Universidad de Alcalá

Estadística

Práctica 2

Estadística descriptiva con una variable

Universidad de Alcalá

Contenido

Universidad de Alcalá

1. Introducción

Con esta práctica se trata de utilizar R y RStudio para calcular medidas estadísticas sobre una variable. Los conceptos teóricos pueden encontrarse explicados en https://hilera.web.uah.es/estadistica/teoria/.

Se usarán en los ejemplos los datos suministrados por los estudiantes de un curso de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá. Con las siguientes variables estadísticas:

- GRUPO: Grupo de laboratorio
- GRADO: Orden de preferencia elegido para el Grado en Ingeniería en Sistemas de Información por la Universidad de Alcalá
- NOTA: Nota final de acceso a la universidad
- VIAJE: Tiempo en llegar a la Escuela Politécnica en minutos
- DORMIR: Horas que se duerme los días laborables
- MOVIL: Compañía de la línea móvil
- SO: Sistema operativo del móvil

GRUPO	GRADO	NOTA	VIAJE	DORMIR	MOVIL	SO
A1	Quinta opcion	8,5	75	7	Movistar/O2/T	Android
A1	Quinta opcion	7,1	90	7	Movistar/O2/T	iOS
A1	Segunda opci	8,63	60	6	Digi Mobil	Android
A1	Segunda opci	8,62	35	8	Movistar/O2/T	iOS
A1	Tercera opcio	8,2	120	7	Movistar/O2/T	Android
A1	Primera opcio	8,7	20	7	Jazztel/Orang	Android
A1	Primera opcio	7,21	20	6	Movistar/O2/T	Android
A1	Segunda opci	7,63	20	8	Movistar/O2/T	Android
A1	Quinta opcion	8,4	60	7	Lowi/Vodafon	iOS
A1	Quinta opcion	8,3	90	6	Lowi/Vodafon	iOS
A1	Tercera opcio	8,2	60	6	Lowi/Vodafon	iOS
A1	Cuarta opcion	9,1	75	7	Jazztel/Orang	Android

Para evitar problemas, se han borrado las filas en las que había algún valor vacío, y el fichero resultante se encuentra en "encuesta.csv".

Se debe cargar el fichero encuesta.csv en una variable de tipo data.frame usando la función read.csv2(), preparada para leer fichero csv con columnas separadas por ";" y decimales con ",".

```
> (encuesta = read.csv2("encuesta.csv"))
```

Se pueden crear vectores indicando con $ el nombre de la columna:

```
> grupo=encuesta$GRUPO
> grado=encuesta$GRADO
```

```
> nota=encuesta$NOTA
> viaje=encuesta$VIAJE
> dormir=encuesta$DORMIR
> movil=encuesta$MOVIL
> so=encuesta$SO
```

El vector nota contiene:

```
> nota
[1]   8.50  7.10  8.63  8.62  8.20  8.70  7.21  7.63  8.40  8.30
[11]  8.20  9.10  9.79 10.11  8.02  7.31  7.50  8.71  8.34  9.21
[21]  7.80 10.30  7.99  6.90  7.80 10.00  8.59  7.00  8.05 10.80
[31]  7.99  8.55  7.34  6.75  9.56  7.42  6.94  7.21  7.68 10.28
[41]  7.86 10.26  7.27  5.80  7.30  7.14  8.60  7.50  8.00  7.54
[51]  7.29  7.83  6.75  9.81  6.80  6.44  6.65  7.80 10.27  7.60
[61]  7.87  7.00  7.08  7.48  8.07  5.82  6.50  9.90  7.50  6.50
[71]  9.46  8.00  7.80  7.65
```

Han respondido 74 estudiantes a la encuesta, pero el número total de estudiantes matriculados es de 108. Por tanto, las 74 observaciones disponibles corresponden a una muestra:

- Muestra: 74 estudiantes que han respondido a la encuesta
- Población: 108 estudiantes matriculados

Las medidas que se calcularán en esta práctica sobre los estudiantes serán, por tanto, **estadísticos** ya que se calculan sobre una muestra (74 estudiantes). Si dispusiéramos de los datos de toda la población (108 estudiantes), las medidas se denominarían parámetros. Por ejemplo, en el caso de la medida conocida como media, existen dos casos:

- Estadístico: “media muestral”. Si se calcula con las 74 observaciones de la muestra.
- Parámetro: “media poblacional”. Si se calcula con las 108 observaciones de la población.

2. Cálculo de medidas estadísticas

En este apartado se utilizará R para calcular medidas estadísticas de centralización, de posición, de dispersión y de forma.

2.1 Medidas de centralización

Las medidas de centralización o de posición de tendencia central indican un valor alrededor del cual se distribuyen las observaciones.

Tabla 1 Cálculo de medidas de centralización con R

Medida	R	Comentarios

Media (aritmética)	`mean (x)`
	
	El resultado sería NA si hay valores NA (no disponibles) en el vector x. Para calcular la media ignorando esos valores NA hay que utilizar el parámetro `na.rm`:
	
	`mean(x, na.rm=TRUE)`
		x es un vector que contiene los valores de la variable estadística (cuantitativa), que deben ser valores numéricos.
		Se obtiene el cociente entre la suma de todos los valores y el número total de valores.

Mediana	`median(x)`	x contiene valores numéricos (variable cuantitativa) Si x tiene un número impar de valores, la mediana es el número de la posición central. Si es par, la mediana es la media de los dos valores centrales.

Moda	No existe una función predefinida en R para la moda. Podemos crear una propia, como la siguiente, adaptada de la definida en statology.org:
	
	`moda = function (x) {`
	`  u = unique(x)`
	`  m = match(x,u)`
	`  t = tabulate(m)`
	`  return (u[t == max(t)])`
	`}`
	
	O se puede instalar un paquete como "modeest" y usar la función `mlv`:
	
	`install.packages("modeest")`
	`library(modeest)`
	`mlv(x, method="mfv")`
		Obtiene el valor del vector x que más se repite.
		Puede devolver un valor, o un vector de valores en el caso de que haya más de una moda, porque existan valores que se repiten el mismo número de veces (variable multimodal).
		x puede ser un vector de valores numéricos (variable cuantitativa) o no numéricos (variable cualitativa).

2.1.1 Ejemplos

1) Media de la nota de acceso a la universidad (con dos cifras decimales)

```
> encuesta = read.csv2("encuesta.csv")
> nota=encuesta$NOTA

> round(mean(nota),2)
[1] 8.02
```

2) Mediana de la nota de acceso a la universidad

```
> round(median(nota),2)
[1] 7.81
```

Al tener el vector un número par de valores (74), podemos comprobar que R ha calculado la mediana haciendo la media de los valores de las posiciones centrales 37 (7.804) y 38 (7.830).

```
> sort(nota)
 [1]  5.80  5.82  6.44  6.50  6.50  6.65  6.75  6.75  6.80  6.90
[11]  6.94  7.00  7.00  7.08  7.10  7.14  7.21  7.21  7.27  7.29
[21]  7.30  7.31  7.34  7.42  7.48  7.50  7.50  7.50  7.54  7.60
[31]  7.63  7.65  7.68  7.80  7.80  7.80  7.80  7.83  7.86  7.87
[41]  7.99  7.99  8.00  8.00  8.02  8.05  8.07  8.20  8.20  8.30
[51]  8.34  8.40  8.50  8.55  8.59  8.60  8.62  8.63  8.70  8.71
[61]  9.10  9.21  9.46  9.56  9.79  9.81  9.90 10.00 10.11 10.26
[71] 10.27 10.28 10.30 10.80
```

3) Moda de la nota de acceso a la universidad

Podemos instalar el paquete `modeest` y usar la función `mlv`:

```
> install.packages("modeest")
> library(modeest)
> mlv(nota, method="mfv")
[1] 7.8
```

La variable nota tiene sólo una moda, es el valor 7.8, que se repite 4 veces.

Se puede calcular el número de veces que se repite la moda, con este comando:

```
> max(tabulate(match(nota,unique(nota))))
[1] 4
```

4) Moda del sistema operativo de los teléfonos móviles de los estudiantes

```
> so=encuesta$SO

> moda(so)
[1] "Android"
```

2.2 Medidas de localización (o posición)

Las medidas de localización o de posición de tendencia no central, permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Son valores de la distribución que la dividen en partes iguales, es decir en intervalos (cuantiles) que comprenden el mismo número de datos como los deciles, cuartiles y percentiles.

Tabla 2 Cálculo de medidas de localización con R

Medida	R	Comentarios
Percentil	`quantile(x,p,type=2)`	p representa el percentil divido entre 100. Por ejemplo, el percentil 35 se calcularía indicando p=0.35. NOTA: Se usa type=2, porque hay hasta 9 formas de calcular cuantiles[1], y el tipo 2 coincide con la explicada en teoría.
Cuartil	`quantile(x,q,type=2)`	El primer cuartil (Q1) se obtiene con q=0.25. El segundo cuartil (Q2) con q=0.5. El tercer cuartil (Q3) con q=0.75. El cuarto cuartil (Q4) con q=1.
Varios percentiles o cuartiles	`quantile(x,c(c1,c2,c3..),type=2)`	Se indican los percentiles o cuartiles en un vector.

2.2.1 Ejemplos

Vamos a redondear todos los resultados con 2 cifras decimales, usando la función `round()`.

[1] No existe un consenso sobre el método de cálculo de un percentil o un cuartil. Existen al menos 9 métodos diferentes, resumidos en un artículo de Hyndman y Fan (1996). La función `quantile` de R permite hacer el cálculo usando los 9 métodos del artículo, indicando el número del método (1 a 9) mediante un parámetro "`type=n`". Si no se utiliza el parámetro `type`, por defecto se aplica el método número 7. El algoritmo aplicado en cada método está explicado en el artículo y en la ayuda de R a la que se accede con `help("quantile")` y en otras fuentes. La diferencia de los resultados no es significativa, y se basa en la forma de considerar los casos en los que el porcentaje es un número exacto.

1) Percentil 10 de la nota de acceso a la universidad

```
> round(quantile(nota,0.1,type=2),2)
 10%
6.75
```

Como hay 74 estudiantes, el 10% serían 7.4. Según la fórmula explicada en teoría, el percentil 10 sería $X_{\lfloor 7.4 \rfloor+1} = X_{7+1} = x_8$. Si ordenamos la muestra y mostramos los 10 primeros valores, podemos comprobar que el de la posición 8 es 6.75 por lo que, en efecto ese es el percentil 10.

```
> sort(nota)[1:10]
 [1] 5.80 5.82 6.44 6.50 6.50 6.65 6.75 6.75 6.80 6.90
```

Lo que indica es que podemos asegurar que el 10% de los estudiantes de la muestra tiene una nota igual o inferior a 6.75, y que el 90% de los estudiantes tiene un nota de acceso igual o superior a 6.75.

2) Primer y segundo cuartiles (Q1 y Q2) de la nota de acceso a la universidad

```
> round(quantile(nota,c(0.25,0.5),type=2),2)
 25%  50%
7.27 7.81
```

Lo que indica es que la cuarta parte (25%) de los estudiantes de la muestra tiene una nota de acceso igual o inferior a 7.27 y que la mitad de los estudiantes (50%) tiene una nota de acceso igual o inferior a 7.81. Ese valor coincide con la mediana.

```
> round(median(nota),2)
[1] 7.81
```

3) Tercer Cuartil (Q3) de la nota de acceso a la universidad

```
> round(quantile(nota,0.75),2)
75%
8.6
```

Lo que indica que las tres cuartas partes (75%) de los estudiantes de la muestra tiene una nota de acceso igual o inferior a 8.6.

2.3 Medidas de dispersión

Las medidas de dispersión reflejan la heterogeneidad de las observaciones y dan una idea sobre la representatividad de las medidas de centralización, de tal forma que a mayor dispersión menor representatividad.

Tabla 3 Cálculo de medidas de dispersión con R

Medida	R	Comentarios

Rango	`range(x)`
	
	`max(x)-min(x)`
		Como vector: `range` obtiene un vector con dos valores: el valor más bajo y el más alto incluido en el vector x.
		Como valor: Diferencia entre el valor máximo y el valor mínimo.

Varianza muestral o cuasivarianza (s^2)	`var(x)`	Obtiene la varianza de la muestra de los n valores incluidos en el vector x, según la fórmula: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$

Varianza poblacional (σ^2)	No existe una función predefinida en R para la varianza poblacional. Podemos crear una propia, como la siguiente:
	`var.pob = function (x) {`
	`  n = length(x)`
	`  return(sum((x-mean(x))^2)/(n))`
	`}`
		Si el vector x contiene los valores de una variable estadística de una población completa, la varianza poblacional es:
		$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$

Desviación estándar o típica muestral (s)	`sd(x)`	Obtiene la desviación estándar de la muestra de los n valores incluidos en el vector x. $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$

Desviación estándar o típica poblacional (σ)	No existe una función predefinida en R para la cuasidesviación estándar. Podemos crear una propia, como la siguiente:
	`sd.pob = function (x) {`
	`  n = length(x)`
	`  return(sqrt(sum((x-mean(x))^2)/(n-1))`
		Si el vector x contiene los valores de una variable estadística de una población completa, la desviación estándar poblacional es:

Medida	R	Comentarios
	`}`	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$
Coeficiente de variación	`sd(x)/mean(x)`	Se obtiene un valor entre 0 y 1. Cuanto menor sea su valor, mayor será la homogeneidad de los datos de la muestra y, por tanto, la media aritmética será una buena representante del conjunto de datos.[2]
Rango intercuartílico	`IQR(x)`	Obtiene la diferencia entre el tercer cuartil (Q3) y el primero (Q1). IQR = Q3-Q1 Los datos atípicos están fuera del intervalo: $[Q_1 - 1.5IQR,\ Q_3 + 1.5IQR]$

2.3.1 Ejemplos

1) Rango de la nota de acceso a la universidad

```
> range(nota)
[1]  5.8 10.8
```

Como diferencia de notas es:

```
> max(nota)-min(nota)
[1] 5
```

2) Varianza de la nota de acceso a la universidad.

[2] Suele ocurrir en el campo de la Estadística que no hay consenso sobre la interpretación de los valores de algunas medidas estadísticas. Por ejemplo, en el caso del coeficiente de variación (CV), hay autores que afirman (también Wikipedia), que si CV≤0.3 la media aritmética es representativa de la muestra de datos, y por tanto es un conjunto de datos "Homogéneo"; mientras que si CV>0.3, la media no es representativa del conjunto de datos por ser un conjunto de datos "Heterogéneo"). Sin embargo, otros autores afinan más y proponen más rangos de valores. Se recomienda consultar este foro de discusión sobre valores aceptables para el coeficiente de variación.

NOTA: Se utiliza la varianza muestral o cuasivarianza, porque no se dispone de las notas de toda la población de estudiantes, sólo de una muestra, de los que respondieron a la encuesta.

```
> round(var(nota),2)
[1] 1.28
```

3) Desviación estándar o típica de la nota de acceso a la universidad

NOTA: Se dispone de una muestra, por lo que hay que calcular la desviación muestral.

```
> round(sd(nota),2)
[1] 1.13
```

4) Coeficiente de variación de la nota de acceso a la universidad

```
> round(sd(nota)/mean(nota),2)
[1] 0.14
```

NOTA: Muchos autores afirman que si es menor a 0.3 el conjunto de datos puede considerarse homogéneo, como ocurre en este caso.

5) Rango intercuartílico de la nota de acceso a la universidad

```
> round(IQR(nota),2)
[1] 1.32
```

Los datos atípicos están fuera del intervalo: $[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR] = [5.29,\ 6.62]$

```
> Q1=round(quantile(nota,c(0.25)),2)
> Q3=round(quantile(nota,c(0.75)),2)
>
> Q1=as.double(Q1)  #se usa as.double() para convertir a número
> Q3=as.double(Q3)
>
> round(Q1-1.5*IQR(nota),2)
[1] 5.29
> round(Q3-1.5*IQR(nota),2)
[1] 6.62
```

2.4 Medidas de forma

Las medidas de forma permiten conocer la forma que tiene la curva que representa la distribución de la frecuencia de los valores de una muestra, normalmente un histograma. Se pueden utilizar para comparar con un posible conjunto de datos con la misma media y desviación estándar pero considerado como "normal". Si imaginamos un histograma, se considera como "normal" una muestra de datos cuyo histograma se ajusta aproximadamente a la curva conocida como la campana de Gauss, que se puede dibujar con el siguiente comando de R: `curve(dnorm(x,mean(muestra),sd(muestra))`.

Las dos medidas de forma más conocidas son los coeficientes de asimetría y de apuntamiento.

Tabla 4 Cálculo de medidas de forma con R

Medida	R	Comentarios

Coeficiente de asimetría[3]	No existe una función predefinida en R, se puede crear una propia:
	`asimetria = function (x) {`
	`  n = length(x)`
	`  return ((1/n)*sum((x-mean(x))^3)/sd(x)^3)`
	`}`
	Y usarla para calcular el coeficiente:
	`asimetria(x)`
	O se puede instalar un paquete como "e1071" y usar la función `skewness`:
	`install.packages("e1071")`
	`library(e1071)`
	`skewness(x)`
		Calcula el coeficiente de asimetría respecto a la media: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^3}{s^3}$
		Si >0, es asimétrica a la derecha, es decir, la mayoría de los valores son menores que la media.
		Si <0, es asimétrica a la izquierda, la mayoría de los valores son mayores que la media.
		Si =0, es simétrica, como la distribución "normal.

[3] Como ocurre con otras medidas, no hay consenso sobre cómo calcular el coeficiente de asimetría. Aquí usamos la fórmula indicada, pero existen otras opciones, como se explican en este artículo de Joanes y Gill (1998). Si instalamos el paquete "e1071" se puede usar la función `skewness` para calcular el coeficiente de asimetría de tres formas distintas, para lo cual se puede utilizar un parámetro type con tres posibles valores: 1, 2 y 3. Por defecto se considera el tipo 3, que coincide con la fórmula usada en esta práctica.

Medida	R	Comentarios

Coeficiente de apuntamiento o curtosis[4]	No existe una función predefinida en R, se puede crear una propia:
	
	`curtosis = function (x) {`
	`  n = length(x)`
	`  return (((1/n)*sum((x-mean(x))^4)/sd(x)^4)-3)`
	`}`
	Y usarla para calcular el coeficiente:
	
	curtosis(x)
	
	O se puede instalar un paquete como "e1071" y usar la función `kurtosis`:
	
	`install.packages("e1071")`
	`library(e1071)`
	`kurtosis(x)`
		Calcula el coeficiente de apuntamiento respecto a la media:
		
		$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}-3$
		
		Si >0, cuanto mayor sea, habrá mucha diferencia de frecuencia entre los valores (histograma con pendiente muy abrupta). Se dice que es leptocúrtica.
		
		Si es <0, cuanto menor sea, la frecuencia de los valores será más parecida (histograma más plano). Se dice que es platicúrtica.
		
		Si =0, tiene un apuntamiento como la distribución "normal". Se dice que es mesocúrtica.

Se puede hacer un estudio de la normalidad de una variable en función del valor de sus coeficientes de asimetría y apuntamiento. Existen propuestas de rangos de valores admitidos para cada coeficiente**.** No se puede asegurar con certeza que una variable sea normal, por lo que su es**t**udio se llevará a cabo en las prácticas sobre estadística inferencial en las que se realizarán pruebas de hipótesis.

[4] Tampoco hay consenso sobre cómo calcular el coeficiente de apuntamiento. Aquí usamos la fórmula indicada, pero existen otras opciones, como se explican en este artículo de Joanes y Gill (1998). Si instalamos el paquete "e1071" se puede usar la función `kurtosis` para calcular el coeficiente de apuntamiento de tres formas distintas, para lo cual se puede utilizar un parámetro type con tres posibles valores: 1, 2 y 3. Por defecto se considera el tipo 3, que coincide con la fórmula usada en esta práctica.

2.4.1 Ejemplos

1) Coeficiente de asimetría de las notas de acceso a la universidad

Podemos instalar el paquete e1071 y usar la función skewness:

```
> install.packages("e1071")
> library(e1071)
> round(skewness(nota),2)
[1] 0.57
```

Al ser un valor positivo, la mayoría de los valores son menores que la media. La distribución es asimétrica a la derecha o positiva.

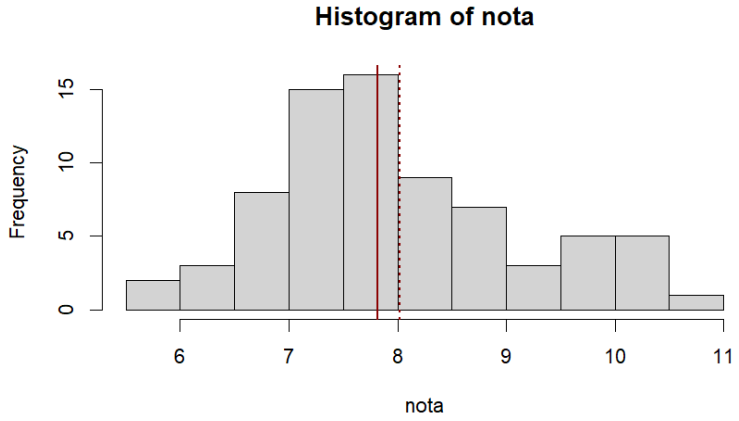
Se puede comprobar que es así, si extraemos los valores menores que la media.

```
> length(nota[nota<mean(nota)])
[1] 45
```

Del total de 74 notas, hay 45 notas menores que la media y 29 mayores.

Podemos comprobar visualmente la asimetría dibujando un histograma, la mediana con una línea continua roja y la media con una línea punteada roja. Se comprueba gráficamente que, en efecto, es asimétrica a la derecha.

```
> hist(nota)
> abline(v=mean(nota), lwd=2, lty=3, col="darkred")
```



2) Coeficiente de apuntamiento o curtosis de las notas de acceso a la universidad

Podemos utilizar la función `kurtosis` del paquete `e1071`:

```
install.packages("e1071")
library(e1071)
> round(kurtosis(nota),2)
[1] -0.28
```

Como es menor que cero, la distribución es platicúrtica, es decir, tiene poco apuntamiento. No hay muchas diferencias entre las frecuencias de los valores, hay menos diferencia que en caso de una variable “normal”.

2.5 Resumen de medidas

Existe en R la función summary que obtiene un resumen de las principales medidas estadísticas descriptivas: valores mínimo y máximo, media, mediana, y primer y tercer cuartil.

```
> round(summary(nota),2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.80    7.27    7.81    8.02    8.60   10.80 
```

3. Tabla de frecuencias

Las principales tablas de frecuencias se obtienen como se indica en la siguiente tabla:

Operación	R	Comentarios
Tabla de frecuencias absolutas	`table(x)`	Número de veces que se repite cada valor de la variable estadística en el vector x.

Tabla de frecuencias relativas	`prop.table(table(x))`
	O bien:
	`table(x)/length(x)`
		Se obtiene el valor de cada frecuencia absoluta dividido por el número total de valores de la variable estadística en el vector x.

Tabla de frecuencias acumuladas	`cumsum(table(x))`	Obtiene una tabla con la Suma todos los valores

Se pueden crear tablas de frecuencias absolutas de cualquier variable estadística, por ejemplo:

```
> frec.abs.nota=table(encuesta$NOTA)
> frec.abs.dormir=table(encuesta$DORMIR)
> frec.abs.so=table(encuesta$SO)
```

Y también tablas de frecuencias relativas:

```
> frec.rel.nota=prop.table(frec.abs.nota)
> frec.rel.dormir=prop.table(frec.abs.dormir)
> frec.rel.so=prop.table(frec.abs.so)
```

Y tablas de frecuencias acumuladas:

```
> frec.acu.nota= cumsum(frec.abs.nota)
> frec.acu.dormir= cumsum(frec.abs.dormir)
> frec.acu.so= cumsum(frec.abs.so)
```

Se puede comprobar que se han creado, por ejemplo, las del sistema operativo del móvil:

```
> frec.abs.so
Android     iOS
     50      24

> frec.rel.so
  Android       iOS
0.6756757 0.3243243

> frec.acu.so
Android     iOS
     50      74
```

O de las de la variable que representa las horas que duermen los estudiantes (redondeamos los valores de las frecuencias relativas para que se muestren pocos decimales, por ejemplo 3):

```
> frec.abs.dormir
 4  5  6  7  8  9
 1  4 22 32 14  1

> round(frec.rel.dormir,3)
    4     5     6     7     8     9
0.014 0.054 0.297 0.432 0.189 0.014

> frec.acu.dormir
 4  5  6  7  8  9
 1  5 27 59 73 74
```

Se puede comprobar con la función `sum` que la suma de las frecuencias relativas es 1:

```
> sum(frec.rel.dormir)
[1] 1
```

3.1 Tabla de frecuencias absolutas para datos agrupado en intervalos de clase

Cuando una variable estadística cuantitativa es continua, es habitual que las frecuencias no se calculen para cada valor sino para una clase o intervalo de valores.

En ese caso, la tabla de frecuencias absolutas se calcula de la siguiente forma:

```
K=número de intervalos o clases
C = diff(range(x))/K
Li=min(x)
L = Li+C*(0:K)
x.agrupada = cut(x, breaks=L, right=FALSE)
frec.abs.agrupada = table(x.agrupada))
```

El significado de cada variable es:

- c: Amplitud de clase o tamaño de cada intervalo
- Li: Límite inferior del intervalo de la primera clase
- L: Vector con los límites de cada clase

Podemos aplicarlo para obtener la tabla de frecuencias absolutas de la nota de acceso a la universidad agrupada en 4 intervalos.

```
> K=4
> C= diff(range(nota))/K
> Li=min(nota)
> L = Li+C*(0:K)
> nota.agrupada = cut(nota, breaks=L, right=FALSE, include.lowest=TRUE)
> frec.abs.nota.agrupada = table(nota.agrupada)
> frec.abs.nota.agrupada
```

```
nota.agrupada
 [5.8,7.05)  [7.05,8.3)  [8.3,9.55) [9.55,10.8)
         13          36          14          10
```

Podemos comprobar que en la tabla se contabilizan 73 valores, cuando en la variable nota hay 74. Se ha perdido el valor más alto, porque los intervalos están abiertos a la derecha, lo que quiere decir que el extremo derecho del intervalo no está incluido en el mismo.

Para evitar perder ese valor, hay que indicar que el último intervalo debe estar cerrado a la derecha, y eso se consigue con el atributo include.lowest=true:

```
> nota.agrupada = cut(nota, breaks=L, right=FALSE, include.lowest = TRUE)
> frec.abs.nota.agrupada = table(nota.agrupada)
> frec.abs.nota.agrupada
nota.agrupada
 [5.8,7.05)  [7.05,8.3)  [8.3,9.55) [9.55,10.8]
         13          36          14          11
```

Ahora sí se han incluido en la tabla los 74 valores de nota, por lo que hay que hacerlo de esta forma..

Para entender el funcionamiento, podemos analizar el resultado de cada comando.

1) El primer comando obtiene la variable K con el número de intervalos que queremos:

```
> (K=4)
[1] 4
```

2) El segundo comando obtiene una variable C que representa la amplitud de clase o tamaño de cada uno de los cuatro intervalos en que hay que dividir el rango de valores disponibles en el vector nota:

```
> (C = diff(range(nota))/K)
[1] 1.25
```

3) El tercer comando obtiene la variable Li, que representa el límite inferior del primer intervalo o intervalo de la primera clase, que debe coincidir con el mínimo valor en el vector nota:

```
> (Li=min(nota))
[1] 5.8
```

4) El cuarto comando obtiene un vector L, con los valores de los límites de cada intercalo. Al haber dividido en 4 intervalos, hay 5 valores:

```
> (L = Li+C*(0:k))
[1]  5.80  7.05  8.30  9.55 10.80
```

5) El quinto comando obtiene un vector nota.agrupada,

```
> (nota.agrupada = cut(nota, breaks=L, right=FALSE, include.lowest = TRUE))
 [1] [8.3,9.55)  [7.05,8.3)  [8.3,9.55)
 [4] [8.3,9.55)  [7.05,8.3)  [8.3,9.55)
 [7] [7.05,8.3)  [7.05,8.3)  [8.3,9.55)
[10] [8.3,9.55)  [7.05,8.3)  [8.3,9.55)
```

```
[13] [9.55,10.8] [9.55,10.8] [7.05,8.3)
[16] [7.05,8.3)  [7.05,8.3)  [8.3,9.55)
[19] [8.3,9.55)  [8.3,9.55)  [7.05,8.3)
[22] [9.55,10.8] [7.05,8.3)  [5.8,7.05)
[25] [7.05,8.3)  [9.55,10.8] [8.3,9.55)
[28] [5.8,7.05)  [7.05,8.3)  [9.55,10.8]
[31] [7.05,8.3)  [8.3,9.55)  [7.05,8.3)
[34] [5.8,7.05)  [9.55,10.8] [7.05,8.3)
[37] [5.8,7.05)  [7.05,8.3)  [7.05,8.3)
[40] [9.55,10.8] [7.05,8.3)  [9.55,10.8]
[43] [7.05,8.3)  [5.8,7.05)  [7.05,8.3)
[46] [7.05,8.3)  [8.3,9.55)  [7.05,8.3)
[49] [7.05,8.3)  [7.05,8.3)  [7.05,8.3)
[52] [7.05,8.3)  [5.8,7.05)  [9.55,10.8]
[55] [5.8,7.05)  [5.8,7.05)  [5.8,7.05)
[58] [7.05,8.3)  [9.55,10.8] [7.05,8.3)
[61] [7.05,8.3)  [5.8,7.05)  [7.05,8.3)
[64] [7.05,8.3)  [7.05,8.3)  [5.8,7.05)
[67] [5.8,7.05)  [9.55,10.8] [7.05,8.3)
[70] [5.8,7.05)  [8.3,9.55)  [7.05,8.3)
[73] [7.05,8.3)  [7.05,8.3)
4 Levels: [5.8,7.05) ... [9.55,10.8]
```

6) El sexto comando obtiene la tabla de frecuencias absolutas:

```
> (frec.abs.nota.agrupada = table(nota.agrupada))
nota.agrupada
 [5.8,7.05)  [7.05,8.3)  [8.3,9.55) [9.55,10.8]
         13          36          14          11
```

3.2 Tabla de frecuencias relativas para datos agrupado en intervalos de clase

Hay dos opciones para calcular las frecuencias relativas a partir de las absolutas.

1) Utilizando la función prop.table

```
> (frec.rel.nota.agrupada = prop.table(frec.abs.nota.agrupada))
nota.agrupada
 [5.8,7.05)  [7.05,8.3)  [8.3,9.55) [9.55,10.8]
  0.1756757   0.4864865   0.1891892   0.1486486
```

2) Dividiendo las frecuencias absolutas por el número de datos

```
> (frec.rel.nota.agrupada = frec.abs.nota.agrupada/length(nota))
nota.agrupada
 [5.8,7.05)  [7.05,8.3)  [8.3,9.55) [9.55,10.8]
  0.1756757   0.4864865   0.1891892   0.1486486
```

3.3 Media de una variable a partir de datos agrupados

Se calcula considerando que hay un valor (marca) que representa a todos los datos incluidos en un mismo intervalo. Ese valor o marca es el punto medio de cada intervalo intervalo:

```
> (marcas = (L[1:K]+L[1:K+1])/2)
[1]  6.425  7.675  8.925 10.175
```

Una vez que obtenemos las marcas, sólo hay que multiplicar cada marca por la frecuencia del intervalo al que representa:

```
> (media.nota.agrupada=sum(marcas*frec.rel.nota.agrupada))
[1] 8.063514
```

Podemos comparar con la media que se obtiene con todos los datos sin agrupar:

```
> mean(nota)
[1] 8.022568
```

Vemos que son muy parecidas, pero no iguales ya que al agrupar se pierde información.

Se podrían calcular de forma similar otras medidas como la varianza o la desviación estándar.

4. Gráficos

Algunos gráficos que se pueden dibujar utilizando funciones de R sobre los valores de una variable estadística son histogramas, diagramas de barras, diagramas de tarta, diagramas de caja (box plot) y diagramas de tallo y hojas.

4.1 Diagrama de tallo y hojas

Se realiza para variables cuantitativas, utilizando la función `stem(variable)`.

Podemos utilizar, por ejemplo, la variable DORMIR sobre las horas que duermen los alumnos:

```
> dormir
 [1] 7 7 6 8 7 7 6 8 7 6 6 7 6 7
[15] 8 6 5 6 6 7 6 9 7 7 7 4 6 8
[29] 6 6 6 7 7 5 7 7 7 5 5 7 8 8
[43] 6 6 6 7 7 7 7 8 7 7 7 7 8 7
[57] 7 8 6 7 8 6 7 7 6 7 6 8 6 6
[71] 7 8 8 8
```

Y realizar el diagrama de tallo y hojas:

```
> stem(dormir)

  The decimal point is at the |

  4 | 0
  5 | 0000
  6 | 00000000000000000000
  7 | 00000000000000000000000000000000
  8 | 00000000000000
  9 | 0
```

Podemos comprobar visualmente que 7 horas es el valor más usual, después 6 horas y luego 8 horas.

Se puede hacer el diagrama de otras variables numéricas, por ejemplo las notas de acceso.

```
> nota
 [1]  8.50  7.10  8.63  8.62  8.20  8.70  7.21  7.63  8.40  8.30
[11]  8.20  9.10  9.79 10.11  8.02  7.31  7.50  8.71  8.34  9.21
[21]  7.80 10.30  7.99  6.90  7.80 10.00  8.59  7.00  8.05 10.80
[31]  7.99  8.55  7.34  6.75  9.56  7.42  6.94  7.21  7.68 10.28
[41]  7.86 10.26  7.27  5.80  7.30  7.14  8.60  7.50  8.00  7.54
[51]  7.29  7.83  6.75  9.81  6.80  6.44  6.65  7.80 10.27  7.60
[61]  7.87  7.00  7.08  7.48  8.07  5.82  6.50  9.90  7.50  6.50
[71]  9.46  8.00  7.80  7.65
```

En este caso, al tener 2 dígitos decimales, la función `stem` de R aplica previamente un redondeo a una sola cifra.

```
> stem(nota)

  The decimal point is at the |

   5 | 88
   6 | 455788899
   7 | 00111223333345555566778888899
   8 | 00000112233456666677
   9 | 1256889
  10 | 0133338
```

Observamos que las notas que más se repiten son las que están en el rango de 7 a 7.9 puntos.

Para que no se aplique el redondeo, se puede utilizar el parámetro `scale`, indicando `scale=3`.

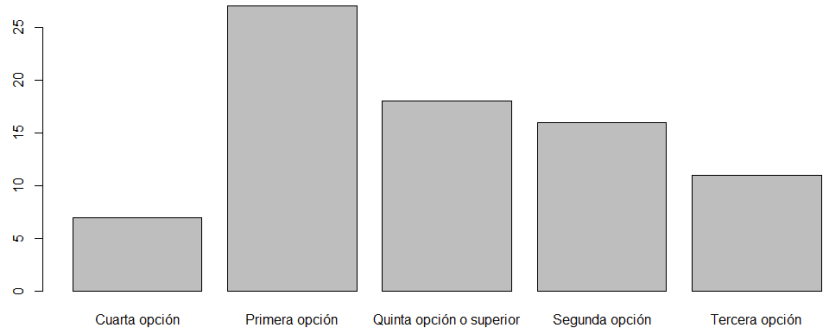
```
> stem(nota,scale=3)

  The decimal point is 1 digit(s) to the left of the |

   58 | 02
   60 |
   62 |
   64 | 400
   66 | 555
   68 | 004
   70 | 00804
   72 | 1179014
   74 | 280004
   76 | 0358
   78 | 000036799
   80 | 00257
   82 | 0004
   84 | 0059
   86 | 02301
   88 |
   90 | 0
   92 | 1
   94 | 66
   96 | 9
   98 | 10
  100 | 01
  102 | 6780
  104 |
  106 |
  108 | 0
```

4.2 Diagrama de barras

Se suele realizar para variables cualitativas, y para variables cuantitativas si tienen un número reducido de valores diferentes.

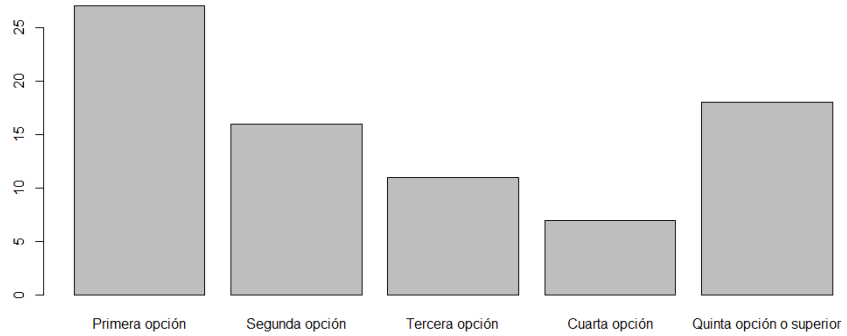
Se realiza con la función `barplot()`, a partir de la tabla de frecuencias absolutas o relativas. Por ejemplo, de la variable estadística GRADO, que representa en qué opción eligió un alumno la titulación de Grado en Ingeniería en Sistemas de Información.

```
> frec.abs.grado=table(grado)
> barplot(frec.abs.grado)
```



Si queremos que aparezcan en otro orden, se puede definir un orden:

```
> orden=c("Primera opción","Segunda opción","Tercera opción","Cuarta
opción","Quinta opción o superior")
> barplot(frec.abs.grado[orden])
```



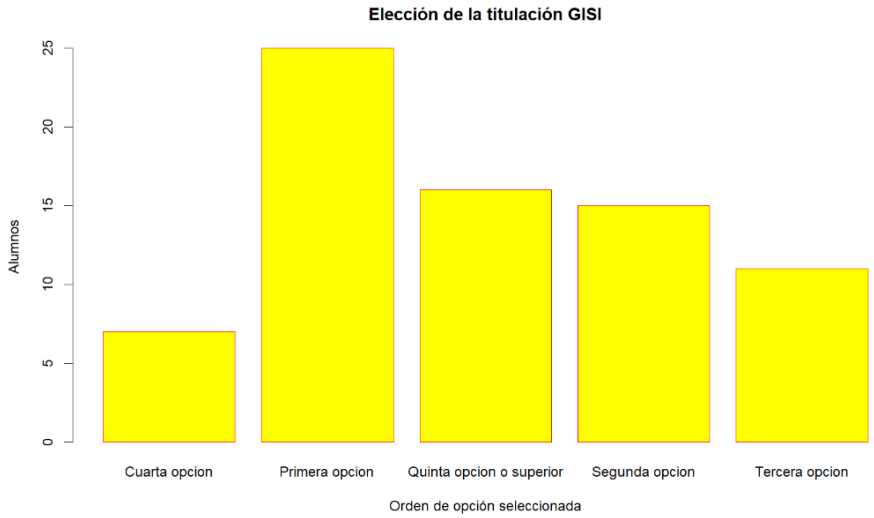
En la función `barplot` se pueden utilizar parámetros como:

- Título del diagrama: `main="título"`

- Subtítulo del diagrama: `sub="subtítulo"`
- Etiqueta para el eje x: `xlab="etiqueta"`
- Etiqueta para el eje y: `ylab="etiqueta"`
- Color de la línea: `col="color"` (ej. “red”, “blue”, “green”, etc., lista completa en https://r-charts.com/es/colores/)
- Color del borde de las columnas: `border="color"`

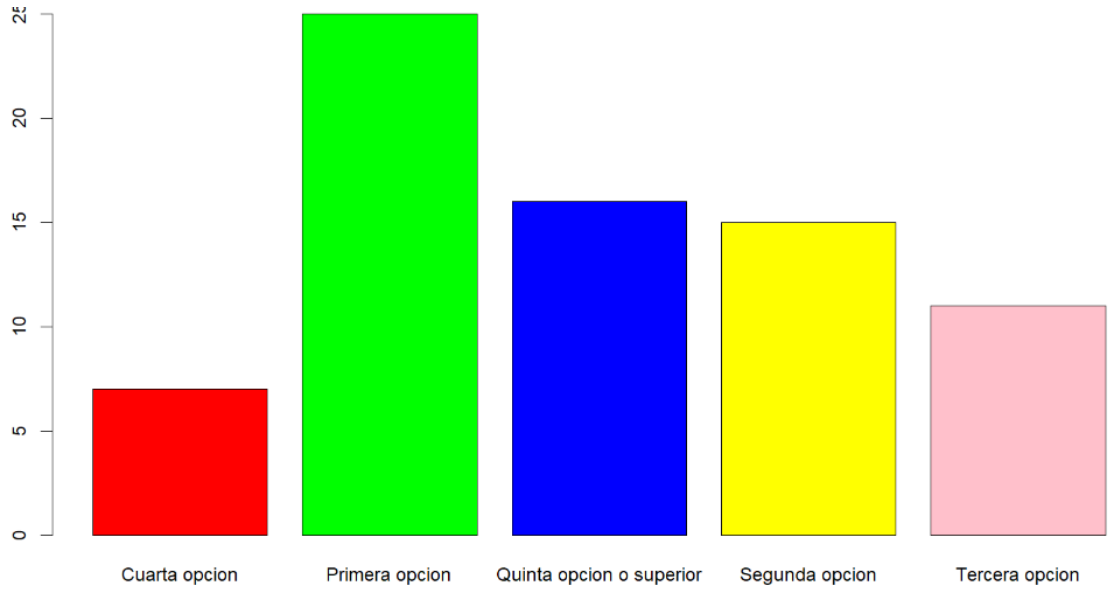
Un ejemplo sería:

```
>barplot (frec.abs.grado, main="Elección de la titulación GISI", xlab="Orden de opción seleccionada", ylab="Alumnos", col="yellow", border = "red")
```



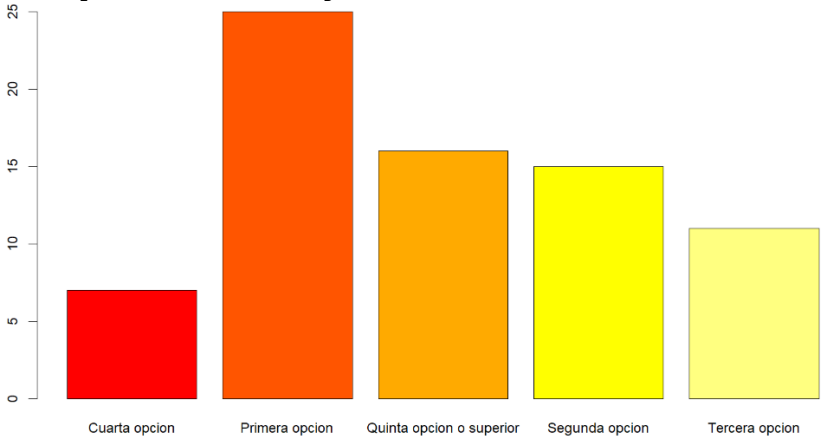
Se puede indicar un color para cada barra.

```
> barplot (frec.abs.grado, col=c("red","green","blue","yellow","pink"))
```



O aplicar una gama de colores usando la función heat.

```
> barplot (frec.abs.grado, col=heat.colors(5))
```



Para conocer más parámetros de la función `barplot` se puede consultar la ayuda de R:

```
> help(barplot)
```

4.3 Diagrama de Pareto

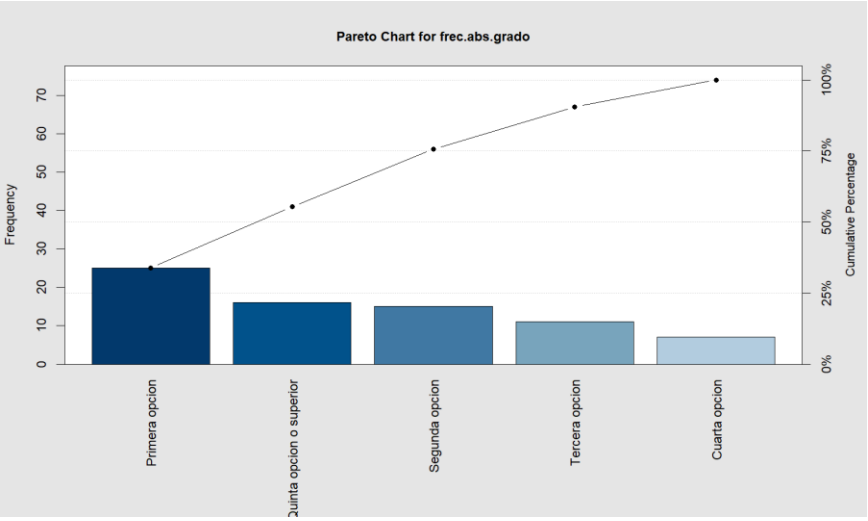
Es un diagrama de barras en orden descendente según frecuencias absolutas, en el que se superpone una línea con las frecuencias acumuladas.

Para dibujarlo se puede importar el paquete “qcc”:

```
> install.packages("qcc")
> library (qcc)
```

Y se obtiene con la función pareto.chart, a la que se pasa como parámetro la tabla de frecuencias absolutas.

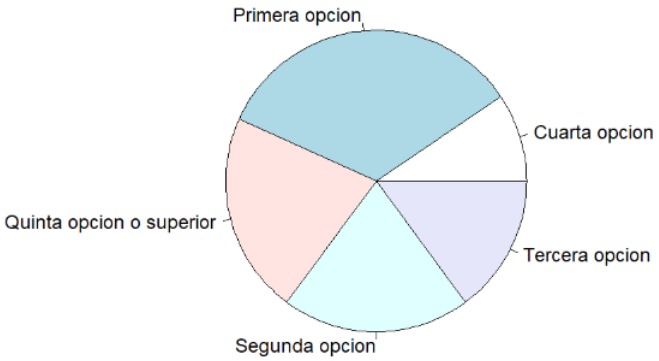
```
> pareto.chart(frec.abs.grado)

Pareto chart analysis for frec.abs.grado
                          Frequency  Cum.Freq. Percentage Cum.Percent.
  Primera opcion           25.000000  25.000000  33.783784    33.783784
  Quinta opcion o superior 16.000000  41.000000  21.621622    55.405405
  Segunda opcion           15.000000  56.000000  20.270270    75.675676
  Tercera opcion           11.000000  67.000000  14.864865    90.540541
  Cuarta opcion             7.000000  74.000000   9.459459   100.000000
```



4.4 Diagrama de tarta

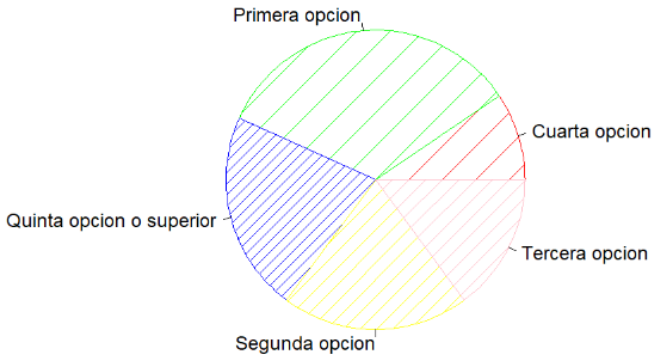
Se utiliza con variables cualitativas.

```
> pie(frec.abs.grado)
```



Se puede indicar un color y una densidad de rallado para cada valor de la variable.

```
> pie(frec.abs.grado, col=c("red","brown","blue","yellow","pink"),
density=c(10,10,30,20,20))
```



Para conocer más parámetros de la función `pie` se puede consultar la ayuda de R:

```
> help(pie)
```

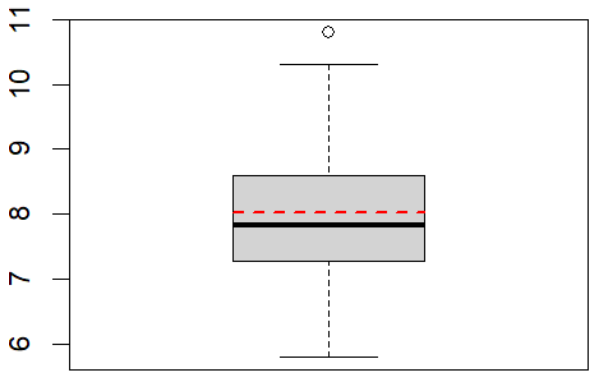
4.5 Diagrama de caja (box plot)

Sólo se puede realizar para variables cuantitativas. En un diagrama de caja se muestra información que también se obtiene con la función summary.

```
> summary(nota)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.80    7.27    7.81    8.02    8.60   10.80
```

El diagrama se dibuja utilizando la función boxplot(variable) y la función segments() para dibujar también la media dentro de la caja, como una línea roja discontinua.

```
> boxplot(nota)
> segments(x0 = 0.8, y0 = mean(nota), x1 = 1.2, y1 = mean(nota), col = "red", lwd
= 2, lty=2)
```



La línea sólida que aparece dentro de la caja representa la mediana (7.81), el borde inferior de la caja es el primer cuartil (Q1=7.27), y el borde superior el tercer cuartil (Q3=8.60). Por tanto la altura de la caja es el rango intercuartílilco (IQR=1.32).

Fuera de la caja hay unas líneas llamadas bigotes (whiskers), que se dibujan en a una distancia de 1.5*IQR de los bordes inferior y superior. Excepto en el caso de que este valor sea, respectivamente, menor que el mínimo de la variable o mayor que el máximo. En este ejemplo:

- El bigote inferior sería Q1-1.5*IQR=7.27-1.5*1.32=5.29, pero como este valor es menor que el mínimo de la variable (5.80), entonces el bigote corresponde al valor mínimo de la variable.
- El bigote superior sería Q3+1.5*IQR=8.6+1.5*1.32=10.58. Como no es mayor que el valor máximo de la variable (10.80), entonces en este caso es el valor calculado.

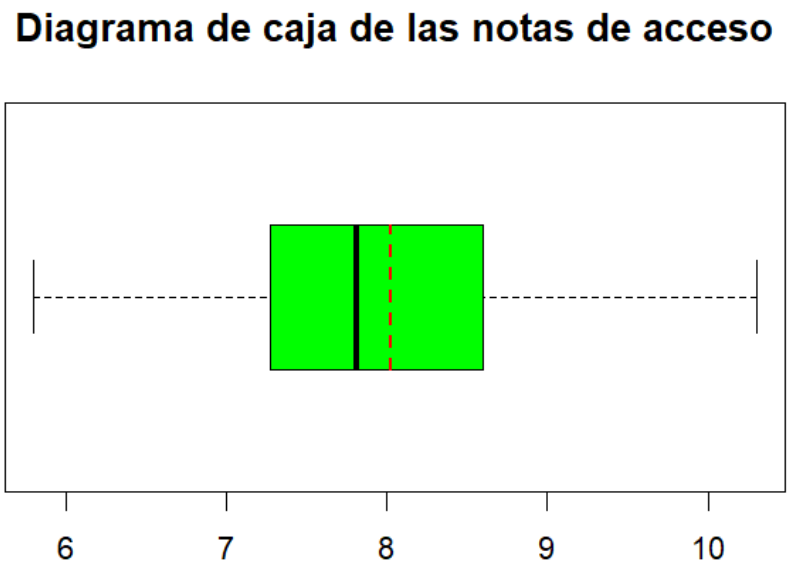
El un diagrama de caja se dibujan con puntos los denominados “datos atípicos” (*outliers*), que son los valores de la variable fueran de los bigotes. En este ejemplo sólo hay uno, se trata de la nota 10.80, es es mayor al bigote superior (10.58).

Se pueden utlizar parámentos de la función boxplot, para mejorar el diagrama. Por ejemplo:

- Con el parámetro horizontal=TRUE, se dibuja en horizontal.

- Con el parámetro outline=FALSE, no se dibujan los datos atípicos
- Con el parámetro main se indica el título del diagrama
- Con el parámetro col se indica el color de rerlleno de la caja

```
> boxplot(nota, horizontal=TRUE, outline=FALSE, main="Diagrama de caja de las notas
de acceso", col="green")
```



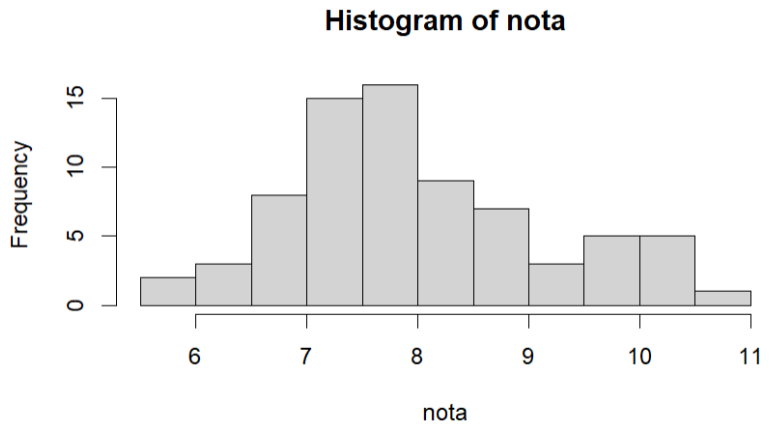
Para conocer más parámetros de la función `boxplot` se puede consultar la ayuda de R:

```
> help(boxplot)
```

4.6 Histograma

Sólo se pueden realizar para variables cuantitativas.

```
> hist(nota)
```



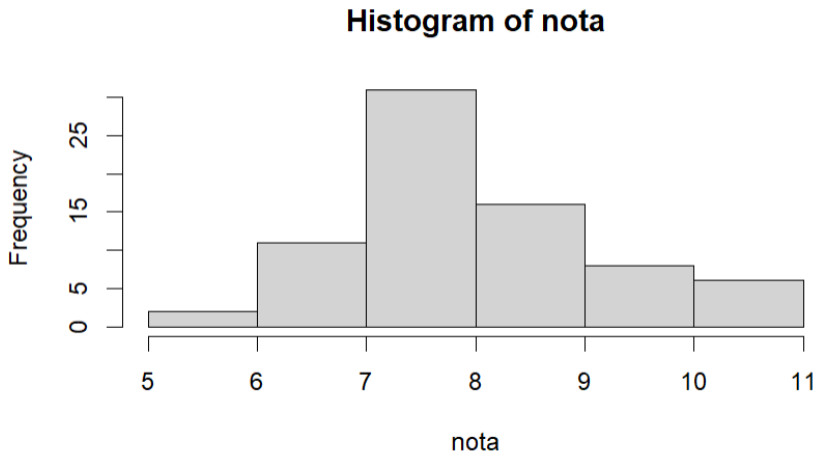
El número de clases o barras (*bins*) del diagrama por defecto es el que establece la fórmula de Sturges[5]. Aunque se pueden utilizar fórmulas propuestas por otros autores, mediante el parámetro breaks:

- Fórmula de Sturges: breaks=”Sturges” (por defecto)
- Fórmula de Scott: breaks=”Scott”
- Fórmula de Freedman-Diaconis: breaks=”FD”

Por ejemplo, usando la propuesta de Scott:

```
> hist(nota,breaks="Scott")
```

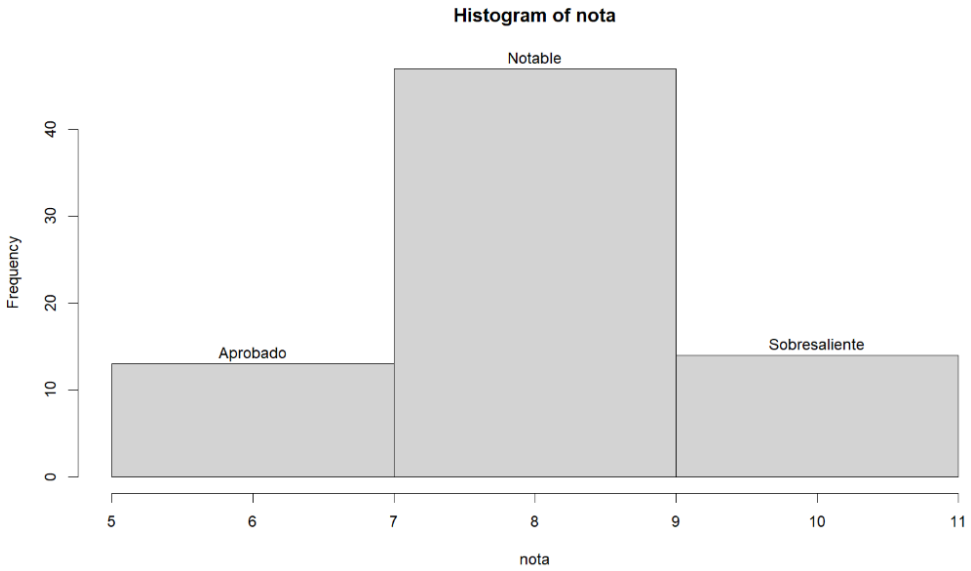
[5] R no aplica estrictamente la fórmula de Sturges, que en este caso sería 1+log2(74)=8. Existe una función en R para calcular la fórmula de Sturges: nclass.Sturges(nota), y también se obtiene 8. Sin embargo, podemos comprobar que en el histograma hay 11 barras y no 8 como debería haber. En cambio la fórmula de Scott si se aplica correctamente, pues con la fórmula se obtiene el valor 6 con nclass.Scott(nota), y en el diagrama en efecto hay 6 barras.



En el primer caso, por defecto se dibujaron 11 barras, mientras que en el segundo caso el diagrama se divide en sólo 6 barras.

Se puede usar el parámetro breaks para indicar un vector con los puntos en los que debe terminar cada columna. Por ejemplo, podríamos dibujar el histograma de notas para ver las notas de aprobado (entre 5 y 7), las de notable (entre 7 y 9) y las de sobresaliente (entre 9 y 11). Y añadir el parámetro labels, para que aparezca un texto sobre cada barra:

```
> hist(nota,breaks=c(5,7,9,11),labels=c("Aprobado","Notable","Sobresaliente"))
```

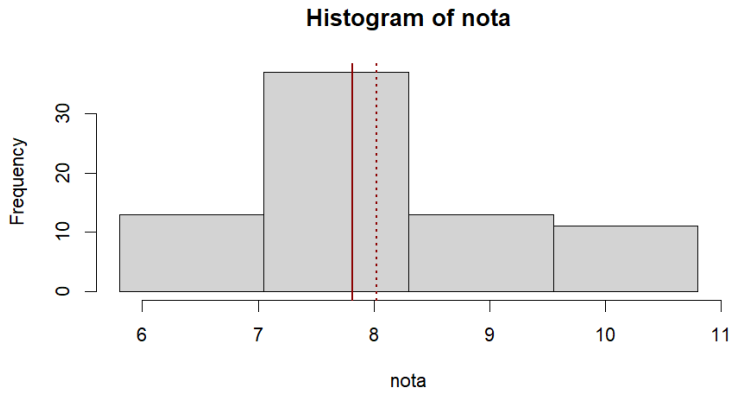


Se podría usar el vector L que se generó con los límites de los intervalos de datos agrupados, por ejemplo, para cuatro intervalos o clases:

```
> K=4
> C = diff(range(nota))/K
> Li=min(nota)
> L = Li+C*(0:K)
> hist(nota,breaks=L)
```

Y si queremos que se vea la posición de la media y la mediana, podemos dibujar sobre el diagrama una línea vertical roja discontinua en el valor de la media y una línea roja continua en el valor de al mediana.

```
> abline(v=mean(nota), lwd=2, lty=3, col="darkred")
> abline(v=median(nota), lwd=2, col="darkred")
```

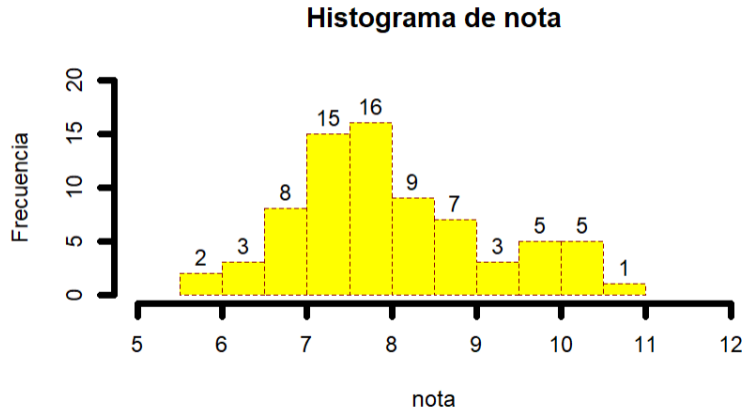


En la función `hist` se pueden parámetros como:

- Título del diagrama: main=”título”
- Subtítulo del diagrama: sub=”subtítulo”
- Etiqueta para el eje x: xlab=”etiqueta”
- Etiqueta para el eje y: ylab=”etiqueta”
- Grosor de la línea: lwd=”número”
- Color de la línea: col=”color” (ej. “red”, “blue”, “green”, etc., lista completa en https://r-charts.com/es/colores/)
- Límites del eje x: xlim=c(mínimo,máximo)
- Límites del eje y: xlim=c(mínimo,máximo)
- Tipo de línea: lty=0,1,2,3,4,5,6 (0=invisible, 1=sólida, 2=rayas, 3=puntos, 4=puntos y rayas, 5= rayas largas, 6=rayas cortas y largas)
- Número de frecuencia encima de cada columna: labels=TRUE
- Color del borde de las columnas: border=”color”

Un ejemplo sería:

```
> hist (nota, main="Histograma de nota", ylab="Frecuencia", lwd=5, col="yellow", lty=2, xlim = c(5,12), ylim=c(0,20), labels=TRUE, border = "darkred")
```

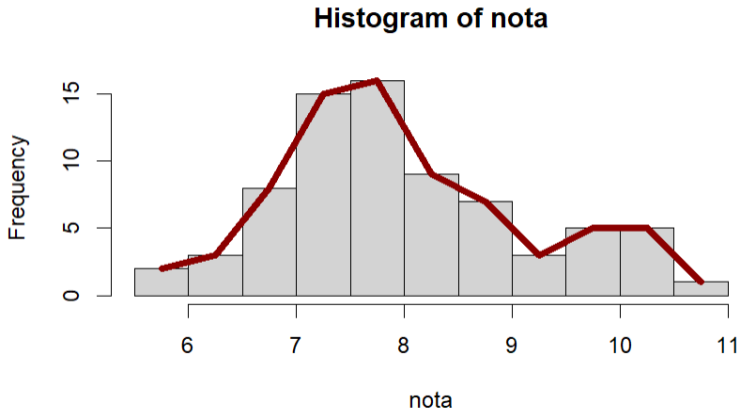


4.7 Polígono de frecuencias

Un polígono de frecuencias se dibuja con líneas que unan los puntos medios de las barras de un histograma.

Puede hacerse de la siguiente forma.

```
> h=hist(nota)
> lines(h$mids,h$counts, col="darkred", lwd=5)
```



Se han incluido los parámetros `col` para el color de la línea y `lwd` para el grosor.

5. Ejercicios propuestos

Resolver los siguientes ejercicios, leyendo previamente el fichero encuesta.csv:

```
encuesta = read.csv2("encuesta.csv")
```

Se pide realizar un análisis de variable estadística VIAJE, que representa el tiempo (en minutos) en llegar a la universidad, convirtiendo previamente los valores de minutos a horas con dos dígitos decimales,

```
viaje=encuesta$VIAJE
viajeH=round(viaje/60,digits=2)
```

Calcular para dicha variable (en horas):

1. Medidas de centralización: media, mediana, moda.
2. Medidas de dispersión: rango, desviación estándar, varianza y coeficiente de variación.
3. Medidas de localización: cuartiles y los percentiles 33 y 66.
4. Medidas de forma: coeficiente de asimetría y coeficiente de apuntamiento o curtosis.
5. Diagrama de caja ¿Hay algún dato atípico?
6. Tabla de frecuencias para los datos agrupados en 10 clases (intervalos)
7. Media para los datos agrupados en esas 10 clases
8. Histograma y diagrama de tarta para esas 10 clases.

6. Referencias recomendadas

- Estadística con R (J.C. Soage)
- Statology (Z. Bobbitt)
- Elementary Statistics with R (C. Yau)
- Estadística descriptiva: representaciones gráficas (D. Molina, A.M. Lara)