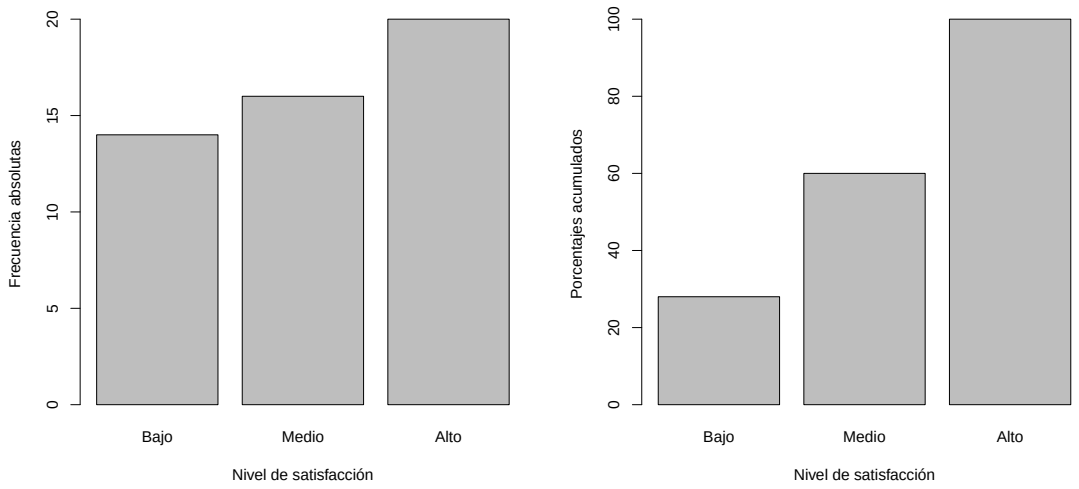



```
par(old.par)
```

En el caso de variables ordinales o discretas también se pueden representar las frecuencias acumuladas, en lo que se conoce como **diagrama de frecuencias acumuladas**.

```
old.par <- par(mfrow = c(1,2))
frec <- table(movil$nsatisfa)
```

```r
barplot(frec, ylab = "Frecuencia absolutas", xlab = "Nivel de satisfacción")
porc <- 100*frec/sum(frec)    # Porcentajes (equiv. frecuencias relativas)
barplot(cumsum(porc), ylab = "Porcentajes acumulados",
        xlab = "Nivel de satisfacción")
```




```r
par(old.par)
```

2.4.3. Ejercicio

Continuando con los datos de vehículos eléctricos `ecars` del ejercicio anterior:

a) Emplear la función `summary()` para obtener descriptivos simples de las variables en el conjunto de datos.

b) Realiza un análisis descriptivo de las variables `segmento`, `cargarapida` y `traccion` (tabla de frecuencias y gráfico de barras o sectores según consideres).

c) Realiza un análisis descriptivo de la variable `asientos`, incluyendo porcentajes acumulados.

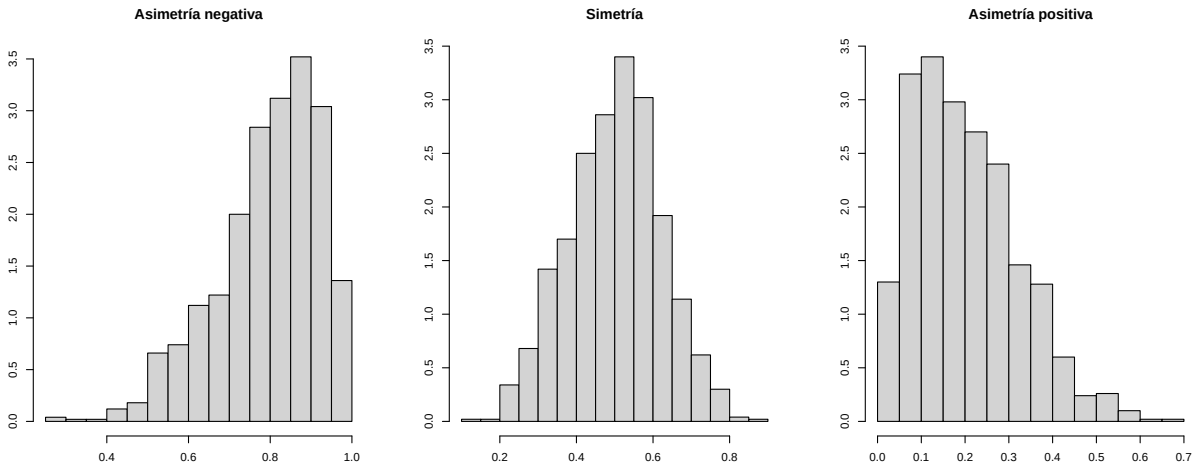
2.5. Descripción de variables continuas

En esta sección se estudia el tratamiento de variables continuas o discretas que toman muchos valores.

2.5.1. Histogramas

La representación gráfica más usual para este tipo de variables es el histograma. Su construcción es similar a la de un gráfico de barras empleando la variable categorizada en intervalos, pero mostrando en el eje x los valores de forma continua (los intervalos están pegados). El área de los rectángulos es proporcional a la frecuencia, normalmente igual a la frecuencia absoluta si los intervalos están equiespaciados. También se suele establecer de forma que el área coincida con la frecuencia relativa (como veremos más adelante, esto se corresponderá con la denominada *densidad*), especialmente recomendable en el caso de intervalos de distinta longitud.

De un histograma podemos obtener mucha información como la forma de la distribución o la detectar la presencia de datos atípicos.



En R podemos generar este gráfico con la función `hist()`:

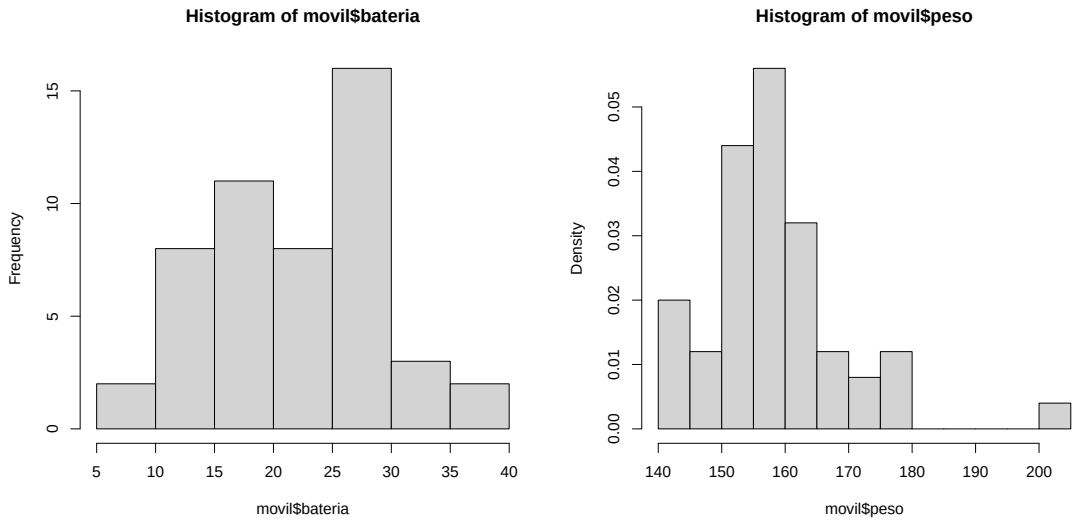
```
hist(x, breaks = "Sturges", freq = NULL, ...)
```

Por ejemplo:

```
old.par <- par(mfrow = c(1,2))
hist(movil$bateria)
res <- hist(movil$peso, breaks = "FD", plot = FALSE)
str(res)
```

```
## List of 6
##  $ breaks  : int [1:14] 140 145 150 155 160 165 170 175 180 185 ...
##  $ counts  : int [1:13] 5 3 11 14 8 3 2 3 0 0 ...
##  $ density : num [1:13] 0.02 0.012 0.044 0.056 0.032 0.012 0.008 0.012 0 0 ...
##  $ mids    : num [1:13] 142 148 152 158 162 ...
##  $ xname   : chr "movil$peso"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"
```

```
plot(res, freq = FALSE, col = "lightgray")
```



```
par(old.par)
```

Esta función devuelve (de forma invisible) una lista con los valores correspondientes a los distintos elementos del gráfico (una forma de obtener tablas de frecuencias asociadas a variables continuas; si no se desea generar el gráfico se puede establecer `plot = FALSE`).

2.5.2. Ejercicio

Continuando con los datos de vehículos eléctricos `ecars` de los ejercicios anteriores, generar histogramas de las variables `precio` y `logprecio` (Nota: si no se creó esta variable se puede cargar el fichero *ecars2.RData*).

2.6. Medidas características

Como ya se comentó, se puede completar la información obtenida a partir de los gráficos con valores numéricos. Estas medidas descriptivas tratan de cuantificar distintas características de la distribución de los datos.

En esta sección supondremos que $x_1, \dots, x_n$ son los valores observados de una variable numérica X. Para calcular algunas de estas medidas es necesario ordenar los datos de menor a mayor, y denotaremos por $x_{(i)}$ el valor en la i-ésima posición en la muestra ordenada (lo que se conoce como *estadístico de orden i*):

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Se clasifican en tres grandes grupos:

- De posición (central y no central).
- De dispersión.
- De forma.

2.6.1. Medidas de posición central

Estas medidas determinan un valor central de la distribución de los datos, que se suele emplear como representante de las observaciones. Entre ellas podemos destacar:

- **Media** (aritmética o media muestral; función `mean()`):

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Mediana** (M_e): es el valor que deja igual número de valores a su izquierda que a su derecha (función `median()`).
  - Si el número de datos es impar, es el valor central

$$M_e = x_{(\frac{n+1}{2})}.$$

  - Si el número de datos es par, se toma la media de los dos valores centrales

$$M_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

- **Moda** (M_o): es el valor de la variable que más veces se repite (el de mayor frecuencia absoluta).
  - Tiene sentido en variables discretas (o cualitativas).
  - En el caso de variables continuas se suelen categorizar en intervalos y se habla de intervalo modal.

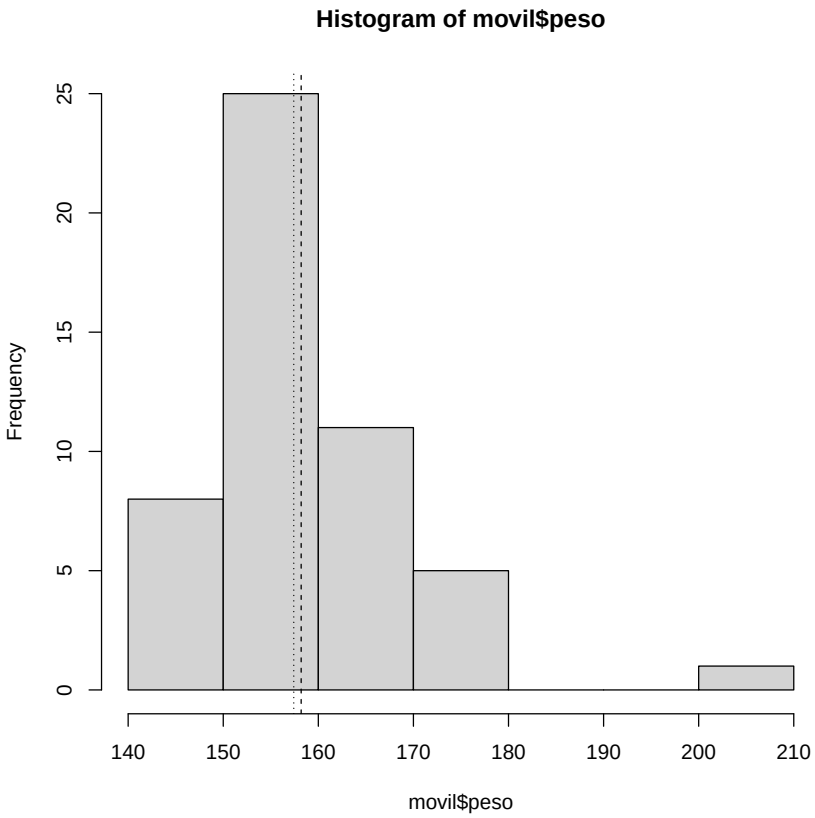
- Puede no ser única, en cuyo caso se dice que la distribución es multimodal (se habla de distribuciones unimodales, bimodales…).

Por ejemplo, continuando con los datos de móviles:

```
media <- mean(movil$peso)
mediana <- median(movil$peso)
c(media = media, mediana = mediana)
```

```
##    media  mediana
## 158.1951 157.4183
```

```
hist(movil$peso)
abline(v = media, lty = 2)
abline(v = mediana, lty = 3)
```



También podemos emplear la función `sapply()` para calcular estas medidas de forma simultánea para varias variables:

```
sapply(movil[c("precio", "bateria", "peso")], mean)
```

```
##   precio  bateria     peso
## 600.0674  22.4400 158.1951
```

Algunas de las propiedades de la media aritmética son de especial interés:

1. Toma valores en el rango de los datos: $x_{(1)} = \text{mín}(x_i) \leq \overline{x} \leq \text{máx}(x_i) = x_{(n)}$.
2. La media de una transformación lineal de los datos es igual a la transformación lineal de la media original: Si $y_i = a + bx_i$ $(i = 1, 2, \ldots, n)$, entonces $\overline{y} = a + b\overline{x}$.

3. El promedio de las desviaciones respecto a la media es cero: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

4. La media minimiza la suma de cuadrados de las distancias a las observaciones: $\bar{x} = \arg\min \sum_{i=1}^{n} (x_i - a)^2$

Por ejemplo podemos ilustrar numéricamente las propiedades 1 y 4 (esta última propiedad está relacionada con el *método de mínimos cuadrados* para el ajuste de modelos):

```
x <- movil$bateria
# Media en minutos hasta apagado + encendido
mean(60*x + 2)
```

```
## [1] 1348.4
```

```
60*mean(x) + 2
```

```
## [1] 1348.4
```

```
# Mínimos cuadrados
fdist <- function(a) sum((x - a)^2)
xdmin <- optimize(fdist, range(x))
xdmin
```

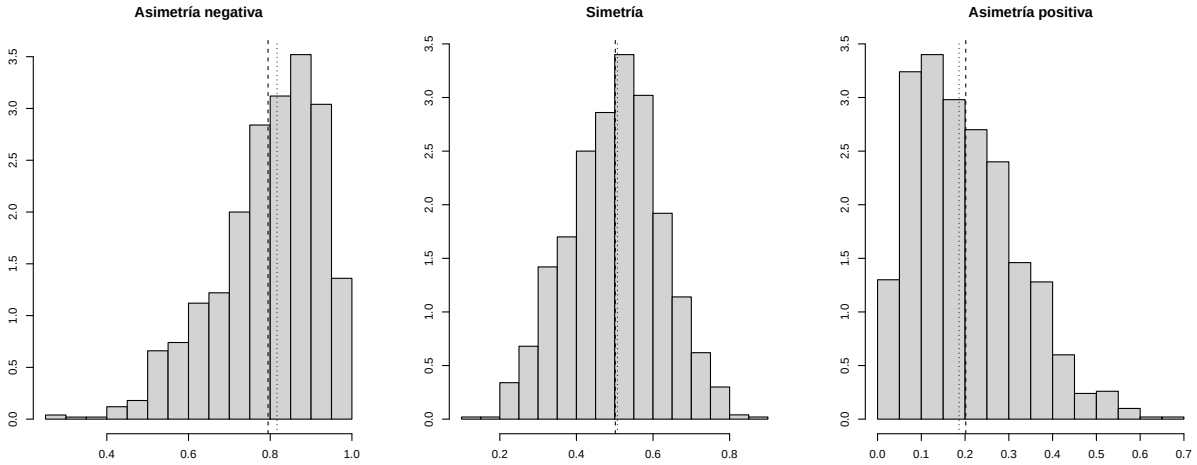
```
## $minimum
## [1] 22.44
##
## $objective
## [1] 2378.32
```

```
mean(x)
```

```
## [1] 22.44
```

2.6.2. Efecto de la asimetría y datos atípicos

Diferencias entre la media y la mediana indicarían una posible asimetría o la posible presencia de datos atípicos (**outliers**).



En el cálculo de la media (aritmética) se emplean todos los valores observados por lo que puede verse afectada por valores atípicos, esto es, no es una medida robusta. La mediana sin embargo es una medida robusta, muy poco sensible a la presencia de observaciones atípicas. Como se mostró en la práctica anterior, también se

emplean otras medidas robustas como la media truncada o la media recortada que tratan de evitar este tipo de problemas.

```
mean(movil$bateria)
```

```
## [1] 22.44
```

```
mean(movil$bateria, trim = 0.2)
```

```
## [1] 22.76667
```

```
median(movil$bateria)
```

```
## [1] 24
```

2.6.3. Medidas de posición no central: Cuantiles

Si ordenamos las observaciones de menor a mayor valor, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, el *cuantil de orden* p, (con $0 < p < 1$) es el valor q_p que deja a lo sumo np observaciones a su izquierda y $n(1-p)$ observaciones a su derecha.

Entre estas medidas destacan:

- Los **cuartiles** son los tres valores, Q_1, $Q_2 = M_e$ y Q_3, que dividen la distribución de las observaciones en cuatro partes de igual frecuencia (cuantiles de orden 0.25, 0.5 y 0.75).
- Los **deciles** dividen la distribución en diez partes de igual frecuencia, cada una conteniendo el 10 % de los valores (cuantiles de orden 0.1, 0.2, ..., 0.9).
- Los **percentiles** dividen la distribución en cien partes iguales, cada una conteniendo el 1 % de los valores (cuantiles de orden 0.01, 0.02, ..., 0.99).

El cálculo de estas medidas en la práctica se realiza de forma análoga al de la mediana. En R podemos obtenerlas empleando la función `quantile()`. Por ejemplo:

```
bateria <- movil$bateria
# Cuartiles
quantile(bateria, probs = c(0.25, 0.5, 0.75))
```

```
##   25%   50%   75%
## 16.25 24.00 28.00
```

```
# Deciles
quantile(bateria, probs = seq(0.1, 0.9, 0.1))
```

```
##  10%  20%  30%  40%  50%  60%  70%  80%  90%
## 13.8 15.8 17.7 20.0 24.0 26.0 27.0 28.0 30.1
```

Como resultado podríamos decir que un 25 % de los móviles tienen una duración de la batería inferior a 16 horas (un 10 % inferior a 13.8 horas) y un 25 % superior a 28 horas (el 90 % es inferior a 30.1 horas).

Estas medidas permiten establecer un rango de valores "normales":

```
quantile(bateria, probs = c(0.05, 0.95))
```

```
##   5%  95%
## 11.0 32.1
```

El 90 % de las duraciones están comprendidas entre 11 y 32.1 horas.

2.6.4. Ejercicio

Continuando con los datos de vehículos eléctricos `ecars` de los ejercicios anteriores:

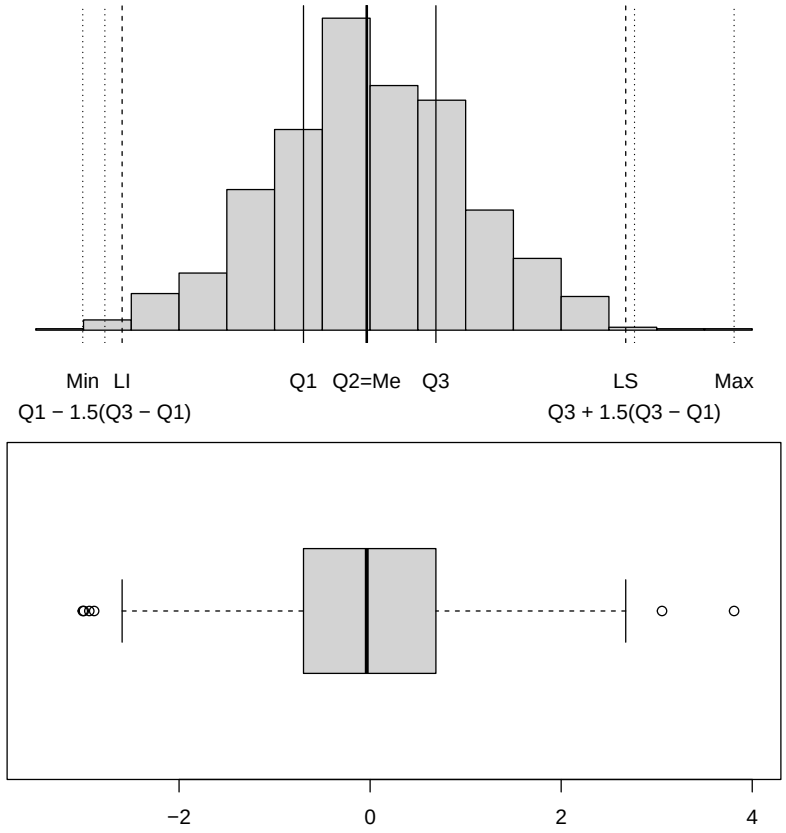
a) Obtener la media y la mediana de las variables `asientos` y `velcarga` (Nota: si aparecen problemas probar a añadir el argumento `na.rm = TRUE`).

b) Obtener los deciles de la variable `velmax` y determinar el intervalo en el que se encuentra el 80 % central de las observaciones.

c) Crear una nueva variable `nacelera` que clasifique a los coches en tres grupos de similar tamaño (porcentaje de observaciones) dependiendo de su aceleración (Baja, Media o Alta).

2.6.5. Gráficos de cajas

Los **Diagramas de caja** o **boxplots** (Tukey, 1977, *Exploratory Data Analysis*) emplean los cuartiles para representar la distribución de los datos de una forma simple. Permiten visualizar la posición y la dispersión de los datos y también detectar posibles valores atípicos. Son muy utilizados en el análisis exploratorio de datos y especialmente útiles para comparar distribuciones (análisis descriptivo multivariante).



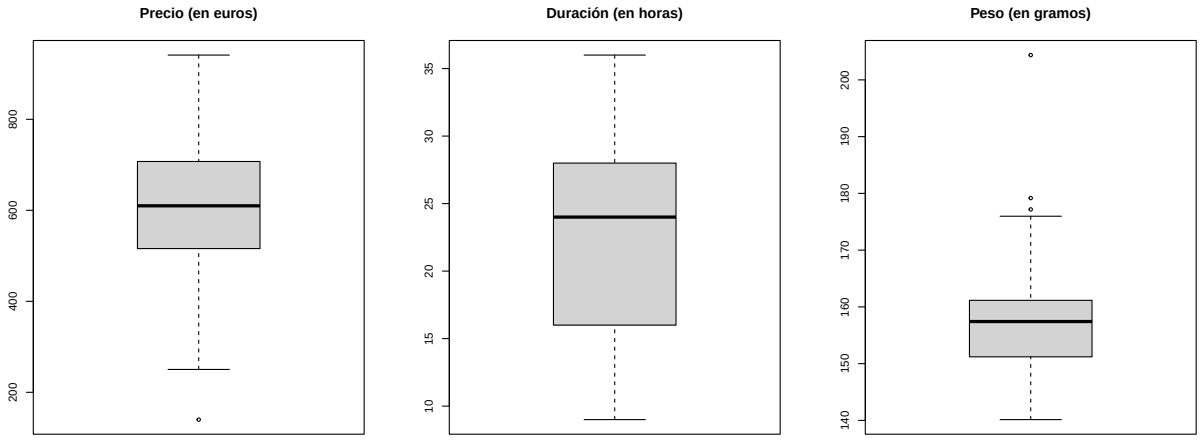
Los extremos de las cajas son el primer Q_1 y el tercer cuartil Q_3 (bisagras de Tukey), marcando con una línea el lugar que ocupa la mediana $Q_2 = M_e$. Dentro de la caja se encontrarían el 50 % central de las observaciones (un 25 % entre la mediana y cada cuartil), esto nos permite ver en torno a que valores se encuentran los datos y la forma en que se distribuyen (asimetría).
Desde la caja se dibujan dos líneas, los bigotes, hasta los límites inferior LI y superior LS (el rango de valores en el que se distribuyen la mayoría de los datos):

$$LI = \text{mín}\,\{x_i : x_i \geq Q_1 - 1{,}5(Q_3 - Q_1)\}$$
$$LS = \text{máx}\,\{x_i : x_i \leq Q_3 + 1{,}5(Q_3 - Q_1)\}$$

Las observaciones que caen fuera del intervalo (LI, LS) se consideran como posibles datos atípicos (*outliers*).

Continuando con el ejemplo de los móviles:

```
old.par <- par(mfrow = c(1,3))
boxplot(movil$precio, main = "Precio (en euros)")
boxplot(movil$bateria, main = "Duración (en horas)")
boxplot(movil$peso, main = "Peso (en gramos)")
```




```
par(old.par)
```

2.6.6. Medidas de dispersión

Son medidas de la variabilidad de los datos (respecto al centro de su distribución). Entre ellas destacan:

- **Varianza**, es la media aritmética de los cuadrados de las desviaciones respecto a la media:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- **Desviación típica**, simplemente es la raíz positiva de la varianza:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

  Está en la misma unidad de medida de la variable original.

No son medidas robustas, son muy sensibles a valores atípicos. También verifican algunas propiedades que pueden resultar de interés:

1. La varianza y la desviación típica toman siempre valores no negativos.
2. Si $y_i = a + bx_i$ $(i = 1, 2, \dots, n)$, se tiene que $s_y^2 = b^2 s_x^2$ y $s_y = |b| s_x$.
3. $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2$.

En R podemos obtener estas medidas con las funciones `var()` y `sd()`, aunque realmente calculan la cuasi-varianza:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

y la cuasi-desviación típica $\hat{s} = \sqrt{\hat{s}^2}$ (más adelante, en la parte de inferencia estadística, se justificará el motivo por el que son preferibles).

Ejemplo:

```
var(movil$bateria)
```

```
## [1] 48.53714
```

```
sd(movil$bateria)
```

```
## [1] 6.96686
```

```
sapply(movil[c("precio", "bateria", "peso")], sd)
```

```
##    precio   bateria      peso
## 163.98653   6.96686  11.25249
```

Otras medidas de dispersión que pueden ser de interés son:

- Coeficiente de variación: $CV = s/\overline{x}$ (cuanto mayor sea, menor es la representatividad de la media).
- Rango (o recorrido): $R = mx(x_i) - mn(x_i)$.
- Rango (o recorrido) intercuartílico: $IQR = Q_3(x) - Q_1(x)$.
- Desviación absoluta mediana, MAD: mediana de las desviaciones absolutas respecto a la mediana $|x_i - Me|$ (reescaladas).

Ejemplo:

```
# Coeficiente de variación
with(movil, sd(bateria)/mean(bateria))
```

```
## [1] 0.3104661
```

```
# Rango
diff(range(movil$bateria))
```

```
## [1] 27
```

```
# Recorrido intercuartílico
IQR(movil$bateria)
```

```
## [1] 11.75
```

```
# IQR(movil$bateria)/1.349 # Recorrido intercuartílico reescalado
# Desviación absoluta mediana
mad(movil$bateria)
```

```
## [1] 7.413
```

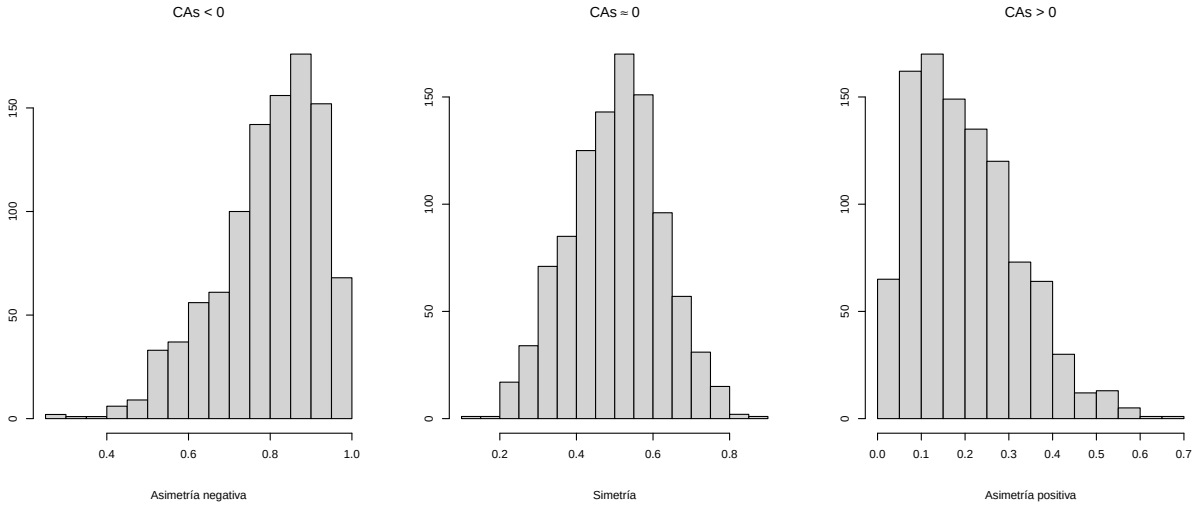
2.6.7. Medidas de forma

Como su nombre indica, estas medidas tratan de medir características de la forma de la distribución de los datos. Entre ellas podemos destacar:

- **Coeficiente de asimetría**: mide la simetría de la distribución respecto a la media

$$CAs = \frac{m_3}{s^3},$$

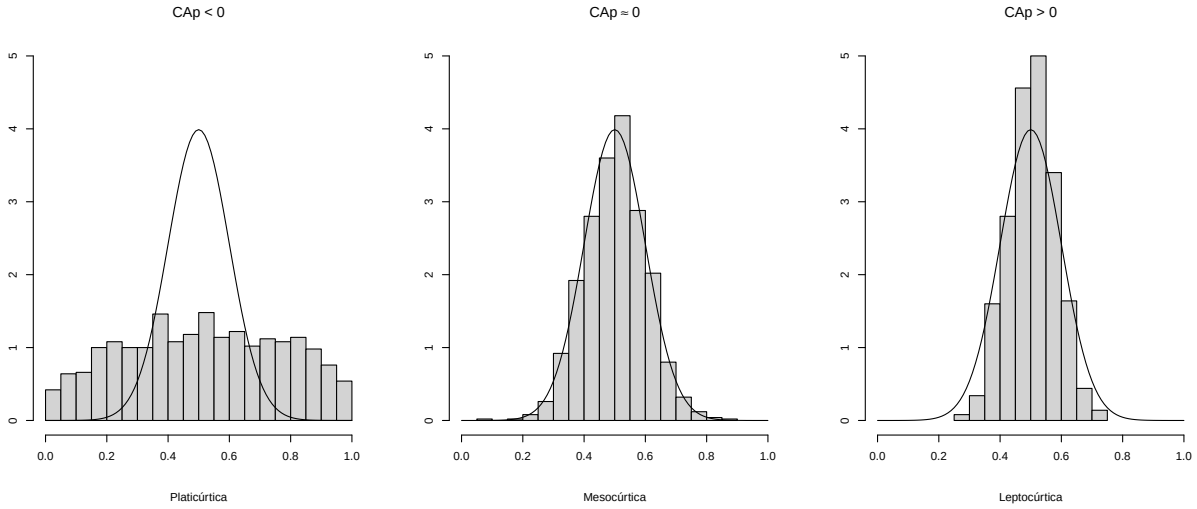
siendo $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$.



- **Coeficiente de apuntamiento o curtosis**: mide la concentración de la distribución en torno a la media

$$CAp = \frac{m_4}{s^4} - 3,$$

siendo $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$.



Para calcular estas medidas se puede emplear el paquete `e1071`:

```
library(e1071)
bateria <- movil$bateria
skewness(bateria)
```

```
## [1] -0.1246566
```

```
kurtosis(bateria)
```

```
## [1] -0.984044
```

2.6.8. Ejercicio

Continuando con los datos de vehículos eléctricos `ecars` de los ejercicios anteriores:

a) Realiza un análisis descriptivo completo de la variable `eficiencia`, incluyendo un gráfico de cajas.

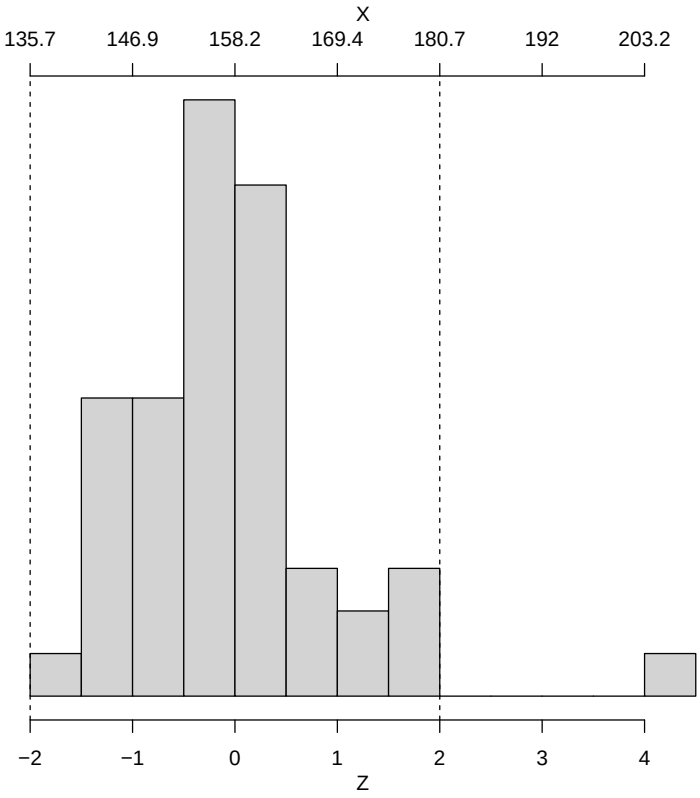
b) Obtén los coeficientes de asimetría y apuntamiento de las variables `precio` y `logprecio`.

2.7. Tipificación de variables

En ocasiones puede ser recomendable transformar la variable a otra escala. Por ejemplo en la escala tipificada (que no depende de la unidad de medida de la variable original) es más fácil identificar si un valor está alejado o próximo al centro de la distribución.

- Se define la **variable tipificada** Z de la variable estadística X como:

$$Z = \frac{X - \overline{x}}{s_x}.$$



La variable Z es adimensional con media cero $\overline{z} = 0$ y desviación típica uno $s_z = 1$. Valores a más de dos o tres desviaciones típicas de la media se suelen considerar posibles valores atípicos (siempre que la distribución de la variable no sea muy asimétrica).

Podemos tipificar una variable empleando la función `scale()`. Por ejemplo:

```
peso <- movil$peso
mean(peso); sd(peso)
```

```
## [1] 158.1951
```

```
## [1] 11.25249
```

```r
z <- scale(peso)
# z <- (peso - mean(peso))/sd(peso) # Equivalente
mean(z); sd(z)
```

```
## [1] 6.534617e-16
```

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

```r
which(abs(z) > 2)
```

```
## [1] 17
```

```r
movil$zpeso <- z
movil[17, ]
```

```
##     sexo  marca nsatisfa ncamaras   precio bateria     peso preciocat    zpeso
## 17 Mujer Huawei    Medio        2 139.9983      20 204.3724   [0,300] 4.103746
```

Transformar estadísticos (medidas, distancias…) a una escala tipificada (donde se sepa fácilmente lo que es grande o pequeño) es algo habitual en los métodos de inferencia estadística.

2.7.1. Ejercicio

Continuando con los datos de vehículos eléctricos `ecars` de los ejercicios anteriores:

a) Genera una nueva variable `zvelmax` que contenga los valores tipificados de la variable `velmax`.

b) Lista los coches con una velocidad máxima a más dos desviaciones típicas de la velocidad media.

c) Prueba a realizar un informe con los ejercicios resueltos en formato html (empleando el menú *File* > *Compile Report* o el botón correspondiente; puede ser de utilidad consultar el apéndice Introducción a RMarkdown).