Estadística climatológica con R

2. Estadística descriptiva básica

2.1 Introducción y conceptos fundamentales de la estadística descriptiva

En general, los investigadores de cualquier índole se ven ante la necesidad de obtener conclusiones a partir del análisis de mucha información. ***La estadística*** a partir del uso de herramientas que van desde el álgebra elemental, hasta el cálculo diferencial, logra la sistematización de dicha información de manera de hacer fácilmente distinguibles aspectos que de otra manera podrían pasar inadvertidos.

Estadística es también el diseño de experimentos, la elección de muestras, la reducción y el proceso de datos y muchas otras cuestiones.

La estadística proporciona instrumentos para la toma de decisiones cuando prevalecen condiciones de ***incertidumbre***.

La estadística como conocimiento puede presentar diferentes aspectos, pero hay dos principales, uno en el que solo se hace una recopilación y presentación de datos sin otro fin que su ordenación, y otro en que haciendo análisis de los datos se llega a deducir leyes o previsiones aplicadas al colectivo motivo de la estadística:

1. *Estadística* ***descriptiva***, que está relacionada con la organización, el resumen y la descripción de los datos
2. *Estadística* ***inferencial o inductiva***, relacionada con el proceso de utilizar datos para tomar decisiones.

Como estado intermedio está el *Análisis Estadístico o Estadística Analítica*.

El proceso de tomar decisiones en situaciones generales, sobre la base de una información incompleta contenida en datos muestrales, es arriesgado y no puede realizarse con certeza; La ***probabilidad*** es una medida de esta incertidumbre. Hay dos tipos de incertidumbre con los que tenemos que enfrentarnos:

- la incertidumbre debida a la aleatoriedad
- la incertidumbre debida a nuestra ignorancia del verdadero estado del sistema.

La colección de toda la información posible que caracteriza un fenómeno determinado se denomina ***población*** o ***universo***. Estadísticamente, ***la población*** se define como un conjunto de ***individuos*** o de ***objetos*** que poseen una o varias características comunes.

Los términos población e individuo deben de ser entendidos como colectivo del que queremos sacar conclusiones y como elemento o unidad que compone la población respectivamente.

La definición de lo que constituye *la población* depende del experimentador y de la naturaleza del problema que se investiga. Una vez definida, de ella se tomaran las observaciones y se deberán sacar conclusiones. Es fundamental fijar la población con toda precisión, ya que solamente la obtención de una muestra representativa de la población permitirá obtener conclusiones fiables sobre ella.

Unos ejemplos de poblaciones pueden ser los tornillos de un tipo determinado que fabrica una empresa o el conjunto de pacientes que van a ser sometidos a un tratamiento. En climatología pueden ser las temperaturas a una hora determinada en un cierto lugar o las precipitaciones anuales ocurridas en una estación.

Lo ideal sería contar con todas las observaciones que componen el universo, para poder observar el comportamiento de una característica relevante. Esto es factible siempre que se trabaje con poblaciones o universos "manejables". En general las poblaciones son de grandes dimensiones y requeriría un muchísimo tiempo, en otros casos el estudio de un elemento es destructivo y no podemos hacer un análisis de la población. Otro problema es que a veces las poblaciones son infinitas con lo cual examinar la población entera es imposible.

Por lo tanto, normalmente nos conformaremos con un conocimiento parcial de la población. Lo conseguiremos observando unos cuantos elementos de la población y viendo como es o cuánto vale esa característica que nos interesa. Este conjunto de elementos que observaremos es lo que llamaremos ***una muestra*** de la población. La muestra debe ser *representativa* de la población y sus elementos *escogidos al azar* para asegurar la objetividad de la investigación.

Una muestras pueden ser, en relación con los ejemplos anteriores de poblaciones: un conjunto de los tornillos fabricados en la fábrica, un número determinado de los pacientes a los que se ha aplicado el tratamiento, la temperatura a la misma hora y lugar o las precipitaciones anuales durante 30 años.

Habitualmente la muestra representativa se obtendrá por un *procedimiento aleatorio* (es decir, *al azar*), lo cual permitirá medir y controlar los posibles errores en términos de probabilidades.

Normalmente la mejor garantía de que la muestra es representativa es el procedimiento al azar o que la población sea suficientemente *homogénea*.

Habitualmente la situación que se plantea es la de un investigador que desea estimar el valor de alguna característica de la población en estudio, como por ejemplo la estatura media de la población de un país, o el determinar el intervalo en el que, con gran probabilidad, se encuentre dicha característica.

La situación pues que se presenta, es la de una característica o valor poblacional objeto de investigación, al que denominaremos ***parámetro poblacional*** o simplemente parámetro, estando este asociado a ***una variable*** en estudio. En el ejemplo anterior asociado a la variable *estatura*, el parámetro en estudio era *la estatura media*. Otro parámetro podría ser *la temperatura media diaria* de una cierta localidad, asociado a la variable temperatura. Esta variable en estudio se corresponde con lo que matemáticamente se denomina ***variable aleatoria*** X o simplemente ***variable***.

Las variables pueden ser de dos tipos ***cualitativas*** y ***cuantitativas***. Variables ***cualitativas*** son aquellas que solo pueden clasificadas en categorías no numéricas. Un ejemplo puede ser el apellido de las personas de una ciudad, la facultad en la que están matriculados los estudiantes de una universidad, el tipo de meteoro atmosférico que ha afectado a una localidad, etc.

Una variable X es ***cuantitativa*** cuando toma valores numéricos. Dentro de las variables cuantitativas se suelen distinguir dos tipos:

- ***Discretas***: solo pueden tomar un conjunto finito o numerable de valores (generalmente enteros). Por ejemplo, el número de páginas de un libro en una biblioteca, el número de días de lluvia o de halo, el número de días con precipitación superior e inferiores a una cantidad, etc.
- ***Continuas***: pueden tomar cualquier valor en un intervalo (finito o infinito). Por ejemplo, el tiempo de duración de una bombilla, la temperatura, precipitación, altura de las nubes.

Utilizaremos la siguiente notación tanto para variables discretas como continuas:

n: tamaño de la muestra = número de elementos observados

$x_1, \ldots, x_n$: representan los *n* valores de la variable estadística obtenidos en la muestra (puede haber repeticiones)

La matriz de datos

Habitualmente, la información primaria sobre los individuos se representa en una matriz, en la que aparecen en la primera columna los individuos y en la siguiente columna las observaciones de los diferentes caracteres o variables en estudio para cada uno de los individuos, de la siguiente forma

	Carácter 1	*Carácter 2*	...	*Carácter p*
Individuo 1	•	•	...	•
Individuo 2	•	•	...	•
....	...	...	...	
Individuo n	•	•	...	•

Cuando en el marco de una investigación se estudia el comportamiento de una sola variable por separado, se dice que el estudio es ***unidimensional***. A veces interesa conocer cómo se comporta una variable en conjunto con otra u otras (simultáneamente). En dicho caso se habla de estadísticas ***multidimensionales***.

Vamos a centrarnos primero en el caso unidimensional y veremos ***las distribuciones unidimensionales de frecuencias***.

2.2 Distribuciones de frecuencias

Agrupación de Datos

La estadística se preocupa de la ordenación y sistematización de datos, para poder apoyar el proceso de toma de decisiones. La forma en que se ordenan los datos varía de acuerdo al tipo de variable de que se trata.

a) Estadísticas de variable discreta

Ejemplo 2.1

Consideremos por ejemplo los datos del número de pacientes diarios de un hospital. Tenemos los siguientes datos

40	42	45	38	45
41	42	45	42	41
39	37	35	40	43
40	44	35	42	41
37	38	37	42	41
39	43	45	45	37
40	41	43	35	45

Donde el total de observaciones de la variable es 35, luego el tamaño de la muestra *n=35*.

Tal como están los datos no dan mucha información. Vamos a proceder primero a ordenar la información.

La variable *X* es el número de pacientes atendidos en un día determinado en un hospital.
Cada una de las observaciones la designaremos por x_i.

Si ordenamos la observaciones de menor a mayor obtendremos

35	35	35	37	37
37	37	38	38	39
39	40	40	40	40
41	41	41	41	41
42	42	42	42	42
43	43	43	44	45
45	45	45	45	45

Es fácil obtener ahora las siguientes conclusiones: la variable asume un total de 10 valores y el menor es 35 y el mayor 45.

Cada uno de estos valores que adopta la variable recibe el nombre de ***clase*** y se denota por ***m***; El número de clases es 10 por haber 10 valores distintos de la variable en la muestra.

Construiremos el cuadro de distribución de frecuencias para esta variable.

Valores de la variable	**Frecuencias absolutas**	**Frecuencias relativas**	**Frecuencias absolutas acumuladas**	**Frecuencias relativas acumuladas**
X_i	n_i	h_i	N_i	H_i
$X_1 = 35$	$n_1 = 3$	$h_1 = 0{,}0857$	$N_1 = 3$	$H_1 = 0{,}0857$
$X_2 = 36$	$n_2 = 0$	$h_2 = 0$	$N_2 = 3$	$H_2 = 0{,}0857$
$X_3 = 37$	$n_3 = 4$	$h_3 = 0{,}1143$	$N_3 = 7$	$H_3 = 0{,}2$
$X_4 = 38$	$n_4 = 2$	$h_4 = 0{,}0571$	$N_4 = 9$	$H_4 = 0{,}2571$
$X_5 = 39$	$n_5 = 2$	$h_5 = 0{,}0571$	$N_5 = 11$	$H_5 = 0{,}3143$
$X_6 = 40$	$n_6 = 4$	$h_6 = 0{,}1143$	$N_6 = 15$	$H_6 = 0{,}4286$
$X_7 = 41$	$n_7 = 5$	$h_7 = 0{,}1429$	$N_7 = 20$	$H_7 = 0{,}5714$
$X_8 = 42$	$n_8 = 5$	$h_8 = 0{,}1429$	$N_8 = 25$	$H_8 = 0{,}7143$
$X_9 = 43$	$n_9 = 3$	$h_9 = 0{,}0857$	$N_9 = 28$	$H_9 = 0{,}80$
$X_{10} = 44$	$n_{10} = 1$	$h_{10} = 0{,}0286$	$N_{10} = 29$	$H_{10} = 0{,}8286$
$X_{11} = 45$	$n_{11} = 6$	$h_{11} = 0{,}1714$	$N_{11} = 35$	$H_{11} = 1$
	$\Sigma\, n_i = 35 = \mathbf{n}$	$\Sigma\, h_i = 1 = 100\%$		

La primera columna contiene los valores que adoptan la variable o las clases. Aunque en ninguno de los días se atendieron 36 pacientes, para la correcta tabulación la tabla debe de contener esta clase.

La segunda columna de la tabla recibe el nombre de ***frecuencias absolutas***; corresponden a las repeticiones que tiene cada valor de la variable para el ejemplo. La nomenclatura de la frecuencia absoluta es $\boldsymbol{n_i}$. Así para la primera clase $n_1=3$, o sea que durante tres días se atendió a *35* pacientes.

Como es lógico, la suma de todas las frecuencias absolutas debe de ser igual al número de observaciones n.

La frecuencia absoluta tiene las siguientes propiedades:

- $0 \leq n_i \leq n$
- $\sum n_i = n$

Muchas veces es más conveniente presenta la información en términos porcentuales. Para ellos se calcula la ***frecuencia relativa*** h_i que viene definida como

$$h_i = \frac{n_i}{n}$$

mostradas en la tercera columna. Así se puede afirmar que el 8,57% de los días el consultorio atiende 35 consultas.

La frecuencia relativa tiene las siguientes propiedades

- $0 \leq h_i \leq 1$
- $\sum h_i = 1$

Una pregunta que puede surgir al analizar los datos, es la cantidad de observaciones que son mayores o menores que un determinado valor de la variable; para ello se calculan las ***frecuencias absolutas acumuladas***. La frecuencia absoluta acumulada de una clase determinada se calcula sumándole a su frecuencia absoluta, las frecuencias absolutas de todas las clases anteriores. En la tabla anterior la frecuencia absoluta acumulada de la clase 4ª es 9 y proviene de la suma de su frecuencia absoluta más todas las anteriores.

La frecuencia relativa acumulada se calcula con el mismo razonamiento anterior pero usando frecuencias relativas en lugar de absolutas.

Calculo con R

```
> x <- c(35,36,37,38,39,40,41,42,43,44,45)  o  x <- 35:45      # se introducen los valores.
> fa <- c(3,0,4,2,2,4,5,5,3,1,6)  # se introducen la frecuencias absolutas
# Si quisiésemos el vector con todos los datos haríamos:
        > datos <- rep(x,fa)   # repite cada valor el número de veces dado en fa
> n <- sum(fa)   # suma de frecuencias absolutas = tamaño de la muestra
> fr <- (fa/n) # frecuencias relativas
> fr <- round(fr,4) # si queremos redondear a 4 decimales
> faac <- cumsum (fa)  # frecuencias absolutas acumuladas
> frac <- cumsum (fr)  # frecuencias relativas acumuladas
Si queremos mostrar la tabla de frecuencias en pantalla (se crea una matriz por columnas)
        > cbind(x,fa,fr,faac,frac)
```

Ejercicio 2.1

El número de días de heladas (temperatura mínima < 0ºC) en la estación de Jaén desde el año 1989 al 2022 fueron:

0, 1, 3, 3, 2, 6, 3, 2, 1, 1, 6, 0, 2, 0, 8, 2, 9, 3, 3, 0, 6, 10, 1, 11, 2, 1, 2, 2, 1, 2, 3, 2, 4, 0

- Constrúyase una tabla de distribución de frecuencias que represente estos datos. (Primero hay que hacer el recuento de datos y obtener las frecuencias absolutas)

b) Estadísticas de variable continua

Las variables continuas también necesitan ser ordenadas, para ser analizadas adecuadamente.

Ejemplo 2.2

Para este caso pondremos otro ejemplo, consideraremos las subvenciones anuales entregadas a ciertas empresas expresadas en millones de euros. Se obtienen los siguientes datos:

8,5	9,3	10,4	6,6	7,9
7,6	8,4	9,3	7,6	10,1
13,3	12,1	10,6	12,4	11,5
11,3	13,1	13,5	14,6	12,3
9,4	7,8	8,3	9,1	7,9
10,2	11,3	11,7	12,8	13,4
11,3	12,1	13,6	14,1	13,2
14,4	11,8	13,8	11,3	10,6

La tabla de distribución de frecuencias para esta muestra es básicamente la misma. La diferencia es que al ser una variable continua se tienen muchos más datos distintos y por lo tanto una enorme cantidad de clases. Es conveniente definir intervalos. (También para las variables discretas se podrían definir intervalos). Esto tiene más sentido antiguamente, cuando no existían los ordenadores, por simplicidad de cálculo.

Intervalos	**Marcas de clase**	**Frecuencias absolutas**	**Frecuencias relativas**	**Frecuencias absolutas acumuladas**	**Frecuencias relativas acumuladas**
$\mathbf{X_{i-1}\text{-}X'_i}$	**Xi**	$\mathbf{n_i}$	$\mathbf{h_i}$	$\mathbf{N_i}$	$\mathbf{H_i}$
6-6,99	6,5	1	0,025	1	0,025
7-7,99	7,5	5	0,125	6	0,15
8-8,99	8,5	3	0,075	9	0,225
9-9,99	9,5	4	0,1	13	0,325
10-10,99	10,5	5	0,125	18	0,45
11-11,99	11,5	7	0,175	25	0,625
12-12,99	12,5	5	0,125	30	0,75
13-13,99	13,5	7	0,175	37	0,925
14-15	14,5	3	0,075	40	1
		$\Sigma\, n_i = 40 = \mathbf{n}$	$\Sigma\, h_i = 1 = 100\%$		

Para la correcta tabulación de los datos en intervalos conviene seguir una serie de reglas:

a) Buscar el valor menor y mayor de la variables
b) Escoger una amplitud de los intervalos *c* adecuada. Debe de ser la misma, en la medida de lo posible, para todos los intervalos
c) Para facilitar la tabulación debe de procurarse que el límite inferior de los intervalos sea un numero entero.

La interpretación de las columnas de esta tabla es la misma que para el caso de la variable discreta. La única diferencia está en la primera y la segunda columna. La primera tiene el encabezado $\mathbf{X_{i-1}\text{-}X_i'}$ donde X_{i-1} representa el límite inferior del intervalo i-esimo y X_i' representa el límite superior de cada intervalo.

La segunda columna recibe el nombre de columnas de ***marcas de clase***. La marca de clase es el punto medio del intervalo y representa a ese intervalo o clase. (Se obtiene sumando los límites de cada intervalo (redondeando el número superior) y dividiendo por 2)

Un número conveniente de intervalos k de amplitud constante puede obtenerse según la fórmula de *Sturges*

$$k = 1 + 3{,}322\ log_{10}\ n$$

siendo *n* el número total de datos.

Una vez determinado el numero *k* de intervalos a considerar, y si es posible tomarlos de igual amplitud, esta será

$$c = \frac{x_n - x_1}{k}$$

Donde x_n es el dato mayor y x_1 el menor.

2.3 Representación gráfica de las distribuciones de frecuencias

La representación gráfica de una distribución de frecuencias depende del tipo de datos que la formen. Aquí nos centraremos en distribuciones unidimensionales de frecuencias.

Datos de un carácter cualitativo

La representación de estos datos está basada en la proporcionalidad de las áreas a las frecuencias absolutas o relativas. Las representaciones más usadas son el diagrama ***de Sectores*** y el ***de Rectángulos***.

Ejemplo 2.3

Consideremos el caso de una encuesta sobre deportes favoritos de la gente. Se llegó a la siguiente tabla.

Deportes	n_i	f_i	N_i	F_i
Futbol	8	0,40	8	0,40
Tenis	5	0,25	13	0,65
Atletismo	2	0,10	15	0,75
Baloncesto	3	0,15	18	0,90
Otros	2	0,10	20	1
	20	**1**		

El diagrama de sectores o de tarta consiste en dividir un círculo en tantos sectores como modalidades presente el carácter cualitativo, en este caso los deportes, asignando a cada sector un ángulo o área proporcional a la frecuencia absoluta.

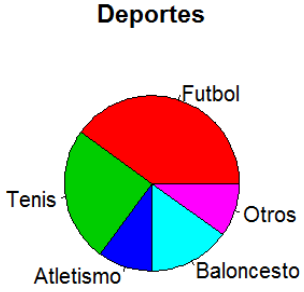
Esta representación es muy fácil en R.

```
> x <- c(8,5,2,3,2)      # se introducen en un vector los valores de las frecuencias absolutas.
> pie (x)                # se pinta el grafico de sectores
```

Así ya se plotea el grafico, pero asigna los colores al azar y en vez de representar el nombre de la clase pone un número. Vamos a crear un vector de nombres y otro de colores.

```
> nombres <- c('Futbol', 'Tenis', 'Atletismo', 'Baloncesto', 'Otros')   # vector nombres
> colores <- c(2,3,4,5,6)                  # vector de colores
> pie (x, labels=nombres, col=colores, main='Deportes')     # se pinta de nuevo el grafico de sectores.
```

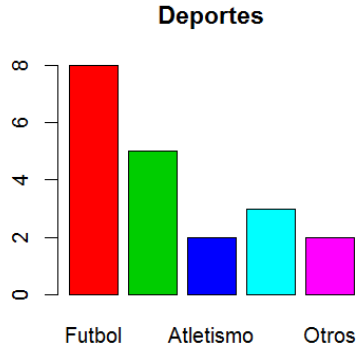
En el grafico hemos un título principal (***main***). Esta propiedad y ***col*** son comunes a cualquier grafico en *R*. Así obtendríamos



El diagrama de barras o rectángulos consiste en construir tantos rectángulos como modalidades presente el carácter cualitativo en estudio. La altura se toma igual a la frecuencia absoluta o relativa (según queramos).

Se crea en R con la función ***barplot*** en lugar de ***pie***. La diferencia es que en vez de usar ***labels*** para los nombres usa ***names***:

> barplot (x, names=nombres, col=colores, main='Deportes')



Ejercicio 2.2

Se analizó el IVA general en distintos países de Europa obteniéndose la siguiente tabla:

PAIS	%
Alemania	19
Belgica	21
Dinamarca	25
España	21
Francia	20
Italia	22
Luxemburgo	15
Suecia	24

- Constrúyase un diagrama de sectores y otro de barras

Datos correspondientes a un carácter cuantitativo

La agrupación por intervalos no se utiliza actualmente porque implica una pérdida de información, ya que todos los datos del intervalo son considerados iguales a la marca de clase. Se emplea para representar los datos en los gráficos, pero en cálculos de estadísticos, como la media, mediana, etc., se incluyen los datos en bruto. Antiguamente se usaba mucho por la dificultad de manejar un gran número de datos, pero hoy en día con el uso de ordenadores no es necesario. Vamos a ver la forma de representar una distribución de frecuencias de una variable cuantitativa.

La representación habitual es un *Histograma* en donde para cada intervalo se levanta un rectángulo con un área igual a la frecuencia absoluta o relativa. Hay que tener en cuenta si los intervalos tienen igual o distinta amplitud. En *R* se consigue con la función ***hist***.

Ejemplo 2.2 (Continuación)

Vamos a aplicarlo al ejemplo 2.2 de las subvenciones a empresas que reproducimos aquí, donde los datos están agrupados en intervalos.

Intervalos	Marcas de clase	Frecuencias absolutas	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
X_{i-1}-X'_i	Xi	n_i	h_i	N_i	H_i
6-6,99	6,5	1	0,025	1	0,025
7-7,99	7,5	5	0,125	6	0,15
8-8,99	8,5	3	0,075	9	0,225
9-9,99	9,5	4	0,1	13	0,325
10-10,99	10,5	5	0,125	18	0,45
11-11,99	11,5	7	0,175	25	0,625
12-12,99	12,5	5	0,125	30	0,75
13-13,99	13,5	7	0,175	37	0,925
14-15	14,5	3	0,075	40	1
		$\Sigma n_i = 40 = \mathbf{n}$	$\Sigma h_i = 1 = 100\%$		

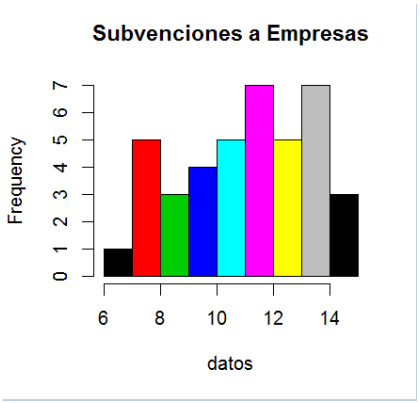
El proceso para representarlo en R seria

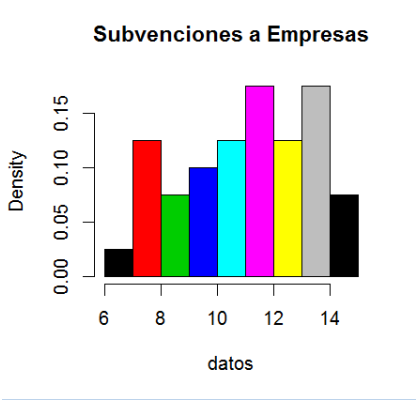
```
> m1 <- c(6.5,7.5,8.5,9.5,10.5,11.5,12.5,13.5,14.5)     #  marcas de clase
> n <- c(1,5,3,4,5,7,5,7,3)     #  frecuencias absolutas
> datos <- rep(m1, n)     #  repetimos las marcas de clase tantas veces como indique la frecuencia absoluta obteniendo así los datos a representar
 > datos
  [1]  6.5  7.5  7.5  7.5  7.5  7.5  8.5  8.5  8.5  9.5  9.5  9.5  9.5 10.5 10.5
 [16] 10.5 10.5 10.5 11.5 11.5 11.5 11.5 11.5 11.5 11.5 12.5 12.5 12.5 12.5 12.5
 [31] 13.5 13.5 13.5 13.5 13.5 13.5 13.5 14.5 14.5 14.5
> d1 <- c(6,7,8,9,10,11,12,13,14,15)     #  indicamos los puntos de corte de los intervalos
> colores <- c(1,2,3,4,5,6,7,8,9)       # colores
> hist(datos, breaks=d1, col=colores, main='Subvenciones a Empresas')                #  histograma de frecuencias absolutas
```

Si queremos el histograma de frecuencias relativas añadimos ***prob=T***

```
> hist(datos, breaks=d1, col=colores, prob=T, main='Subvenciones a Empresas')
```

Si suprimimos la opción ***breaks***, R ajustaría los intervalos.





Ejercicio 2.3

Un fabricante de neumáticos ha recopilado información sobre la cantidad de miles de km recorridos por un modelo concreto de esos neumáticos hasta que se ha producido un pinchazo o un reventón de neumático. Los concesionarios han proporcionados los siguientes datos:

52.452	50.432	37.748	51.831	73.808	61.065	35.807	57.277
48.698	65.854	75.850	36.949	75.548	69.010	61.477	65.585
44.411	41.886	34.754	59.888	59.449	67.632	89.116	69.483
63.692	70.003	65.996	55.989	49.677	46.502	67.467	64.398
84.588	40.709	50.238	61.390	85.720	45.313	46.724	61.752
55.643	55.912	46.681	66.519	59.168	66.313	35.884	28.625
47.012	71.360	78.635	41.715	72.635	41.463	48.996	48.172
79.426	67.662	53.324	49.011	29.480	41.128	30.252	33.412
48.240	57.884	55.257	84.656	48.662	10.504	60.951	38.420
74.239	60.727	56.155	86.070	90.565	53.751	76.580	68.629
51.179	74.582	58.708	48.035	67.124	41.830	61.030	58.267
61.979	4.3068	41.539	62.215	51.269	82.919	34.182	37.654
80.502	35.342	44.719	37.402				

- Constrúyase una tabla de frecuencias para estos datos tomando como numero de intervalos el que proporciona la fórmula de Sturges. Interpretar la tabla.
- Dibujar el histograma de frecuencias relativas.

Para no tener que escribirlos los datos se pueden meter en *R* con:

```
datos <- c(52.452, 50.432, 37.748, 51.831, 73.808, 61.065, 35.807, 57.277, 61.979, 4.3068,
41.539, 62.215, 51.269, 82.919, 34.182, 37.654, 51.179, 74.582, 58.708, 48.035, 67.124,
41.830, 61.030, 58.267, 74.239, 60.727, 56.155, 86.070, 90.565, 53.751, 76.580, 68.629,
48.240, 57.884, 55.257, 84.656, 48.662, 10.504, 60.951, 38.420, 79.426, 67.662, 53.324,
49.011, 29.480, 41.128, 30.252, 33.412, 47.012, 71.360, 78.635, 41.715, 72.635, 41.463,
48.996, 48.172, 55.643, 55.912, 46.681, 66.519, 59.168, 66.313, 35.884, 28.625, 84.588,
40.709, 50.238, 61.390, 85.720, 45.313, 46.724, 61.752, 63.692, 70.003, 65.996, 55.989,
49.677, 46.502, 67.467, 64.398, 44.411, 41.886, 34.754, 59.888, 59.449, 67.632, 89.116,
69.483, 48.698, 65.854, 75.850, 36.949, 75.548, 69.010, 61.477, 65.585, 80.502, 35.342,
44.719, 37.402)
```

2.4 Medidas de tendencia central

Las medidas de tendencia central, de posición o promedios examinan propiedades de las distribuciones de frecuencias relacionadas con sus valores medios o más frecuentes y nos permiten comparar entre varias de ellas. Describiremos las más comunes.

Media aritmética

Llamando $x_1, x_2, \ldots, x_k$ a los datos distintos de un carácter cuantitativo en estudio, o las marcas de clase de los intervalos en los que se han agrupado los datos, y $n_1, n_2, \ldots, n_k$ a las correspondientes frecuencias absolutas de dichos valores o marcas de clase, llamaremos media aritmética de la distribución de frecuencias o de la muestra al valor

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

en donde *n* es la frecuencia total. Si los datos no están tabulados sería:

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

En R se calcula por medio de la función ***mean()***

```
> media <- mean(datos)  # media de los datos, la calcula R directamente si tenemos todos los datos en el vector
```

En el caso de agrupar los datos por intervalos y considerar la marca de clase la media calculada por la formula anterior difiere ligeramente de cuando calculamos la media con ***mean*** de todos los datos. Esta última es más exacta ya que en el caso anterior asignamos todos los datos de un intervalo al valor de la marca de clase.

Mediana

La mediana es otra medida de posición. Se define como aquel valor de la variable tal que, supuestos ordenados los valores de esta en orden creciente, la mitad son menores o iguales y la otra mitad mayores o iguales. La mediana divide a la población o muestra que se está investigando en dos partes iguales, con igual número de observaciones cada una.

Una vez ordenados, la mediana es el valor de la muestra que deja a izquierda y derecha el mismo número de observaciones. Tenemos dos posibilidades, si el número de observaciones es impar la mediana es el valor central mientras que si es par la mediana es el punto medio de los dos valores centrales.

Por ejemplo en la siguiente distribución de frecuencias

x_i	n_i	N_i
0	3	3
1	2	5
2	2	7
	7	

Ordenamos los valores en orden creciente
0 0 0 1 1 2 2

El **1** será el valor de la mediana.

En *R* la mediana se calcula con la función ***median()***

```
> x <- c(0,0,0,1,1,2,2)      #  valores (no hace falta que estén ordenados)
> mediana <- median(x)  # mediana de los datos
```

Más tarde veremos que al igual que con la mediana, se pueden encontrar valores de la variable que dividen a la población en 3, 4,…, 10 o cualquier cantidad que queramos. Se verán con los *cuantiles*.

Si los datos se distribuyen según una *distribución normal o de gauss*, que veremos en el capítulo 4, en forma de campana aproximadamente simétrica, la media es una medida óptima. A esta curva se aproximan más o menos fielmente las frecuencias o probabilidades de ocurrencia de los datos de las series de ***temperatura***, ***precipitación anuales, humedad***, ***presión, las componentes del viento por separado***, etc., cuando el número de observaciones crece indefinidamente, en especial cuando se refieren a observaciones o datos obtenidos a una misma hora del día, un mismo día del año en un mismo lugar.

Los datos de ***precipitación*** se distribuyen presentando una cola o rama de curva más larga hacia la derecha, al igual que ocurre con las ***temperatura máximas***, es decir tienen asimetría positiva. En el caso de las ***temperaturas mínimas*** lo es hacia la izquierda o asimetría negativa.

Para situaciones donde la distribución es asimétrica o distinta de la normal otras medidas tal como la mediana pueden ser más adecuadas. Las medidas que no son sensibles a asunciones particulares sobre la distribución de los datos son llamadas medidas *robustas*.

La media puede estar influenciada fuertemente por valores extremos, así que el uso de la mediana es también preferible si hay valores anómalos o valores extremos (*outliers*). Medidas que no son muy influenciadas por unos pocos valores extremos son conocidas como medidas *resistentes*.

La mediana es más robusta y resistente que la media, pero también puede mostrar una sorprendente sensibilidad a pequeños cambios en los datos.

Moda

La ***moda*** se define como aquel valor de la variable al que corresponde la máxima frecuencia (absoluta o relativa), o sea el valor más repetido de la muestra.

En el ejemplo anterior se ve a simple vista que la moda es 0.

Series o muestras con *una sola moda* son todas las que corresponden a datos de temperaturas tanto anuales como mensuales; a precipitaciones anuales; presiones anuales, etc.

Algunas distribuciones presentan *2 modas*, llamadas *bimodales*, son por ejemplo las muestras de precipitación o temperaturas anuales en zonas tropicales.

Ejemplo 2.3

Supongamos que tenemos unos datos sobre un estudio de población de una localidad aislada. Tenemos en concreto la edad de una muestra de individuos de la población.

Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad
1	22	10	29	19	34	28	42	37	59
2	22	11	29	20	35	29	44	38	66
3	23	12	29	21	35	30	44	39	67
4	24	13	29	22	35	31	45	40	69
5	25	14	29	23	36	32	45	41	69
6	25	15	31	24	38	33	45		
7	26	16	31	25	39	34	47		
8	27	17	32	26	39	35	48		
9	28	18	33	27	42	36	52		

Calculo con R:

```
# Primero introducimos los datos
> x <- 22:69 # valores o clases
> fa <-
c(2,1,1,2,1,1,1,5,0,2,1,1,1,3,1,0,1,2,0,0,2,0,2,3,0,1,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,1,0,2) #
frecuencias absolutas
> datos <- rep(x,fa) # vector con todos los valores
```

O directamente:

```
> datos <- c(22, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 29, 29, 29, 31, 31, 32, 33, 34, 35, 35, 35, 36, 38, 39, 39, 42, 42, 44, 44, 45, 45, 45, 47, 48, 52, 59, 66, 67, 69, 69)

> mean (datos)  # media de la muestra
> mean(datos)
[1] 38.26829
> median (datos)  # mediana de la muestra
> median(datos)
[1] 35
```

2.5 Medidas de dispersión

Las medidas de tendencia central de la sección anterior servían para resumir la distribución de frecuencias en un solo valor. *Las medidas de dispersión* tienen como propósito estudiar lo concentrada que esta la distribución en torno a algún promedio. Estudiaremos las medidas de dispersión más utilizadas.

Recorrido

Si x_{max} es el dato mayor, o la última marca de clase, si los datos vienen agrupados en intervalos, y x_{min} el dato menor o primera marca de clase, llamaremos *Recorrido* a

$$R = x_{max} - x_{min}$$

La principal ventaja del recorrido es el dar una medida rápida de la dispersión de los datos.

En R calcularíamos el recorrido como

```
> x <- c(22, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 29, 29, 29, 31, 31, 32, 33, 34, 35, 35, 35, 36, 38, 39, 39, 42, 42, 44, 44, 45, 45, 45, 47, 48, 52, 59, 66, 67, 69, 69)
                # datos o marca de clase
> recorrido <- max(x) - min(x)    #  máximo – mínimo
```

Otra opción es a través de ***range():***

```
> rango <- range(x)
> recorrido <- rango[2] - rango[1]
```

Varianza

Una idea que surge para cuantificar las dispersiones con respecto a la media es la de restar a cada valor de la variable la media aritmética previamente calculada y posteriormente calcular un promedio de estas diferencias. Si simplemente restamos a cada valor el valor de la media nos da una serie de valores que al sumar da 0 por la propia definición. Este problema se supera si elevamos al cuadrado los valores de las diferencias. Si hacemos el promedio dividiendo por el número de la muestra obtendremos *la varianza*.

Si denotamos por $x_1, x_2, \ldots, x_k$ los datos o las marcas de clase para datos agrupados en intervalos, llamaremos ***varianza*** denotada por σ^2

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

siendo $\bar{x}$ la media aritmética de la distribución.

Si los datos no están tabulados seria:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2$$

Así en el ejemplo anterior (2.3) la varianza

$$\sigma^2 = 1631{,}83 - (38{,}27)^2 = 167{,}37$$

En *R* lo haríamos así:

```
> datos <- c(22, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 29, 29, 29, 31, 31, 32, 33, 34, 35, 35, 35, 36, 38, 39, 39, 42, 42, 44, 44, 45, 45, 45, 47, 48, 52, 59, 66, 67, 69, 69)
> num.datos <- length(datos)  # numero de datos (41)
> varianza <- (1/num.datos) * sum(datos*datos) - mean(datos)**2
```

Cuando en lugar de considerar la serie completa o colectivo total, se considera una muestra del mismo, el parámetro es sustituido por ***s***, y en el denominados aparece *n-1* en vez de *n*; aunque si la muestra tiene un número grande de elementos es casi igual uno que otro.

Al valor s^2 se le suele denominar ***cuasivarianza***.

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

o para datos no tabulados

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

En *R* la cuasivarianza se calcula con la función ***var***

```
> var(datos)  #cuasivarianza
> var(datos)
[1] 171.5512
```

Desviación típica

La varianza esta expresada en unidades al cuadrado y a veces da una falsa imagen de dispersión. En su lugar se suele utilizar *la raíz cuadrada de la varianza*, llamada ***desviación típica σ*** para la población y ***cuasidesviacion típica s*** para la muestra (en inglés: *standard deviation* ***sd***)

Así en el ejemplo anterior

$$\sigma = \sqrt{\sigma^2} = 12{,}94$$

En *R* usaremos ***sd*** para obtener la cuasidesviación típica (raíz cuadrada de cuasivarianza).

Siguiendo el ejemplo anterior:

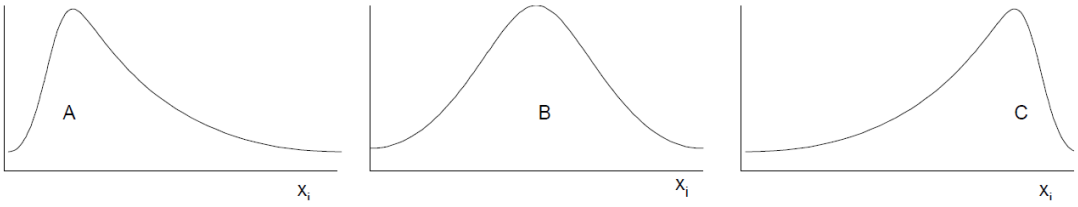
```
> desv.tipica <- sqrt(varianza)  # desviación típica (raíz cuadrada de la varianza)
> cuasi.varianza <- var(datos)  # cuasivarianza
> cuasi.desviacion.tipica <- sd(datos)  # cuasidesviación típica (raíz cuadrada de la cuasivarianza)
> cuasi.desviacion.tipica
[1] 13.09776
```

2.6 Simetría y apuntamiento

Hasta ahora hemos descrito el comportamiento de una variable describiendo en torno a que valores se concentra (medidas de tendencia central) y como se dispersa o se concentra alrededor de estos valores (medidas de dispersión). Ahora vamos a estudiar como la forma de la distribución, su simetría o lo apuntada o aplanada que sea.

Asimetría

Un aspecto muy importante al describir una variable es determinar *la simetría* (o *asimetría*) en el comportamiento de la misma. Observamos las siguientes representaciones de unas distribuciones de frecuencias



La distribución A tiene una 'cola' más larga hacia la derecha o hacia valores positivos de las variables, diremos que es una variable con *asimetría o sesgo positivo*. La C representa una distribución con *asimetría o sesgo negativo* y la B una distribución *simétrica*.

Otra forma de definir cuando una distribución es simétrica es cuando su mediana, su moda y su media aritmética coincidan.

Para ver ciertas propiedades de las distribuciones es muy útil la definición de momentos. ***Los momentos*** se definen como el valor medio que toma una magnitud o variable con relación al origen de medida, al cero, o respecto a un valor medio de los valores conocidos de la variable. Según el grado de esta relación los momentos pueden ser de primero, segundo,… o enésimo orden.

El momento de *primer orden* con respecto al origen es el valor medio de los datos

$$\mu_1 = \frac{\sum x_i}{n} = \bar{x} = E(x)(esperanza\ de\ la\ variable\ x)$$

y es la *media aritmética*, *esperanza matemática* o valor esperado de los valores de la variable, o sea de los datos.

En general, se define el *momento de orden r* de una variable con respecto a su media como:

$$\mu_r = \mathbf{E}\left(\mathbf{X} - \overline{\mathbf{X}}\right)^r = \frac{\sum\left(\mathbf{x} - \bar{\mathbf{x}}\right)^r}{\mathbf{n}}$$

El momento de segundo orden respecto a la media es *la varianza* σ^2

El tercer momento de una variable con respecto a su media cuantifica *la simetría* de una variable.

Así si $\mu_3 = 0$ sería una simetría perfecta, si $\mu_3 > 0$ estamos en el caso de asimetría o sesgo positivo y si $\mu_3 < 0$ tendríamos una asimetría o sesgo negativo.

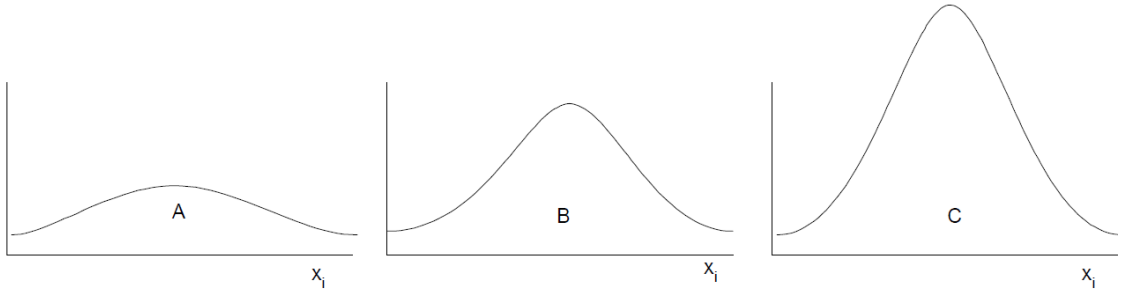
Se puede definir un tercer momento adimensional de la variable con respecto a su media, llamado *coeficiente de asimetría*, $\boldsymbol{\alpha_3}$,

$$\alpha_3 = \frac{\mu_3}{\sqrt[2]{\mu_2^3}}$$

Este momento tiene las mismas propiedades que μ_3 en cuanto la cuantificación de la asimetría.

Curtosis

La curtosis no es más que la cuantificación de lo puntiaguda o aplanada que puede ser la forma que representa el comportamiento de la variable. Se pueden dar los siguientes casos simplificados



La distribución A, la más plana, recibe el nombre de *platocúrtica*. La C, más puntiaguda, la llamaremos *leptocúrtica*. Y la B, intermedia entre las dos anteriores, *mesocúrtica o isocúrtica*. La Figura B corresponde a la distribución normal que veremos más adelante.

La curtosis de una variable se mide con el cuarto momento con respecto a la media.

$$\mu_4 = \mathbf{E}\left(\mathbf{X} - \overline{\mathbf{X}}\right)^4 = \frac{\sum\left(\mathbf{x} - \overline{\mathbf{x}}\right)^4}{\mathbf{n}}$$

El coeficiente de curtosis es el siguiente: $\alpha_4 = \frac{\mu_4}{\mu_2^2}$ de tal manera que por convención:

- Si $\alpha_4 = 3 \rightarrow$ mesocúrtica (normal)
- Si $\alpha_4 < 3 \rightarrow$ platocúrtica
- Si $\alpha_4 > 3 \rightarrow$ leptocúrtica

Mientras mayor sea α_4 mas leptocúrtica será la distribución de frecuencias.

Calculo con R

Para calcular la asimetría y achatamiento de una distribución a través de *R* se puede usar el paquete '***moments***' que facilita notablemente el cálculo de los momentos y coeficientes.

Primero habría que instalarlo la primera vez con:

```
> install.packages('moments ') # para instalarlo la primera vez
```

Se carga el paquete antes de usarlo

```
> library('moments')  # carga el paquete moments
```

Los momentos de orden k con respecto a su media se hallarían con

```
> moment(datos,order=k,central=T)
```

El coeficiente de asimetría lo hallaríamos con:

```
> coeficiente.asimetria <- skewness(datos)
```

El coeficiente de curtosis:

```
> coeficiente.curtosis <- kurtosis(datos)
```

Para los datos del ejemplo 2.3 sería:

```
> library('moments')  # carga el paquete moments

> datos <- c(22, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 29, 29, 29, 31, 31, 32, 33, 34, 35, 35, 35, 36, 38, 39, 39, 42, 42, 44, 44, 45, 45, 45, 47, 48, 52, 59, 66, 67, 69, 69)

> coeficiente.asimetria <- skewness(datos)
> coeficiente.asimetria
[1] 0.9927962
```

Al ser $\mu_3 > 0$ estamos en el caso de asimetría o sesgo positivo.

```
> coeficiente.curtosis <- kurtosis(datos)
>   coeficiente.curtosis
[1] 3.216295
```

El coeficiente de curtosis es casi 3 por lo que la curva es *normal*, aunque ligeramente *leptocúrtica* o más puntiaguda de la normal.

2.7 Percentiles, quintiles, cuartiles, etc.

Cuantiles es el nombre que reciben las divisiones de la población en partes iguales. Veremos algunos de los más usados, como cuartiles, quintiles o percentiles.

Al igual que la mediana la dividía la población en dos partes iguales, *los terciles* la dividen en 4 partes iguales, cada una de las cuales posee la cuarta parte de las observaciones. En este caso es necesario hacer 3 cortes, cada corte es un cuartil que representamos por Q_1, Q_2 y Q_3.

De la misma forma se definen *los quintiles*, que dividen los datos ordenados en cinco partes, *los deciles* (D_i), en diez y *los percentiles*, que dividen la población en 100 partes iguales (P_1,P_2, …, P_{99})

En R los cuantiles se calculan con la función ***quantile***.

Un gráfico muy útil con el que se puede visualizar fácilmente la dispersión y simetría de los datos, es *el diagrama de cajas* (*box-plot* inventado por Tukey en 1977), ejecutando en R la función de dibujo ***boxplot***. Esta función representa una caja vertical en donde el límite inferior es el primer cuartil, el límite superior es el tercer cuartil, y la caja aparece dividida por la mediana. Se añaden dos segmentos a la caja para para unirla al máximo y mínimo valor. Se llama recorrido o *rango intercuartílico* (***IQR*** en inglés) a la diferencia entre el tercer y el primer tercil de la distribución.

Al igual que la mediana, *el rango intercuartílico* es una medida que es resistente a la influencia de valores extremos y puede ser más apropiada que la desviación típica cuando la distribución es asimétrica.

Los datos inferiores al primer cuartil menos 1,5 veces el recorrido intercuartílico, o superiores al tercer cuartil más 1,5 veces el recorrido intercuartílico se consideran anómalos (outliers) y se representan por pequeños círculos fuera del diagrama de cajas.

Ejemplo 2.3 (continuación)

Siguiendo con el ejemplo 2.3 cuya tabla reproducimos

Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad
1	22	10	29	19	34	28	42	37	59
2	22	11	29	20	35	29	44	38	66
3	23	12	29	21	35	30	44	39	67
4	24	13	29	22	35	31	45	40	69
5	25	14	29	23	36	32	45	41	69
6	25	15	31	24	38	33	45		
7	26	16	31	25	39	34	47		
8	27	17	32	26	39	35	48		
9	28	18	33	27	42	36	52		

> edad <- c(22, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 29, 29, 29, 31, 31, 32, 33, 34, 35, 35, 35, 36, 38, 39, 39, 42, 42, 44, 44, 45, 45, 45, 47, 48, 52, 59, 66, 67, 69, 69)

Para calcular los cuantiles usamos la función ***quantile***. Con ella podemos calcular el valor de cualquier combinación de cuantiles, sean cuartiles, quintiles, etc. Por ejemplo el primer cuartil seria:
> quantile (edad, prob=0.25) # Se dividen la muestra en cuatro partes. Solo mostramos el primer cuartil
> quantile (edad, prob=c(0.25, 0.5, 0.75)) # Nos daría los 3 cuartiles (0,25,0,5 y 0,75)

Si solo ponemos ***quantile(edad)*** nos daría el mínimo, los tres cuartiles y el máximo

```
> quantile(edad)
  0%  25%  50%  75% 100%
  22   29   35   45   69
```

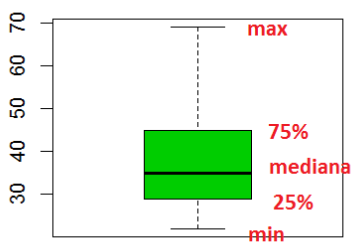
La mediana es el caso especial del cuantil 50%
> quantile (edad, prob=0.5) #mediana o segundo cuartil o cuantil 50%

Los quintiles los mostraríamos con
> quantile (edad, prob = c(0.2, 0.4, 0.6, 0.8)) # que dividen los datos en 5 partes
Los deciles:
> quantile (edad, prob = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)) # Deciles: 10 partes

Con la función de resumen ***summary*** obtenemos el mínimo, los cuartiles, la media y el máximo.
> summary(edad)

```
> summary(edad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.00   29.00   35.00   38.27   45.00   69.00
```

Si dibujamos el grafico de cajas ***> boxplot(edad, col=3)*** (col=32 color verde)



Nos muestra en el grafico la distribución de la muestra.

Ejercicio 2.4

Con los datos del ejercicio 2.3 calcular:

- Calcular la media aritmética, mediana. Comentar.
- Obtener la (cuasi) varianza y la desviación típica.
- Obtener el rango y los cuartiles de la distribución.
- Analizar la simetría y el apuntamiento de la distribución de frecuencias resultante. Calcular los coeficientes de asimetría y de curtosis mediante el paquete ***moments***.
- Optativo: Calcular los coeficientes de asimetría y de curtosis por las formulas.