Estadística

Tema 1. Estadística descriptiva

María Dolores Frías Domínguez
Jesús Fernández Fernández
Carmen María Sordo

Departamento de Matemática Aplicada y Ciencias de la Computación

TEMA1: Estadística descriptiva

Tablas

Estadísticos

Gráficos

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Estadística descriptiva

Se ocupa del análisis de **muestras** de datos procedentes de experimentos, encuestas etc, que contienen una componente aleatoria no predecible.

POBLACIÓN: todos los estudiantes de la Universidad de Cantabria.

MUESTRA: alumnos de 1º de la Universidad de Cantabria.

Estadística descriptiva

Los datos que estudiamos pueden ser de diferentes tipos:

Atendiendo a su naturaleza:

Cualitativas, se dividen en categorías no numéricas (sexo de los individuos, fumadores o no...)

Semi-cuantitativas, valores no numéricos pero que admiten clasiricación (calidad de un servicio: malo, regular, bueno)

Cuantitativas, son numeros reales (edad, altura...). Estas a su vez pueden ser discretas si toman un número finito o numerable de valores (edad) y continuas si toman un número infinito de valores dentro de un cieerto intervalo (altura y peso).

Atendiendo al número de observaciones:

Unidimensionales, bidimensionales, multidimensionales.

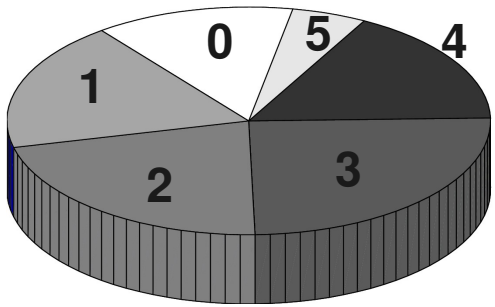
Estadística descriptiva

Encuesta a 60 familias de una ciudad sobre el número de hijos:

0	2	2	4	0	3	3	2	5	2	3	2	4	3	4
3	1	4	1	1	0	4	1	1	4	2	4	2	0	3
1	3	0	5	2	2	3	0	3	0	5	1	1	4	0
3	2	3	2	3	3	1	2	4	2	3	1	3	1	4

Hijos	n_i
0	8
1	11
2	13
3	15
4	10
5	3

Tablas



Gráficos

media=2.283
varianza=2.005
desv.stand.=1.416
moda=3
mediana=2

Estadísticos

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Tabla de frecuencias

Una tabla de frecuencias resume la información contenida en los datos de una muestra. Las columnas de la tabla muestran distintas variables dependiendo de si los datos son discretos o continuos.

Caso discreto (con pocos valores posibles):

x_i: posibles valores que pueden aparecer en los datos

n_i: frecuencia absoluta. Número de ocurrencias en la muestra de cada posible valor

f_i: frecuencia relativa $$f_i = \frac{n_i}{n}$$

N_i: frecuencia absoluta acumulada $$N_i = \sum_{j \leq i} n_j$$

F_i: frecuencia relativa acumulada $$F_i = \sum_{j \leq i} f_j = \frac{1}{n} \sum_{j \leq i} n_j$$

Tabla de frecuencias

Ejemplo

En una encuesta a 60 familias de una ciudad sobre el número de hijos.

0	2	2	4	0	3	3	2	5	2	3	2	4	3	4
3	1	4	1	1	0	4	1	1	4	2	4	2	0	3
1	3	0	5	2	2	3	0	3	0	5	1	1	4	0
3	2	3	2	3	3	1	2	4	2	3	1	3	1	4

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

R tip

```
ni <- table(data)
Ni <- cumsum(ni)
fi <- mitabla/length(ni)
Fi <- cumsum(ni)/length(ni)
```

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Ejercicio

En una obra se han ido anotado el número de metros que los albañiles azulejan por hora, obteniéndose la tabla de frecuencias siguiente:

Metros	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Albañiles	*5*	*7*	*12*	*10*	*11*	*6*	*3*	*3*	*2*	*1*

Completar esa tabla de frecuencias.

Tabla de frecuencias

Una tabla de frecuencias resume la información contenida en los datos de una muestra.

Caso continuo (o discreto con muchos valores posibles): Los datos han de agruparse por **clases**.

$(L_{i-1}, L_i]$: límites de clase. Valor inferior y superior del intervalo que define las clases

x_i: marcas de clase. Valor medio de los límites de clase.

n_i: frecuencia absoluta. Número de ocurrencias en la muestra de cada posible valor

f_i: frecuencia relativa $f_i = \frac{n_i}{n}$

N_i: frecuencia absoluta acumulada $N_i = \sum_{j \leq i} n_j$

F_i: frecuencia relativa acumulada $F_i = \sum_{j \leq i} f_j = \frac{1}{n} \sum_{j \leq i} n_j$

Tabla de frecuencias

Ejemplo

El tiempo de acceso al disco duro (milisegundos) medido en 30 instantes de tiempo distintos ha sido:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1.000	

El **Criterio de Sturges** nos dice cuántas clases definir:

$$c = \left\lfloor \frac{3}{2} + \frac{log(n)}{log(2)} \right\rfloor$$

$$c = \left\lfloor \frac{3}{2} + \frac{log(30)}{log(2)} \right\rfloor = \lfloor 6.41 \rfloor = 6$$

6.2	7.2	5.3	4.4	6.3	9.1	6.4	9.9	6.7	8.7
5.5	8.4	6.9	4.1	8.5	7.3	8.5	7.2	9.1	4.4
7.3	8.8	5.8	7.5	4.4	7.8	6.9	6.1	8.2	6.6

Ejercicio

En un cierto colectivo de personas se toma una muestra de 30 personas a las que se observa el peso, obteniéndose los siguientes datos:

$$57.2, 92.5, 72.8, 74.8, 60.1, 96.1, 74.3, 89.1, 69.2, 77.7,$$
$$65.0, 82.1, 66.2, 51.3, 83.9, 71.3, 84.8, 62.5, 103.2, 64.1,$$
$$73.1, 87.3, 58.9, 76.1, 45.8, 79.1, 68.9, 62.5, 81.5, 65.7$$

Representar este conjunto de datos mediante una tabla, agrupando los datos por clases.

Estadísticos

Cualquier función de los datos de la muestra, por lo que solo se definen para datos cuantitativos (valores numéricos).

$$T(x_1, x_2 ... x_{1n})$$

Sirven para cuantificar ciertas características de la muestra:

Estadísticos de tendencia central o localización
Estadísticos de posición
Estadísticos de dispersión
Estadísticos de forma

Estadísticos de posición

Indican valores que parten la muestra en proporciones dadas: cuantiles, percentiles, cuartiles y deciles.

Todos ellos tienen las unidades de la variable observada.

Cuantil de orden α (C_α): Se define para cualquier valor α entre 0 y 1 que verifique:

$$\sum_{\forall i | x_i \leq C_\alpha} f_i \geq \alpha \qquad \sum_{\forall i | x_i \geq C_\alpha} f_i \geq 1-\alpha$$

Ejemplo

Alturas (cm): 160, 165, 172, 174, 174, 176, 179, 180, 180, 180, 180, 187

$C_{0.5} = Med = [176, 179] \rightarrow (176+179)/2 = 177.5$ cm

$C_{0.5}$ deja por debajo al 50% de los datos y por encima al 50%.

Estadísticos de posición

Cuantil de orden α (C_α): Para datos agrupados se calcula como:

$$C_\alpha = L_{i-1} + a_i \frac{\alpha n - N_{i-1}}{n_i}$$

α orden del cuantil
i intervalo que contiene al cuantil
L_{i-1} limite inferior del intervalo i
a_i amplitud del intervalo i
n_i frecuencia absoluta del intervalo i
N_{i-1} frecuencia absoluta acumulada del intervalo i

Ejemplo

¿$C_{0.5}$?

$$0.5 \times 31 = 15.5$$

$[L_{i-1}, L_i)$	x_i	n_i	N_i
$[5, 10)$	7.5	3	3
$[10, 15)$	12.5	6	9
$[15, 20)$	17.5	13	22
$[20, 25)$	22.5	7	29
$[25, 30)$	27.5	2	31

Puntuaciones test

$$C_{0.5} = 15 + 5\frac{0.5 \times 31 - 9}{13} = 17.5 \text{ ptos}$$

Estadísticos de posición

Percentil de orden 100α: Es el cuantil de orden α

Deciles: Son los cuantiles de orden $C_{0.1}$ $C_{0.2}$ ..... $C_{0.8}$ $C_{0.9}$

Cuartiles (Q): Dividen a la muestra en 4 grupos con frecuencias similares.

Primer cuartil $Q_1 = C_{0.25}$ = Percentil 25

Segundo cuartil $Q_2 = C_{0.50}$ = Percentil 50 = Mediana

Tercer cuartil $Q_3 = C_{0.75}$ = Percentil 75

Ejemplo

Alturas (cm): 160, 165, 172, 174, 174, 176, 179, 180, 180, 180, 180, 187

$C_{0.25}$ = [172,174] cm → (172+174)/2 = 173 cm

$C_{0.5}$ = *Med* = 177.5 cm $C_{0.75}$ = 180 cm

Estadísticos de localización

Indican valores con respecto a los que los datos parecen agruparse: media, mediana y moda.

Todos ellos tienen las unidades de la variable observada.

Media: Es la media aritmética (promedio) de los datos

Datos sin agrupar:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Suma de los valores dividido por el tamaño de la muestra

Ejemplo

Alturas de 5 personas en metros: 1.72 1.65 1.60 1.84 1.58

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1.72 + 1.65 + 1.60 + 1.84 + 1.58}{5} = 1.68 \text{ m}$$

Estadísticos de localización

Media: Es la media aritmética (promedio) de los datos.

Datos agrupados:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{c} x_i n_i = \sum_{i=1}^{c} x_i f_i$$

La media es un estadístico muy sensible a valores extremos.

Encuesta a 60 familias sobre el número de hijos:

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

Ejemplo

$$\bar{x} = \frac{0\cdot 8 + 1\cdot 11 + 2\cdot 13 + 3\cdot 15 + 4\cdot 10 + 5\cdot 3}{60}$$

$$\bar{x} = 2.28 \text{ hijos}$$

Estadísticos de localización

Mediana: Valor que divide a los datos en dos grupos con el mismo número de elementos. Es el Q_2 y el $C_{0.50}$

La mediana es un estadístico robusto ya que no es sensible a valores extremos.

Ejemplo

{1,4,6,10,12} → Mediana = 6

{1,4,6,10,30} → Mediana = 6

{1,4,6,8,10,12} → Mediana = (6+8)/2=7

Estadísticos de localización

Ejemplo

Número de hijos de 60 parejas estudiadas:

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

60/2 = 30 → **la mediana del número de hijos es 2 hijos**

Estadísticos de localización

Moda: Es el valor que más se repite, el de mayor frecuencia relativa o absoluta.

Clase Modal: Es el clase que tiene mayor frecuencia relativa por unidad de amplitud.

Ejemplo

1 3 5 5 7 10 → **5**

1 3 5 5 7 7 10 → **5 y 7 (bimodal)**

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1.000	

[6-7) clase modal

Ejercicio

En un cierto colectivo de personas se toma una muestra de 30 personas a las que se observa el peso, obteniéndose la siguiente tabla:

$(L_{i-1} - L_i]$	x_i	n_i	N_i	f_i	F_i
45-55	50	2	2	0.067	0.067
55-65	60	7	9	0.233	0.300
65-75	70	9	18	0.300	0.600
75-85	80	7	25	0.233	0.833
85-95	90	3	28	0.100	0.933
95-105	100	2	30	0.067	1.000

a) Calcular la media, la mediana y la clase modal.

b) Calcular el valor del peso que puede considerarse indicativo de anormalmente alto y bajo (representativo del 5% de la población con mayor y menor peso, respectivamente).

Estadísticos de dispersión

Ejemplo

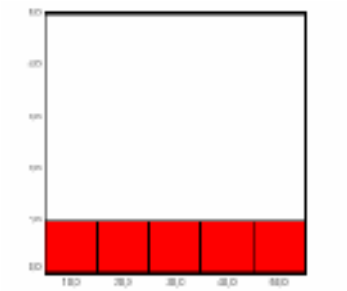
Conjunto 1: 10 20 30 40 50 → media=30 mediana=30 moda=no tiene

Conjunto 2: 10 30 30 30 50 → media=30 mediana=30 moda=30

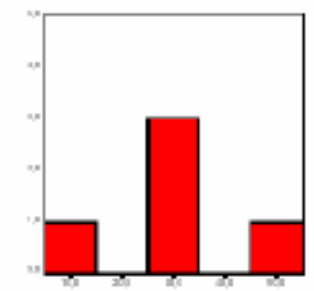
Conjunto 3: 30 30 30 30 30 → media=30 mediana=30 moda=30

Sin embargo los datos son totalmente distintos!!

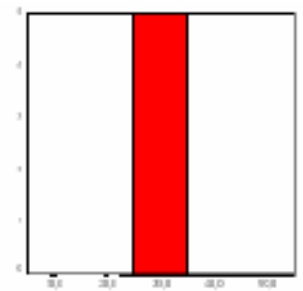
Conjunto 1



Conjunto 2



Conjunto 3



Los estadísticos de localización no caracterizan completamente los datos son necesarios los **estadísticos de dispersión**.

Estadísticos de dispersión

Indican la mayor o menor concentración de los datos con respecto a las medidas de localización: rango, rango intercuartílico, varianza, cuasi-varianza, desviación típica, cuasi-desviación típica y coeficiente de variación.

Rango: Diferencia entre el máximo y el mínimo. Muy sensible a valores extremos.

Rango intercuartílico (RIC): Diferencia entre el tercer y el primer cuartil.

$$RIC = C_{0.75} - C_{0.25}$$

Ambos tiene las mismas unidades que la variable.

Ejemplo

Alturas (cm): 160, 165, 172, 174, 174, 176, 179, 180, 180, 180, 180, 187

Rango=187-160=27 cm

RIC=$C_{0.75}$- $C_{0.25}$ = 180-173 = 7 cm

Estadísticos de dispersión

Varianza (s^2_n): Unidades de la variable al cuadrado

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{c}(x_i - \bar{x})^2 n_i \qquad S_n^2 = \overline{x^2} - \bar{x}^2$$

Cuasi-varianza (s^2): Unidades de la variable al cuadrado

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{c}(x_i - \bar{x})^2 n_i = \frac{n}{n-1}S_n^2$$

Desviación típica (s_n): Unidades de la variable.

$$S_n = \sqrt{S_n^2}$$

Cuasi-desviación típica (s): Unidades de la variable

$$S = \sqrt{S^2}$$

Todos son sensibles a valores extremos.

Estadísticos de dispersión

Ejemplo

El tiempo de acceso al disco duro (milisegundos) medido en 30 instantes de tiempo distintos ha sido:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1.000	

Estadísticos de dispersión

Ejemplo

El tiempo de acceso al disco duro (milisegundos) medido en 30 instantes de tiempo distintos ha sido:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i	$x_i n_i$	$x_i^2 n_i$
4-5	4.5	4	4	0.133	0.133	18	81
5-6	5.5	3	7	0.100	0.233	16.5	90.75
6-7	6.5	8	15	0.267	0.500	52	338
7-8	7.5	6	21	0.200	0.700	45	337.5
8-9	8.5	6	27	0.200	0.900	51	433.5
9-10	9.5	3	30	0.100	1.000	28.5	270.75
Total		30		1.000		211	1551.5

$$S_n^2 = \overline{x^2} - \bar{x}^2 = \frac{1551.5}{30} - \left(\frac{211}{30}\right)^2 = 2.25 \text{ ms}^2$$

$$S_n = \sqrt{S_n^2} = \sqrt{2.25} = 1.5 \text{ ms}$$

Estadísticos de dispersión

Coeficiente de variación (CV): Razón entre la cuasi-desviación típica y la media.

$$CV = \frac{S}{\bar{x}}$$

También se denomina variabilidad relativa y es frecuente usarla en porcentaje.

Es adimensional, por lo que resulta interesante para comparar la variabilidad de variables diferentes.

Ejemplo

Si el peso de los individuos de una muestra tiene CV=30% y la altura CV=10% los individuos presentan más dispersión en peso que en altura.

Estadísticos de forma

Momentos de orden r (m_r):

Se llama momento muestral m_r de orden *r*, respecto de una constante *a*, a la siguiente medida:

$$m_r = \frac{1}{n}\sum_{k=1}^{n}(x_k - a)^r$$

Cuando a=0 se habla de **momentos respecto del origen.**

Si a=$\bar{x}$ se dice que son **momentos centrales.**

La media muestral es el momento de primer orden (r=1) respecto del origen (a=0).

La varianza es el momento muestral de segundo orden (r=2) respecto de la media (a=$\bar{x}$)

Estadísticos de forma

Dan idea de la forma de la distribución: coeficiente de asimetría o sesgo y coeficiente de curtosis o apuntamiento. Son adimensionales.

Coeficiente de asimetría o sesgo (C_A): Indica si la distribución es simétrica o no.

$$C_A = \frac{m_3}{S_n^3} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \bar{x}}{S_n}\right)^3$$

C_A=0, la distribución es simétrica (media = mediana)
C_A>0, la distribución es asimétrica por la derecha
C_A<0, la distribución es asimétrica por la izquierda

Coeficiente de curtosis o apuntamiento (C_C): Indica el grado de apuntamiento de la distribución con respecto a distribución normal o gaussiana.

$$C_C = \frac{m_4}{S_n^4} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \bar{x}}{S_n}\right)^4 - 3$$

C_c=0, distribución mesocúrtica (Normal)
C_c>0, distribución letpcúrtica o apuntada
C_c<0, distribución platicúrtica o aplanada

Ejercicio

En un cierto colectivo de personas se toma una muestra de 30 personas a las que se observa el peso, obteniéndose la siguiente tabla:

$(L_{i-1} - L_i]$	x_i	n_i	N_i	f_i	F_i
45-55	50	2	2	0.067	0.067
55-65	60	7	9	0.233	0.300
65-75	70	9	18	0.300	0.600
75-85	80	7	25	0.233	0.833
85-95	90	3	28	0.100	0.933
95-105	100	2	30	0.067	1.000

Calcular la cuasi-desviación típica, la varianza, el rango intercuartílico, el coeficiente de variación, el coeficiente de asimetría y el de curtosis.

Gráficos

Los gráficos son una herramienta de resumen de la información contenida en los datos que permiten sacar conclusiones acerca de la muestra de un solo vistazo.

Veremos distintos tipos de gráficos, algunos de los cuales dependen del tipo de variable: si es discreta o continua o si es cuantitativa o cualitativa.

- Diagrama de sectores
- Gráfico de barras
- Histograma
- Diagrama de cajas

Gráficos

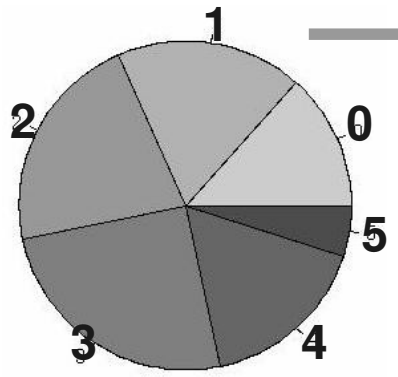
Diagrama de sectores: Es una representación circular o con forma de tarta en la que cada sector del círculo tiene un ángulo directamente proporcional a la frecuencia relativa de cada posible valor de la variable.

Está indicado para variables cualitativas o discretas con un número pequeño de posibles valores.

Ejemplo

Encuesta a 60 familias de una ciudad sobre el número de hijos:

Hijos	n_i
0	8
1	11
2	13
3	15
4	10
5	3



1 hijo: 11 x 360/60=66º

R tip

```
pie(table(data))
```

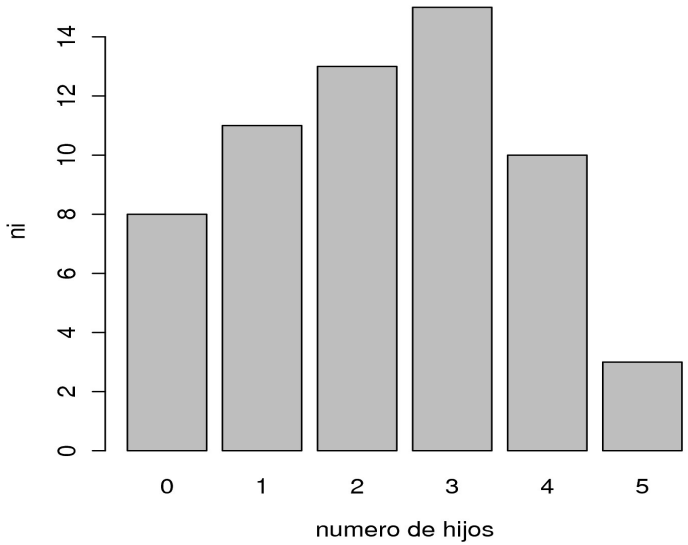
Gráficos

Diagrama de barras: Representa mediante barras la información contenida en la tabla de frecuencias, ya sea la frecuencia absoluta o la relativa.

Está indicado para variables cualitativas o discretas con un número pequeño de posibles valores.

Ejemplo

Encuesta a 60 familias de una ciudad sobre el número de hijos:



Hijos	n_i
0	8
1	11
2	13
3	15
4	10
5	3

R tip

```
barplot(table(data), xlab="numero de
   hijos", ylab="ni")
```

Gráficos

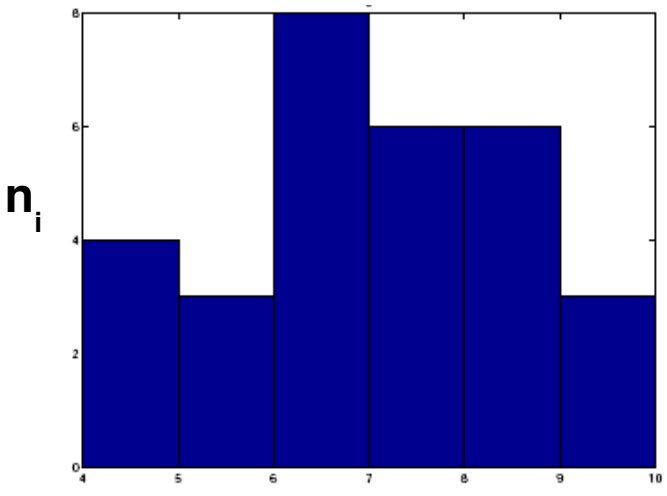
Histograma de frecuencias: Muestran la distribución de una serie de datos de variables cuantitativas continuas o agrupadas en intervalos de clase.

Se trata de un gráfico de barras verticales en el que el ancho de cada barra corresponde con el rango del intervalo mientras que la altura respresenta la frecuencia absoluta o relativa.

Ejemplo

El tiempo de acceso al disco duro (milisegundos) medido en 30 instantes de tiempo distintos ha sido:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1.000	

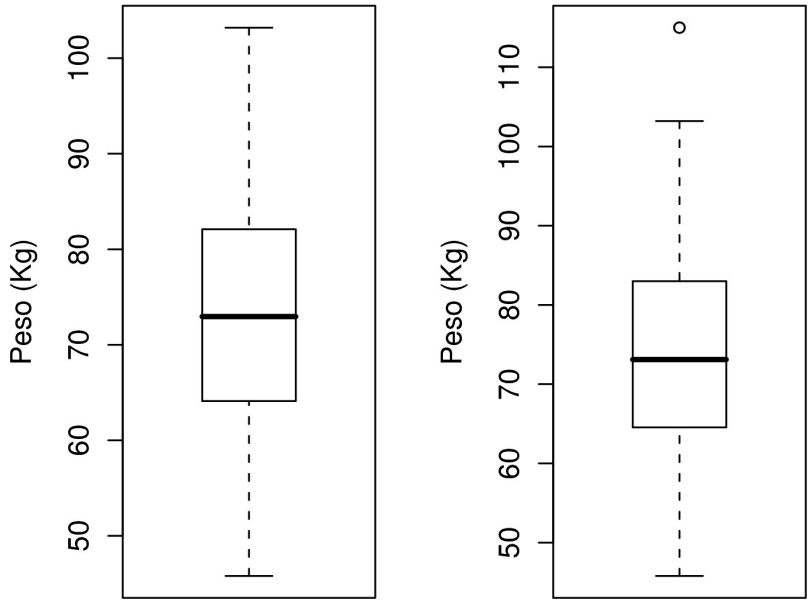


R tip

```
Hist(data,scale="frequency",breaks="Sturges",
   col="darkgray",xlab="tiempo",ylab="ni")
```

Gráficos

Diagrama de cajas o box and wiskers: Resumen gráficamente 5 datos: máximo, mínimo, $C_{0.25}$, $C_{0.5}$ y $C_{0.75}$



R tip

```
boxplot(data,ylab='Peso (Kg)')
```

La zona central (caja) contiene el 50% de las observaciones (RIC).

Los outliers son datos anómalos que se representan fuera de los bigotes. Son valores mayores que Q_3+1.5RIC o valores menores Q_1-1.5RIC.

Ejercicio

Jaime llevaba toda la tarde analizando los datos de altura de un grupo de personas (en centímetros) y ya tenía listo su diagrama de cajas. Lamentablemente, se le ha derramado un cafe corrosivo sobre él y ha borrado parte del diagrama. Ayúdale a dibujarlo de nuevo con los datos que había recogido. Viendo el diagrama, ¿podrías decir si los datos presentan asimetría?

$(L_{i-1}, L_i]$	n_i
(155,160]	3
(160,165]	6
(165,170]	18
(170,175]	21
(175,180]	12
(180,185]	42
(185,190]	12
(190,195]	3