

# Manual de Estadística Descriptiva con OpenOffice Calc

## Índice

<b>1. Tablas y gráficas estadísticas</b>	<b>1</b>
1.1. Utilización de la base de datos del INE . . . . .	18
<b>2. Cálculo de medidas de tendencia central</b>	<b>22</b>
<b>3. Cálculo de medidas de dispersión</b>	<b>33</b>
<b>4. Cálculo de las medidas de simetría y apuntamiento</b>	<b>39</b>
<b>5. Análisis bivariante</b>	<b>46</b>

## 1. Tablas y gráficas estadísticas

El cálculo de tablas de frecuencias y porcentajes así como su representación gráfica puede realizarse de manera sencilla con la ayuda de herramientas informáticas. Estudios estadísticos simples pueden realizarse mediante hojas de cálculo (tipo Microsoft Excel o OpenOffice Calc). Análisis más complejos requieren el uso de herramientas más sofisticadas, como el software especializado en estadística SPSS o R.

En esta sección aprenderemos a obtener tablas y gráficas mediante hojas de cálculo. Utilizaremos el programa OpenOffice Calc, que es la versión de software libre de hoja de cálculo. El programa puede obtenerse de forma gratuita de <http://es.openoffice.org/> y se instala fácilmente en cualquier sistema operativo. La versión utilizada en los siguientes ejemplos es la 3.0.

### Ejemplo 1

Consideramos los siguientes datos obtenidos de la web del Instituto Nacional de Estadística. Calcularemos la tabla de frecuencias y porcentajes y haremos varias representaciones gráficas.

Estadísticas judiciales 2005	
Estadística de lo Penal. Condenados. Resultados autonómicos	
Condenados según edad y sexo	
Unidades: nº de condenados	
Ambos sexos	
Baleares (Illes)	
De 18 a 20 años	155
De 21 a 25 años	543
De 26 a 30 años	653
De 31 a 35 años	619
De 36 a 40 años	515
De 41 a 50 años	636
De 51 a 60 años	248
De 60 y más	100

Fuente: Instituto Nacional de Estadística

Los pasos a seguir para calcular las tablas de frecuencias y porcentajes son los siguientes:

1. Abrir la aplicación OpenOffice Calc.

Se abrirá una ventana como la que se muestra en la figura 1.

2. Para introducir los datos del problema nos situamos sobre la casilla A1 (columna A, fila 1) moviéndonos con el cursor del ratón y escribimos en ella el título de la tabla: *Condenados Illes Balears por edad (año 2005)*. La fila 2 la dejamos en blanco para facilitar la lectura de la tabla.

A continuación escribimos en la casilla A3 *Edad (años)*, y en las posiciones inferiores de la misma columna: 18 – 20, 21 – 25, …, 51 – 60, *más de 60*. Para desplazarnos de una casilla a la siguiente podemos utilizar el ratón, las flechas del teclado o la tecla *Tab*.

Repetiremos la operación en la columna B. En la casilla B3 escribiremos *Nº condenados* y en las casillas inferiores: 155, 543, …, 100.

Si en algún momento deseamos rectificar alguno de los datos introducidos deberemos hacer doble clic sobre la casilla correspondiente y reintroducir el valor.

Después de este paso la hoja de cálculo tendrá el aspecto que se muestra en la figura 2.

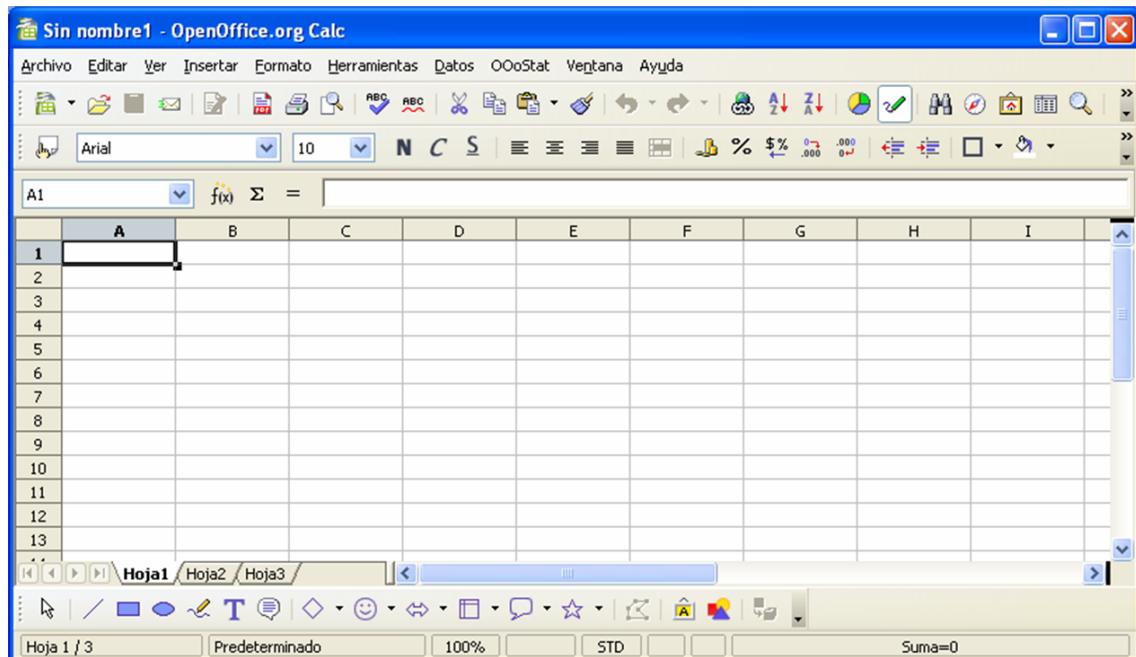


Figura 1: Ventana de inicio de OpenOffice Calc

	A	B	C	D
1	Condenados Illes Balears por edad (año 2005)			
2				
3	Edad (años)	Nº condenados		
4	18-20	155		
5	21-25	543		
6	26-30	653		
7	31-35	619		
8	36-40	515		
9	41-50	636		
10	51-60	248		
11	Más de 60	100		
12				

Figura 2: Hoja de cálculo trás la introducción de los datos del ejemplo 1

- Los valores de la columna B ( $n^o$  condenados) son las frecuencias absolutas de la variable *Edad*. Deseamos calcular las frecuencias relativas y los porcentajes. Además, como la variable *Edad* es cuantitativa podemos calcular también las frecuencias y porcentajes acumulados.

Empezamos por dar nombre a las columnas que mostrarán los valores calculados. Nos situamos sobre la casilla *C3* y escribimos *Frecuencia relativa*. Utilizando la tecla *Tab* o el ratón nos desplazaremos a la siguiente casilla de la misma fila (casilla *D4*) y escribiremos *Porcentaje*. Repitiendo el proceso escribiremos en las casillas *E5* a *H5* los valores: *Frecuencia absoluta acumulada*, *Frecuencia relativa acumulada* y *Porcentaje acumulado*.

Si el tamaño del texto escrito es mayor que la anchura de la columna el texto se sobreescibirá sobre las columnas vecinas. Para evitarlo podemos aumentar la anchura de las columnas situándonos sobre las líneas que separan las letras de las columnas



y desplazándolas con el cursor.

También podemos ajustar el texto automáticamente al tamaño de la columna situándonos sobre la columna a modificar y siguiendo los siguientes pasos: acceder a la opción *Formato* del menú principal, hacer clic sobre la opción *Celdas...*, se abrirá una nueva ventana en la que seleccionaremos la pestaña *Alineación* y haremos clic sobre la opción *Ajustar texto automáticamente* dentro del campo *Propiedades*.

Tras estos ajustes la fila 3 de la hoja de cálculo contiene los siguientes valores:

3	Edad (años)	Nº condenados	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
---	-------------	------------------	------------------------	------------	-------------------------------------	-------------------------------------	-------------------------

- Calcularemos primero la suma de los valores de frecuencias absolutas, es decir, el número total de condenados. Escribiremos este valor al final de la columna *B* (casilla *B12*). Para ello nos situaremos sobre esta casilla, escribiremos `=SUMA(B4:B11)` y pulsaremos la tecla *Enter*. El valor 3469 se mostrará en la casilla. La función *SUMA* es una función de Calc que permite sumar los valores de las casillas que se le indican (en nuestro caso desde la casilla *B4* hasta la *B11*).
- Para calcular las frecuencias relativas debemos dividir las frecuencias absolutas entre la suma de las frecuencias. Para ello nos situaremos sobre la casilla *C4*, escribiremos `=B4/$B$12` y pulsaremos *Enter*. En la casilla aparece el valor 0,04, resultado de dividir el valor de las casillas *B4* y *B12*.

Podemos repetir la operación con el resto de las casillas de la columna pero Calc ofrece una manera más sencilla de hacer estas operaciones. Basta situarnos con el cursor sobre la esquina inferior derecha de la casilla  $C4$ , hacer clic con el botón izquierdo del ratón y, manteniendo el botón pulsado, arrastrar el cursor hasta la casilla  $C11$ . Al soltar el botón aparecen en las casillas los valores calculados (ver columna  $C$  en la figura 3), ya que Calc reescribe automáticamente la fórmula de la primera casilla para adaptarla a las casillas seleccionadas.

6. Los porcentajes se obtienen multiplicando las frecuencias relativas por 100. Para ello nos situamos sobre la casilla  $D4$ , escribimos  $=C4*100$  y pulsamos *Enter*. A continuación nos situamos con el cursor en la esquina inferior derecha de la casilla y, manteniendo el botón izquierdo del ratón pulsado, arrastramos el cursor hasta la casilla  $D11$ . Al soltar el botón los resultados se escriben en las casillas correspondientes (ver columna  $D$  en la figura 3).
7. Las frecuencias absolutas acumuladas se calculan sumando a la frecuencia absoluta del valor considerado las frecuencias absolutas de los valores anteriores. La frecuencia absoluta acumulada del primer valor (18 – 20) es igual a su frecuencia absoluta, por lo que en la casilla  $E4$  escribimos  $=B4$  y pulsamos *Enter*. En la casilla siguiente,  $E5$ , escribimos  $=E4+B5$  y pulsamos *Enter*. Las restantes casillas se calcularán automáticamente si situamos el cursor en la esquina inferior derecha de la casilla  $E5$  y, manteniendo el botón izquierdo del ratón pulsado, arrastramos el cursor hasta la casilla  $E11$ . Al soltar el botón se muestran los valores calculados (ver columna  $E$  en la figura 3).
8. Las frecuencias relativas acumuladas se calculan dividiendo las frecuencias absolutas acumuladas entre la frecuencia absoluta total. Para ello escribimos en la casilla  $F4$   $=E4/\$B\$12$  y pulsamos *Enter*. Repitiendo el procedimiento explicado en los casos anteriores extendemos el cálculo hasta la casilla  $F11$  (el resultado se muestra en la columna  $F$  en la figura 3).
9. Finalmente, los porcentajes acumulados se obtienen multiplicando por 100 las frecuencias relativas acumuladas. Para ello escribimos  $=F4*100$  y pulsamos *Enter* en la casilla  $G4$ . Repitiendo el procedimiento explicado en los casos anteriores extendemos el cálculo hasta la casilla  $G11$ .

El tabla final obtenida se muestra en la figura 3.

10. Podemos imprimir la tabla calculada o guardarla como un fichero .pdf para su posterior impresión utilizando los iconos y respectivamente, del menú de Calc.

En todo caso, la visualización de la tabla mejora si separamos las filas y las columnas mediante líneas. Para ello, antes de imprimir o guardar el fichero seleccionaremos todas las casillas de la tabla situándonos sobre la casilla  $A3$  y, manteniendo pulsado el botón izquierdo del ratón, arrastrando el cursor

	A	B	C	D	E	F	G	
1	Condenados Illes Balears por edad (año 2005)							
2								
3	Edad (años)	Nº condenados	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado	
4	18-20	155	0,04	4,47	155	0,04	4,47	
5	21-25	543	0,16	15,65	698	0,2	20,12	
6	26-30	653	0,19	18,82	1351	0,39	38,94	
7	31-35	619	0,18	17,84	1970	0,57	56,79	
8	36-40	515	0,15	14,85	2485	0,72	71,63	
9	41-50	636	0,18	18,33	3121	0,9	89,97	
10	51-60	248	0,07	7,15	3369	0,97	97,12	
11	Más de 60	100	0,03	2,88	3469	1	100	
12		3469						
13								

Figura 3: Frecuencias y porcentajes obtenidos a partir de los datos del ejemplo 1

hasta la casilla G12. A continuación accederemos a la opción *Formato* del menú principal, haremos clic sobre la opción *Celdas...*, se abrirá una nueva ventana en la que seleccionaremos la pestaña *Borde* y haremos clic sobre el ícono .

Si imprimimos la tabla o visualizamos el fichero .pdf en la pantalla obtendremos el resultado de la figura 4:

Condenados Illes Balears por edad (año 2005)

Edad (años)	Nº condenados	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
18-20	155	0,04	4,47	155	0,04	4,47
21-25	543	0,16	15,65	698	0,2	20,12
26-30	653	0,19	18,82	1351	0,39	38,94
31-35	619	0,18	17,84	1970	0,57	56,79
36-40	515	0,15	14,85	2485	0,72	71,63
41-50	636	0,18	18,33	3121	0,9	89,97
51-60	248	0,07	7,15	3369	0,97	97,12
Más de 60	100	0,03	2,88	3469	1	100
	3469					

Figura 4: Tabla final del ejemplo 1

La tabla anterior puede incluirse fácilmente en informes escritos con Microsoft Word o OpenOffice Writer. Para ello seleccionaremos con el cursor todas las casillas que componen la tabla y utilizaremos la combinación de teclas *Ctrl+C*. La tabla queda copiada en el portapapeles de Windows y puede ser pegada en otros documentos mediante la combinación de teclas *Ctrl+V*.

A continuación explicamos como representar gráficamente los valores calculados

1. **Creación de un diagrama de barras.** Obtendremos en primer lugar un diagrama de barras que represente el número de condenados en función de su edad. Los pasos a seguir son los siguientes:

- Seleccionamos la opción *Insertar* del menú principal y hacemos clic sobre **Gráfico...**.
- Se abrirá una ventana titulada *Asistente de gráficos* (ver Figura 5). En primer lugar debemos seleccionar el tipo de diagrama, en nuestro caso un diagrama de barras con texto, por lo que hacemos clic sobre el icono . En la parte derecha de la ventana se elige una variante del diagrama de barras, nosotros nos quedamos con la opción por defecto (*Normal*) y pulsamos *Siguiente*.
- Se muestra una nueva ventana donde seleccionaremos las casillas de datos a representar. Para ello escribiremos **A3:A11;B3:B11** en *Rango de datos*.

Los datos de la primera columna seleccionada se representarán sobre el eje horizontal y los de la segunda sobre el eje vertical. Haciendo clic sobre el botón *Siguiente* pasamos a la siguiente ventana (*Series de datos*). Aplicamos las opciones por defecto por lo que volvemos a pulsar *Siguiente*.



Figura 5: Inserción de un diagrama de barras. Ventana inicial del *Asistente de gráficos*

- En la última ventana debemos añadir el texto de la gráfica. Escribimos el *Título del diagrama*: “Condenados Illes Balears (año 2005)”, los títulos de los ejes X e Y (“Edad” y “Nº condenados”, respectivamente) y desactivamos la opción *Mostrar leyenda*. Finalmente pulsamos el botón *Finalizar*.

El gráfico creado se muestra sobre la hoja de cálculo. Podemos variar su posición y tamaño mediante el ratón. El resultado final se muestra en la figura 6.

Al igual que para la tabla de frecuencias este diagrama puede imprimirse o bien guardarse como un fichero .pdf. Además, haciendo clic sobre el mismo y utilizando la combinación de teclas *Ctrl+C* es posible copiarlo en el portapapeles de Windows. De esta forma puede ser pegado fácilmente (combinación de teclas *Ctrl+V*) en un documento de Microsoft Word o OpenOffice Writer para la elaboración de un informe.

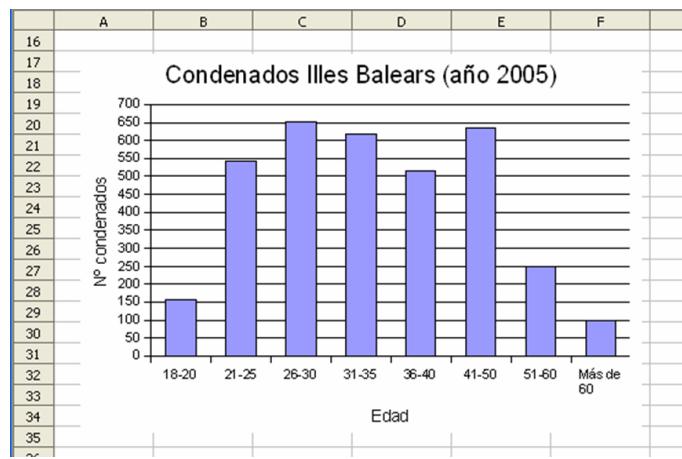


Figura 6: Diagrama de barras de frecuencias absolutas del ejemplo 1

2. **Creación de un diagrama de tarta.** Obtendremos a continuación un diagrama de tarta que represente los porcentajes de condenados para cada intervalo de edad. Los pasos a seguir son prácticamente idénticos a los del diagrama de barras, con las siguientes modificaciones
  - a) En la ventana inicial del *Asistente de gráficos* seleccionamos el tipo de diagrama haciendo clic sobre el icono
  - b) En la siguiente ventana el rango de datos es A3:A11;D3:D11.
  - c) Una vez creado el diagrama es posible cambiar el tamaño del texto de la leyenda haciendo doble clic sobre el diagrama, seleccionando la opción *Formato* del menú principal y a continuación la opción *Leyenda*. Aparece una nueva ventana en la que hay que escoger la pestaña *Caracteres* y el *Tamaño* deseado. El resultado final se muestra en la figura 7.
  - d) Este diagrama puede ser imprimido o insertado en otro documento al igual que el diagrama de barras.

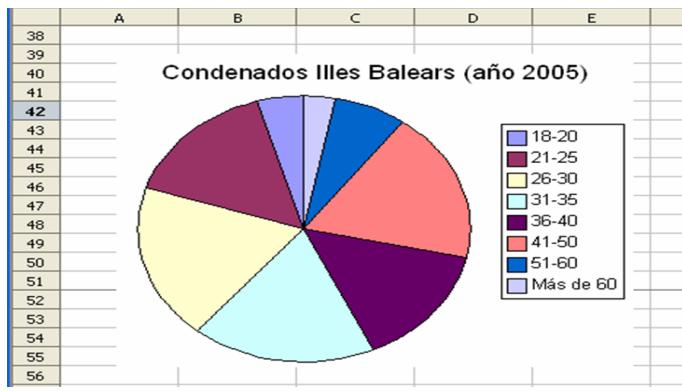


Figura 7: Diagrama de tarta de porcentajes del ejemplo 1

### Ejemplo 2

En este ejemplo aprenderemos a crear un diagrama de barras dobles a partir de los siguientes datos sobre población reclusa menor de edad en Baleares:

Edad (años)	Varón	Mujer
14	62	6
15	78	10
16	134	21
17	332	29

1. Introducimos los datos en una hoja de cálculo tal como se ha explicado para el ejemplo 1. Supongamos por ejemplo que los datos de *Edad* ocupan las casillas A4 a A7, los de *Varón* las casillas B4 a B7 y los de *Mujer* de C4 a C7 (las casillas A3, B3 y C3 contienen los títulos de las columnas).
2. Seguimos los pasos explicados para la creación de diagramas de barras en el ejemplo 1 pero seleccionando ahora las tres columnas de datos (rango de valores A3:A7;B3:B7;C3:C7). Las opciones a elegir son las mismas que en el caso del ejemplo 1 con la diferencia de que en la última ventana seleccionamos la opción *Mostrar leyenda* y que los títulos del diagrama y los ejes X e Y son, respectivamente: “Población reclusa menor de edad en Baleares”, “Edad” y “Nº reclusos”.

Al pulsar sobre el botón *Finalizar* obtenemos el resultado que se muestra en la figura 8.



Figura 8: Diagrama de barras dobles del ejemplo 2

### Ejemplo 3

En este ejemplo mostramos cómo calcular un histograma de frecuencias absolutas a partir de los datos siguientes sobre el peso de un grupo de personas:

Peso (Kg)	Frecuencia absoluta
45-49	20
50-54	35
55-59	40
60-64	55
65-69	45
70-74	50
75-79	35
80-84	30
85-89	25
90-94	15
95-99	5

1. En primer lugar creamos un documento OpenOffice Calc con los datos de la tabla anterior, tal como se ha explicado en el ejemplo 1. Supongamos que los datos sobre *Peso* ocupan las casillas *A4* a *A14* y los de frecuencia absoluta las casillas *B4* a *B14* (ver figura 9).
2. OpenOffice Calc no proporciona ninguna herramienta para la creación automática de histogramas en un caso general. Sólo en el caso de que todos los intervalos de valores sean de la misma amplitud (como en este ejemplo) es posible crear un histograma de manera sencilla.

En el caso del ejemplo todos los intervalos son de longitud 5 y podemos representar el histograma como un diagrama de barras modificado. En primer lugar debemos calcular la altura de las barras.

Sabemos que el área de las barras del histograma es igual al valor representado (en este caso la frecuencia absoluta). Por ejemplo, la primera barra debe tener área 20, como su anchura es 5 su altura deberá ser  $\frac{20}{5} = 4$ . Razonando de la misma manera podemos calcular el resto de alturas. Podemos hacerlo de forma automática con Calc: escribimos en la casilla  $C4 =B4/5$ , pulsamos *Enter* y a continuación extendemos el cálculo hasta la casilla  $C14$  utilizando el método explicado en el ejemplo 1. Al final de esta operación en la columna  $C$  aparecen los valores de altura calculados (ver figura 9).

	A	B	C	D
1				
2				
3	Peso (Kg)	Frecuencia absoluta	Altura barras histograma	
4	45-49	20	4	
5	50-54	35	7	
6	55-59	40	8	
7	60-64	55	11	
8	65-69	45	9	
9	70-74	50	10	
10	75-79	35	7	
11	80-84	30	6	
12	85-89	25	5	
13	90-94	15	3	
14	95-99	5	1	
15				

Figura 9: Tabla de datos del ejemplo 3

3. Ahora el histograma se puede calcular como un diagrama de barras. Seguimos el procedimiento descrito para el ejemplo 3 seleccionando las casillas  $A3$  a  $A14$  y  $C3$  a  $C14$  en el rango de datos. Los títulos del diagrama y del eje X son, respectivamente, “Histograma pesos” y “Pesos (Kg)” y la opción *Mostrar leyenda* no se selecciona.

Al crear el diagrama obtenemos un diagrama de barras con las barras separadas. Para unir las barras y darle la forma típica de un histograma debemos hacer doble clic sobre una de las barras del diagrama hasta que aparece la ventana que se muestra en la figura 10. Escogemos la pestaña *Opciones* y ponemos a 0% el valor de *Espacio* en la opción *Configuración*. Al pulsar sobre el botón *Aceptar* de esta ventana obtenemos un histograma como el que se muestra en la figura 11.

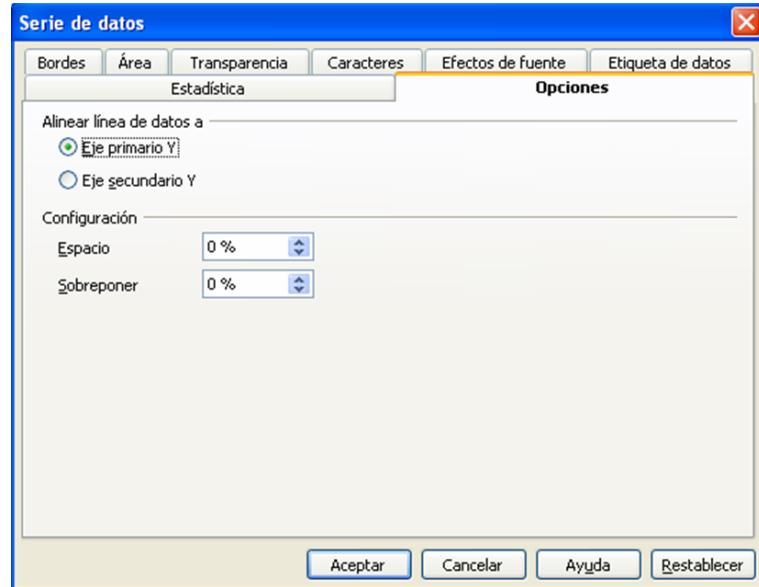


Figura 10: Ventana de diálogo para ajustar la anchura de las barras del diagrama de barras

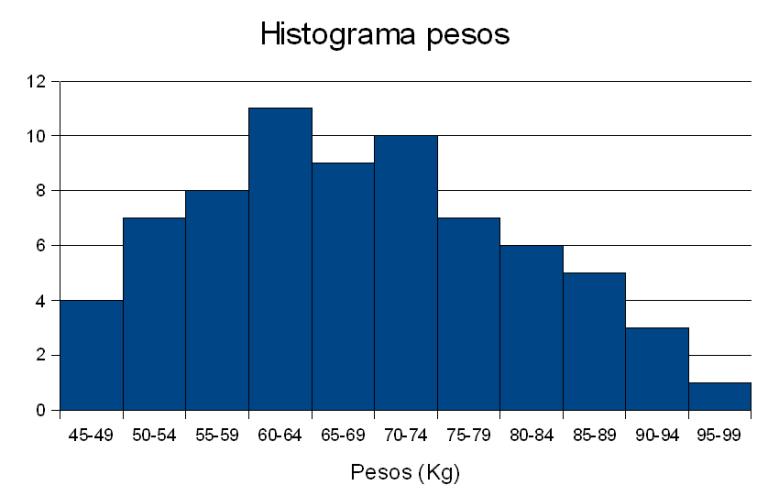


Figura 11: Histograma del ejemplo 3

#### Ejemplo 4

En este ejemplo mostramos cómo crear un diagrama lineal que representa la evolución del PIB español (en miles de millones de dólares) desde 1992 hasta 2007. Los datos proceden del Fondo Monetario Internacional.

Año	PIB
1992	612
1993	513
1994	515
1995	597
1996	622
1997	573
1998	601
1999	618

Año	PIB
2000	582
2001	609
2002	688
2003	885
2004	1045
2005	1131
2006	1231
2007	1414

1. En primer lugar creamos un documento OpenOffice Calc con los datos de la tabla anterior, tal como se ha explicado en el ejemplo 1. Supongamos que los datos sobre *Años* ocupan las casillas *A4* a *A19* y los de PIB las casillas *B4* a *B19* (en *A3* y *B3* se encuentran los nombres de las columnas).
2. Creamos un diagrama siguiendo el procedimiento explicado en anteriores ejemplos. Las casillas de datos a seleccionar son de *A3* a *A19* y *B3* a *B19*.

Seleccionamos el tipo de diagrama representado por el icono Línea y la variante representada por el icono .

No seleccionamos la opción de *Mostrar leyenda* y los títulos del diagrama y los ejes X e Y son, respectivamente, “Evolución PIB de España”, “Año” y “PIB (miles millones dólares)”. La gráfica obtenida se muestra en la figura 12.

#### Ejemplo 5

En todos los ejemplos anteriores hemos partido de datos de frecuencias absolutas a partir de las cuales hemos calculado frecuencias acumuladas, porcentajes, etc. Es habitual sin embargo disponer de datos *en bruto* que deben organizarse primero en tablas de frecuencias absolutas antes de realizar cualquier otro cálculo. En este último ejemplo explicamos cómo organizar este tipo de datos.

Partimos de los datos que se muestran en el siguiente gráfico (fuente LFP):

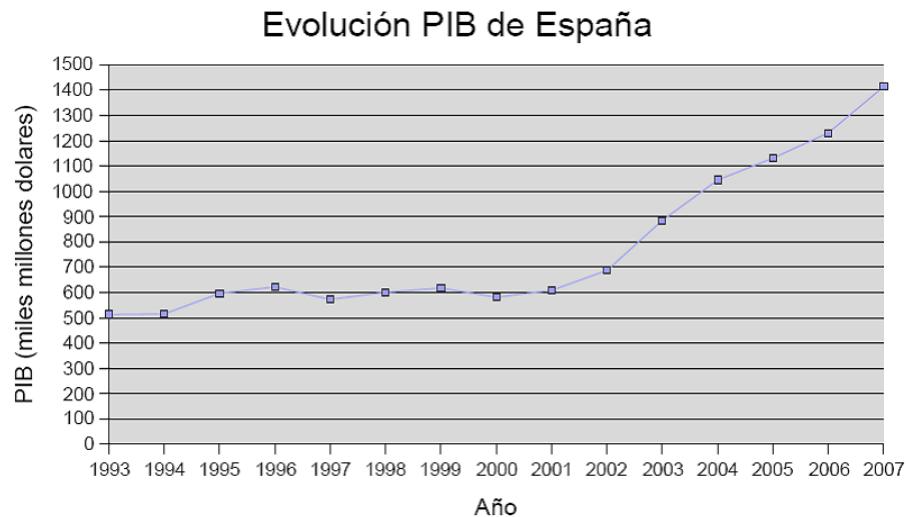
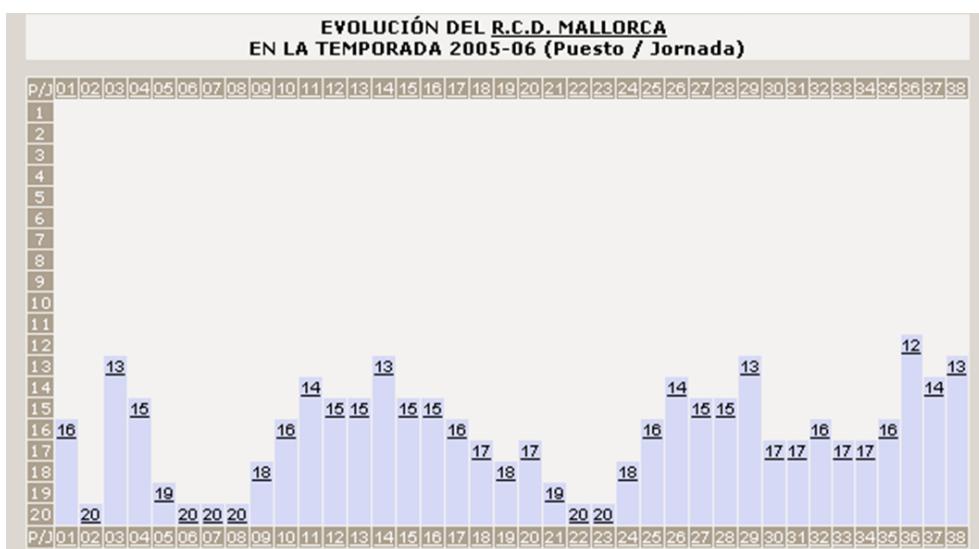


Figura 12: Diagrama lineal del ejemplo 4



Para la variable “clasificación del RCD Mallorca durante la temporada 2005-2006” deseamos calcular las frecuencias absolutas y acumuladas. Los pasos a seguir son los siguientes:

1. Creamos un documento OpenOffice Calc y escribimos en la primera columna los datos *brutos* del gráfico. Los datos ocupan las casillas A1 a A38. El resultado se muestra en la figura 13.
  2. A continuación creamos dos nuevas columnas en el documento (por ejemplo, las columnas B y C). En la parte superior de la primera columna escribimos el nombre de la variable (“Clasificación”) y a continuación escribimos, en orden creciente, los distintos valores que toma la variable. En la parte superior de

	A	B
1	16	
2	20	
3	13	
4	15	
5	19	
6	20	
7	20	
8	20	
9	18	
10	16	
11	14	
12	15	
13	15	
14	13	
15	15	
16	15	
17	16	
18	17	
19	18	
20	17	
21	19	
22	20	
23	20	
24	18	
25	16	
26	14	
27	15	
28	15	
29	13	
30	17	
31	17	
32	16	
33	17	
34	17	
35	16	
36	12	
37	14	
38	13	

	A	B	C
	12	Clasificación	Frecuencia absoluta
1		12	
2		13	12
3		13	13
4		13	14
5		13	15
6		14	16
7		14	17
8		14	18
9		15	19
10		15	20
11		15	
12		15	
13		15	
14		15	
15		15	
16		16	
17		16	
18		16	
19		16	
20		16	
21		16	
22		17	
23		17	
24		17	
25		17	
26		17	
27		17	
28		18	
29		18	
30		18	
31		19	
32		19	
33		20	
34		20	
35		20	
36		20	
37		20	
38		20	

Figura 13: Izquierdo: documento OpenOffice Calc con los datos *en bruto* del ejemplo 5 (paso 1). Derecha: documento preparado para el cálculo de las frecuencias absolutas (paso 2)

la segunda columna escribimos “Frecuencia absoluta”, que calcularemos a continuación.

Para facilitar la tarea de escribir en orden creciente los distintos valores de la variable podemos **ordenar** los valores *brutos* del siguiente modo:

a) Seleccionamos las casillas *A1* a *A38* manteniendo el botón izquierdo del ratón pulsado.

b) Pulsamos el icono . Los valores se ordenan de menor a mayor.

(Con la opción los valores se ordenarían de mayor a menor).

Ahora es sencillo ver qué valores toma la variable y escribirlos de forma ordenada en la columna *C*.

Al final de este paso el documento Calc tiene la forma que se muestra en la figura 13-derecha.

3. Para calcular las frecuencias absolutas nos situamos en la casilla *C2*, correspondiente a la frecuencia absoluta del valor 12. Escribimos

=CONTAR.SI(\$A\$1:\$A\$38; "="&B2)<sup>1</sup> y pulsamos *Enter*. Un 1 aparece escrito en la casilla, lo que significa que el valor 12 aparece un única vez en la lista de datos *brutos* (su frecuencia absoluta es 1). Extendemos el cálculo a las casillas *C3* a *C10* del siguiente modo: situamos el cursor en la esquina inferior derecha de la casilla *C5* y, manteniendo el botón izquierdo del ratón pulsado, arrastramos el cursor hasta la casilla *C10*. Al soltar el botón se mostrarán los valores calculados. Al final de este paso la hoja de cálculo tiene el aspecto que se muestra en la figura 14-izquierda.

B	C
Clasificación	Frecuencia absoluta
12	1
13	4
14	3
15	7
16	6
17	6
18	3
19	2
20	6

Clasificación	Frecuencia absoluta	Frecuencia acumulada
12	1	1
13	4	5
14	3	8
15	7	15
16	6	21
17	6	27
18	3	30
19	2	32
20	6	38

Figura 14: Izquierda: frecuencias absolutas del ejemplo 5 (paso 3). Derecha: tabla final.

<sup>1</sup>La instrucción =CONTAR.SI(A1:A38; “=”&B2) examina las columnas A1 a A38 y cuenta cuántas de ellas tienen un valor igual al de la casilla B2

4. Las frecuencias acumuladas se calculan siguiendo el procedimiento explicado en los ejemplos anteriores. La tabla final se muestra en la figura 14.

## 1.1. Utilización de la base de datos del INE

El Instituto Nacional de Estadística (INE) ofrece a través de su página web ([www.ine.es](http://www.ine.es)) gran cantidad de información sobre distintos temas: Educación, Cultura, Salud, Economía, Justicia, etc. Los datos de los ejemplos de la sección anterior se han obtenido de esta web. También la Dirección General de Tráfico en su página web ([www.dgt.es](http://www.dgt.es)) ofrece datos estadísticos sobre seguridad vial. En Baleares, una fuente importante de datos oficiales es el Institut d'Estadística de les Illes Balears (IBESTAT, <http://www.caib.es/ibae/ibae.htm>).

En esta sección explicamos cómo utilizar la base de datos del INE. Por ejemplo, supongamos que deseamos hacer un estudio sobre Educación. Los pasos a seguir para obtener los datos del INE son los siguientes:

1. abrir en el navegador la dirección <http://www.ine.es>
2. hacer clic sobre la opción  que aparece a la izquierda de la página principal
3. las diferentes opciones de la base de datos se muestran en la nueva página (ver Figura 15)
4. para acceder a los datos sobre Educación hacemos clic sobre *Educación* (bajo el epígrafe *Sociedad*) en el menú principal
5. en la nueva página se ofrecen varios estudios estadísticos relacionados con la educación. Supongamos que deseamos conocer los datos sobre *Gasto público en educación*, haremos clic sobre este concepto.
6. en la nueva página que se abre se explica en qué consisten los datos recopilados y se permite al usuario acceder a la información de un año en concreto. En el menú desplegable que aparece al hacer clic sobre *Seleccione un año* escogemos por ejemplo la opción *2004-2005*
7. se nos ofrecen varios tipos de datos (ver Figura 16). Elegimos por ejemplo la opción *Becas e importe de las mismas por universidad en la que está matriculado el becario, número, entidad que las concede y tipo de beca*, bajo el epígrafe *Enseñanzas universitarias*
8. la nueva página que se abre nos permite escoger los datos a mostrar y la manera de mostrarlos mediante una serie de menus (ver Figura 17). Podemos seleccionar por ejemplo las universidades Autónoma de Barcelona, Complutense de Madrid, Illes Balears, Pública de Navarra, Sevilla y Deusto.

Para ello haremos clic sobre las opciones del menú *Universidad en la que está matriculado el becario*, manteniendo pulsada la tecla *Ctrl*, lo que nos permitirá la selección simultánea de varias opciones.

<b>Entorno físico y medio ambiente</b>	<b>Economía</b>
<a href="#">Entorno físico</a>	<a href="#">Empresas</a>
<a href="#">Estadísticas sobre el medio ambiente</a>	<a href="#">Cuentas económicas</a>
<a href="#">Cuentas ambientales</a>	<a href="#">Estadísticas financieras y monetarias</a>
<a href="#">Indicadores ambientales</a>	<a href="#">Comercio exterior</a>
<a href="#">Otros estudios ambientales</a>	<a href="#">Información tributaria</a>
<b>Demografía y población</b>	<b>Ciencia y tecnología</b>
<a href="#">Cifras de población</a>	<a href="#">Investigación y desarrollo tecnológico</a>
<a href="#">- Padrón municipal</a>	<a href="#">Nuevas tecnologías de la información y la comunicación</a>
<a href="#">- Estimaciones y proyecciones</a>	
<a href="#">- Censos de Población</a>	
<a href="#">- Datos históricos</a>	
<a href="#">Movimiento natural de la población</a>	
<a href="#">Migraciones</a>	
<a href="#">Análisis y estudios demográficos</a>	
<b>Sociedad</b>	<b>Agricultura</b>
<a href="#">Educación</a>	<a href="#">Agricultura, ganadería, silvicultura y pesca</a>
<a href="#">Cultura y ocio</a>	
<a href="#">Salud</a>	
<a href="#">Justicia</a>	
<a href="#">Nivel, calidad y condiciones de vida.(IPC, ...)</a>	
<a href="#">Mercado laboral</a>	
<a href="#">Análisis sociales</a>	
<a href="#">Elecciones</a>	
	<b>Industria y construcción</b>
	<a href="#">Industria</a>
	<a href="#">Energía</a>
	<a href="#">Construcción y vivienda</a>
	<b>Servicios</b>
	<a href="#">Encuestas globales del sector servicios</a>
	<a href="#">Comercio</a>
	<a href="#">Transporte y actividades conexas, comunicaciones</a>
	<a href="#">Hostelería y turismo</a>
	<a href="#">Otros servicios empresariales, personales y comunitarios</a>
	<b>Clasificaciones</b>
	<a href="#">Clasificaciones nacionales</a>
	<a href="#">Clasificaciones internacionales</a>
	<a href="#">Proceso de revisión de clasificaciones</a>
	<b>Internacional</b>
	<a href="#">Internacional</a>
	<b>Historia</b>
	<a href="#">Fondo documental</a>

Figura 15: Opciones de la base de datos del INE

<b>Becas y ayudas. Curso 2004-2005</b>
<b>Todas las enseñanzas</b>
<a href="#">■ 1.1 Becas, ayudas, becarios, beneficiarios e importe de las mismas por CCAA de destino, entidad que las concede, número y tipo de enseñanza.</a>
<a href="#">■ 1.2 Becas, ayudas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de beca o ayuda (1).</a>
<b>Enseñanzas obligatorias, educación infantil y educación especial</b>
<a href="#">■ 2.1 Ayudas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de ayuda.</a>
<a href="#">■ 2.2 Ayudas e importe de las mismas por CCAA de destino, número, entidad que las concede y tipo de ayuda.</a>
<a href="#">■ 2.3 Ayudas, beneficiarios e importe de las mismas concedidas por CCAA de destino, número, entidad que las concede y nivel educativo del beneficiario.</a>
<b>Enseñanzas postobligatorias no universitarias</b>
<a href="#">■ 3.1 Becas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de beca.</a>
<a href="#">■ 3.2 Becas e importe de las mismas por CCAA de destino, número, entidad que las concede y tipo de beca.</a>
<a href="#">■ 3.3 Becas, becarios e importe de las mismas por CCAA de destino, número, entidad que las concede y nivel educativo del becario.</a>
<b>Enseñanzas universitarias</b>
<a href="#">■ 4.1 Becas e importe de las mismas por administración educativa financiadora, número y tipo de beca.</a>
<a href="#">■ 4.2 Becas e importe de las mismas por universidad en la que está matriculado el becario, número, entidad que las concede y tipo de beca.</a>
<a href="#">■ 4.3 Becas, becarios e importe de las mismas (1) por universidad en la que está matriculado el becario, entidad que las concede y número.</a>

Figura 16: Datos sobre gasto público en educación en la base de datos del INE

A continuación seleccionaremos las opciones Becas e Importe en el menú *Número* (manteniendo pulsada la tecla *Ctrl*), la opción Todas las Administraciones en el menú *Entidad que las concede* y Total en *Tipo de beca*.

Finalmente elegiremos la forma de visualizar los datos. Podemos elegir qué variables se mostrarán por filas y cuales por columnas. Por defecto el menú nos ofrece visualizar por filas la variable *Universidad en la que está matriculado el becario* y por columnas el número de becas, la entidad que las concede y el tipo. Aceptamos las opciones por defecto y generamos el resultado pulsando el botón **Consultar selección**. Obtendremos el resultado que se muestra en la figura 18

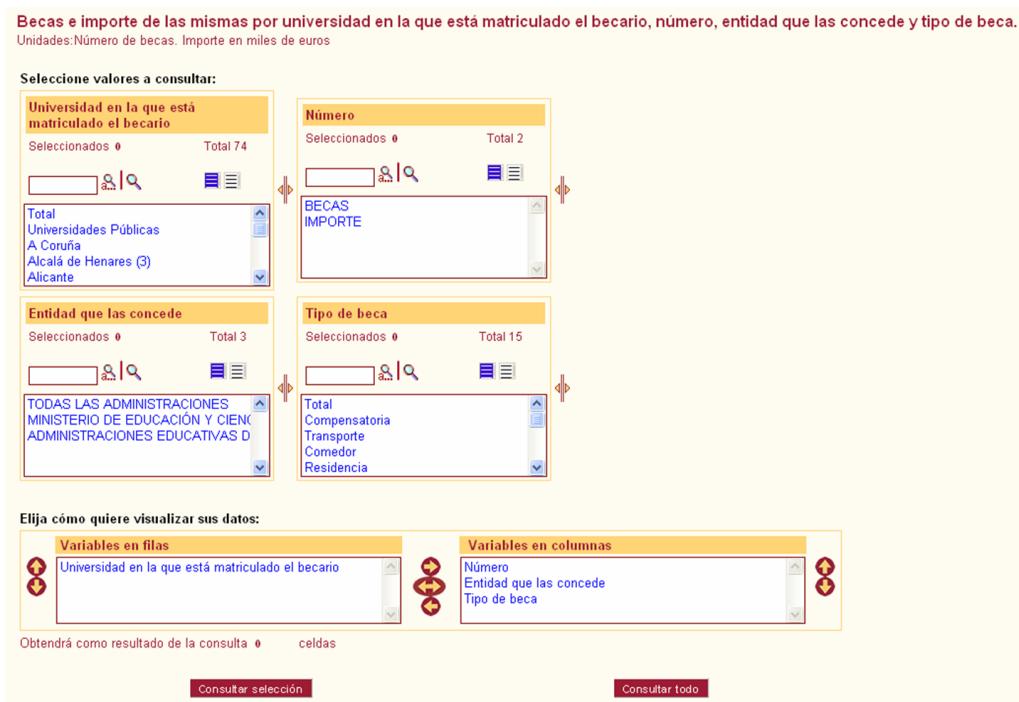


Figura 17: Menus para la selección de datos sobre número de becas e importe de las mismas en centros universitarios

- la tabla obtenida se puede imprimir pulsando el icono de la página de resultados. También se puede guardar en distintos formatos para ser utilizada por distintos programas estadísticos pulsando sobre el botón **Descargar como:** .

Nosotros elegiremos descargar como fichero Excel y guardaremos el fichero obtenido (extensión .xls) en alguna carpeta de nuestro ordenador. Este tipo de ficheros pueden leerse desde la aplicación Excel de Microsoft y también desde la aplicación Calc de OpenOffice, que es la que utilizamos en nuestros ejemplos.

Becas y ayudas. Curso 2004-2005		
Enseñanzas universitarias		
Becas e importe de las mismas por universidad en la que está matriculado el becario, número, entidad que las concede y tipo de beca.		
Unidades Número de becas. Importe en miles de euros		
	BECAS	IMPORTE
	TODAS LAS ADMINISTRACIONES	TODAS LAS ADMINISTRACIONES
Total	Total	Total
Autónoma de Barcelona	13.528	10.514,4
Complutense de Madrid (3)	24.416	26.248,1
Illes Balears (2)	8.237	6.146,6
Pública de Navarra	3.267	3.200,5
Sevilla	33.592	32.506,0
Deusto	2.622	1.609,8

Notas:

1) No se contabilizan como becas las de exención de precios académicos a familias numerosas de 3 hijos que afectan a 102.670 alumnos y ascienden a 33.912,4 miles de euros.

2) Se ha supuesto que los becarios y por tanto las becas son para los alumnos de la Universidad de Baleares.

3) En la Comunidad de Madrid, en el caso de las becas de exención de precios concedidas a alumnos con discapacidad superior al 50%, figura su importe, pero se desconoce el número de ayudas.

Fuente: Ministerio de Educación y Ciencia

Figura 18: Tabla de datos sobre número de becas e importe de las mismas en centros universitarios

Al abrir el fichero obtenido con Calc los nombres de las universidades seleccionadas aparecen en la columna A de la hoja de cálculo mientras que el número de becas obtenidas por cada universidad así como su importe aparecen en las columnas B y C. Siguiendo los pasos explicados en la sección anterior podremos calcular las tablas y gráficos asociados a estos datos.

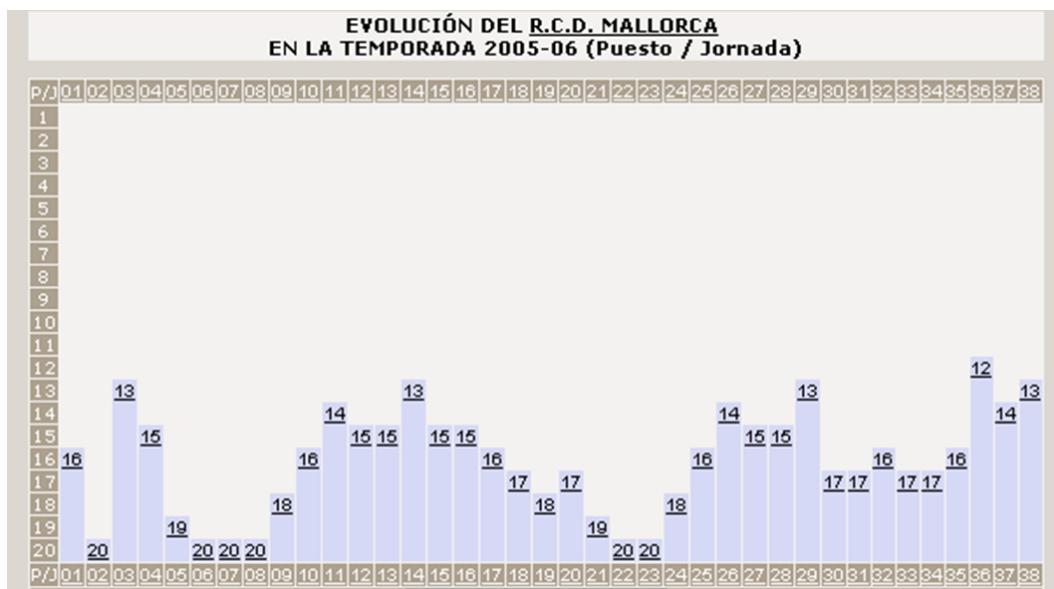
De manera similar se puede acceder a datos sobre la evolución anual de diferentes parámetros (población, ocupación turística, producción industrial, etc) haciendo clic sobre la opción   Banco de series temporales que aparece a la izquierda de la página principal del INE.

## 2. Cálculo de medidas de tendencia central

Las operaciones matemáticas que deben realizarse para calcular los estadísticos asociados a una distribución de datos son muy sencillas y pueden realizarse con una simple calculadora. No obstante, si la cantidad de datos es elevada, programas como el OpenOffice Calc permiten calcular fácilmente los estadísticos más habituales. Explicaremos cómo hacerlo en los siguientes ejemplos.

### Ejemplo 1

Calcular la media, la moda, la mediana y los cuartiles 1º y 3º de la clasificación del RCD Mallorca durante la temporada 2005-2006 a partir de los datos del siguiente gráfico (fuente LFP):



OpenOffice Calc permite calcular de forma muy sencilla la moda, mediana y media de un conjunto de datos *brutos*, es decir, no organizados en un tabla de frecuencias. Para resolver este ejemplo seguiremos los pasos siguientes:

1. Al igual que en el ejemplo 5 del tema anterior, en primer lugar escribiremos los datos *brutos* en la primera columna de la hoja de cálculo (casillas A1 a A38). El resultado se muestra en la figura 13-izquierda del tema anterior.
2. A continuación escribimos las palabras “Moda”, “Mediana”, “Media”, “1er cuartil” y “3er cuartil” en las casillas C15, C16, C17, C18 y C19 de la hoja de cálculo (o en otras casillas cualesquiera)
3. **Moda.** Nos situamos en la casilla D15 y escribimos =Moda(A1:A38). Al pulsar *Enter* obtenemos el valor de la moda.

4. **Mediana.** Nos situamos en la casilla *D16* y escribimos =Mediana(A1:A38). Al pulsar *Enter* obtenemos el valor de la mediana.
5. **Media.** Nos situamos en la casilla *D17* y escribimos =Promedio(A1:A38). Al pulsar *Enter* obtenemos el valor de la media.
6. **1<sup>er</sup> cuartil.** Nos situamos en la casilla *D18* y escribimos =Cuartil(A1:A38;1). Al pulsar *Enter* obtenemos el valor del 1<sup>er</sup> cuartil.
7. **3<sup>er</sup> cuartil.** Nos situamos en la casilla *D19* y escribimos =Cuartil(A1:A38;3). Al pulsar *Enter* obtenemos el valor del 3<sup>er</sup> cuartil.

El resultado obtenido es:

Moda	15
Mediana	16
Media	16,34
1 <sup>er</sup> cuartil	15
3 <sup>er</sup> cuartil	18

Los tres valores centrales (moda, media y mediana) son muy similares, aunque se cumple que media > mediana > moda, por lo que podemos afirmar que se trata de una distribución casi simétrica con una ligera asimetría hacia la derecha.

**Nota:** al utilizar las fórmulas *Mediana* y *Cuartil* de Calc para el cálculo de la mediana y el primer y tercer cuartiles los resultados obtenidos pueden ser ligeramente diferentes a los que obtendríamos con el método que se explica en los ejemplos siguientes. La razón es que Calc utiliza unas fórmulas diferentes a las nuestras para el cálculo de cuartiles de datos *brutos*.

## Ejemplo 2

Repetir los cálculos del ejemplo anterior pero a partir de los datos organizados en la tabla de frecuencias siguiente (correspondiente al ejemplo 5 del tema 2):

Clasificación	Frecuencia absoluta
12	1
13	4
14	3
15	7
16	6
17	6
18	3
19	2
20	6

En este caso el cálculo de la media es muy sencillo pero el cálculo de moda y mediana es algo más complicado. A continuación explicamos cómo hacerlo:

1. En primer lugar escribimos los valores de la variable y sus frecuencias en una tabla de OpenOffice Calc. A continuación calculamos las frecuencias y los porcentajes acumulados, tal como se ha explicado en el tema anterior. La tabla que obtenemos es como la de la figura 19.

A Clasificación	B Frecuencia absoluta	C Frecuencia acumulada	D Porcentaje acumulado
1	12	1	2,63
2	13	4	13,16
3	14	3	21,05
4	15	7	39,47
5	16	6	55,26
6	17	6	71,05
7	18	3	78,95
8	19	2	84,21
9	20	6	100
10			
11			
12			

Figura 19: OpenOffice Calc con los del ejemplo 2 y las frecuencias y porcentajes acumulados calculadas

2. **Moda.** En casos como el del ejemplo, en los que la tabla de datos es pequeña basta observar los valores de frecuencia absoluta para descubrir para qué valor de la variable tenemos un máximo. En este ejemplo la moda es 15, ya que tiene el valor de frecuencia absoluta mayor (7).

Cuando la tabla de datos es muy grande, la moda se puede encontrar siguiendo el siguiente procedimiento:

- a) Primero copiamos los datos originales en unas nuevas columnas para evitar perderlos:
  - 1) seleccionamos, manteniendo el botón del ratón pulsado, las casillas con los valores de la variable y sus frecuencias absolutas (en nuestro ejemplo las casillas *A2* a *A10* y *B2* a *B10*),
  - 2) pulsamos la combinación de teclas *Ctrl-C* para copiar estos datos,
  - 3) situamos el cursor en alguna otra casilla del documento (por ejemplo, *A14*),
  - 4) pulsamos el botón derecho del ratón y seleccionamos la opción *Pegado especial...*,
  - 5) activamos la opción *Números* y desactivamos todas las demás,
  - 6) finalmente pulsamos *Aceptar*. Los datos (sólo los valores numéricos, no las fórmulas) quedan copiados en las casillas *A14* a *A22* y *B14* a *B22*.
- b) Seleccionamos las casillas *A14* a *A22* y *B14* a *B22* manteniendo el botón izquierdo del ratón pulsado.

- c) En el menú principal escogemos la opción *Datos*, y a continuación *Ordenar...*
- d) Se abre una ventana en la que se definen los criterios de ordenación. En nuestro caso escogemos las opciones: *Ordenar según: Columna B* y *Ascendente*. Pulsamos *Aceptar*.
- e) En la columna A (casillas A14 a A22) aparecen los valores de la variable ordenados en orden decreciente de frecuencia absoluta (casillas B14 a B22), ver figura 20. El primer valor de la columna es la moda de la distribución. En nuestro caso: moda=15. Si la distribución fuera bimodal (el máximo ocurre en dos valores de la variable), deberíamos tomar como moda los dos primeros valores de la columna A.

	A	B	
13			
14	15	7	
15	20	6	
16	16	6	
17	17	6	
18	13	4	
19	14	3	
20	18	3	
21	19	2	
22	12	1	
23			

Figura 20: Ilustración del cálculo de la moda en el ejemplo 2

3. **Mediana.** La mediana es el primer valor de la columna *Clasificación* cuyo porcentaje acumulado es superior o igual al 50 %. Observando la tabla de frecuencias de la figura 19 vemos que la mediana es 16, ya que su porcentaje acumulado es 55, 26 %.
4. **1<sup>er</sup> cuartil.** El cálculo es similar al de la mediana. El 1<sup>er</sup> cuartil es el primer valor de la columna *Clasificación* cuyo porcentaje acumulado es superior o igual al 25 %. Observando la tabla de frecuencias de la figura 19 vemos que el 1<sup>er</sup> cuartil es 15, ya que su porcentaje acumulado es 39, 47 %.
5. **3<sup>er</sup> cuartil.** Igual que en el caso anterior pero buscando un porcentaje acumulado mayor o igual que 75 %. En este caso el 3<sup>er</sup> cuartil es 18, ya que su porcentaje acumulado es 78, 95 %.
6. **Media.** La media en este caso se calcula de manera muy sencilla del siguiente modo:
  - a) Situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla F2.
  - b) Escribimos la fórmula

=SUMA . PRODUCTO(A2:A10;B2:B10)/SUMA(B2:B10)<sup>2</sup> y pulsamos *Enter*. El valor de la media se escribe en la casilla *F2*. En este caso, media=16,34.

Por último comentar que los valores de moda, mediana y media obtenidos son los mismos que en el ejemplo anterior.

### Ejemplo 3

Calcular la mediana y la moda de la calificación de los alumnos de *Fonaments Matemàtics II* (Ingeniería Telemática) del curso 2003-2004, a partir de los datos de la siguiente tabla (fuente UIB). ¿Es posible calcular la media?

Assignatura:	2485 - Fonaments Matemàtics II
Any acadèmic:	2003-04
Convocatòria:	Juny
Qualificació	Núm. alumnes
Suspens	17
Aprovat	16
Notable	12
Excel·lent	1
Matricula d'honor	3

En este ejemplo, la variable “Qualificació” es una variable ordinal por lo que no es posible calcular su media, pero sí su moda y mediana.

Al tratarse de una tabla con muy pocos valores la moda se encuentra fácilmente observando la tabla: moda=*Suspens*, ya que es el valor con la máxima frecuencia absoluta.

Para calcular la mediana debemos calcular primero los porcentajes acumulados. Lo podemos hacer en una hoja de cálculo de OpenOffice Calc tal como se ha explicado en el tema anterior y obtendríamos el resultado de la figura 21.

A continuación seguimos el procedimiento explicado en el ejemplo 2 para el cálculo de la mediana: buscamos el primer valor de la columna *Qualificació* cuyo porcentaje acumulado sea superior o igual a 50 %. En este caso la solución es *Aprovat* (porcentaje acumulado=67,35).

Este ejemplo muestra como el valor de la moda no explica suficientemente bien la distribución de valores. En este caso la moda era *Suspens*, sin embargo más de la mitad de los estudiantes han aprobado (de hecho  $16 + 12 + 1 + 3 = 32$  estudiantes

---

<sup>2</sup>Esta fórmula calcula la media empleando la fórmula vista en clase: multiplica los valores de las casillas *A2* y *B2*, *A3* y *B3*, etc, suma los productos y finalmente divide el total por la suma de los valores de las casillas *B2* a *B10*

	A	B	C	D	
1	Qualificació	Frec. Absoluta	Frec. acumulada	Porcentaje acumulado	
2	Suspens	17	17	34,69	
3	Aprovat	16	33	67,35	
4	Notable	12	45	91,84	
5	Excellent	1	46	93,88	
6	Matricula H.	3	49	100	
7					

Figura 21: Izquierda: frecuencias absolutas y acumuladas para el ejemplo 3.

aprueban, exactamente el doble de alumnos suspendidos). La mediana describe de manera mejor los valores de la variable al dar un valor de *Aprovat*.

#### Ejemplo 4

Calcular la moda y la mediana de la edad de los condenados en Illes Balears en 2005 a partir de los datos de la siguiente tabla. Calcular los cuartiles primero y tercero. ¿Es posible calcular la media? Si la respuesta es negativa, ¿cómo estimarías de manera aproximada el valor de la media?

Estadísticas judiciales 2005	
Estadística de lo Penal. Condenados. Resultados autonómicos	
Condenados según edad y sexo	
Unidades: nº de condenados	
Ambos sexos	
Balears (Illes)	
De 18 a 20 años	155
De 21 a 25 años	543
De 26 a 30 años	653
De 31 a 35 años	619
De 36 a 40 años	515
De 41 a 50 años	636
De 51 a 60 años	248
De 60 y más	100

Fuente: Instituto Nacional de Estadística

Se trata de datos agrupados en forma de intervalos, de manera que calcularemos los estadísticos siguiendo el procedimiento explicado en clase:

1. **Moda.** Observando la tabla vemos que el valor máximo se da en el intervalo 26 – 30. La moda se calcula como el valor medio del intervalo, es decir:  $\text{moda} = \frac{26+30}{2} = 28$ .
2. **Mediana.** Calculamos las frecuencias acumuladas para cada intervalo y seguimos el procedimiento descrito en el ejemplo 2 para hallar la mediana. En este caso, la mediana se encuentra en el intervalo 31 – 35 (ver figura 22).

	A	B	C	D	
1	Edad	Frec. Absoluta	Frec. Acumulada	Porcentaje acumulado	
2	18-20	155	155	4,47	
3	21-25	543	698	20,12	
4	26-30	653	1351	38,94	
5	31-35	619	1970	56,79	
6	36-40	515	2485	71,63	
7	41-50	636	3121	89,97	
8	51-60	248	3369	97,12	
9	60-70	100	3469	100	
10	...				

Figura 22: Tabla de frecuencias absolutas y acumuladas y porcentajes acumulados para el ejemplo 3.

Para calcular de modo más preciso la mediana utilizamos la fórmula dada en clase:

$$\text{mediana} = L_i + \frac{50\% \cdot N - N_{i-1}}{n_i} \cdot (L_{i+1} - L_i)$$

donde  $L_i$  y  $L_{i+1}$  denotan los límites inferior y superior del intervalo,  $n_i$  es la frecuencia del intervalo,  $N_{i-1}$  es la frecuencia acumulada en el intervalo anterior y  $N$  es la suma de todas las frecuencias

En nuestro caso:  $L_i = 31$ ,  $L_{i+1} = 35$ ,  $n_i = 619$ ,  $N_{i-1} = 1351$  y  $N = 3469$ , por tanto

$$\begin{aligned} \text{mediana} &= 31 + \frac{50\% \cdot 3469 - 1351}{619} \cdot (35 - 31) = \\ &= 31 + \frac{1734,5 - 1351}{619} \cdot 4 = 33,48 \end{aligned}$$

El cálculo de los cuartiles es muy similar al de la mediana. Primero hallamos los intervalos en los que se encuentra cada uno de ellos, siguiendo un procedimiento similar al explicado en el ejemplo 2. A continuación aplicamos la fórmula de los cuartiles:

- **Primer cuartil.**  $L_i = 26$ ,  $L_{i+1} = 30$ ,  $n_i = 653$ ,  $N_{i-1} = 698$  y  $N = 3469$

$$26 + \frac{25\% \cdot 3469 - 698}{653} \cdot (30 - 26) = 26 + \frac{867,25 - 698}{653} \cdot 4 = 27,04$$

- **Tercer cuartil.**  $L_i = 41$ ,  $L_{i+1} = 50$ ,  $n_i = 636$ ,  $N_{i-1} = 2485$  y  $N = 3469$

$$41 + \frac{75\% \cdot 3469 - 2485}{636} \cdot (50 - 41) = 41 + \frac{2601,75 - 2485}{636} \cdot 9 = 42,65$$

3. **Media.** Para calcular la media de unos valores agrupados en intervalos el primer paso consiste en calcular el valor medio de cada intervalo. En este

ejemplo sin embargo tenemos el problema de que para el último intervalo no podemos calcular el valor medio, ya que está definido como *60 y más* y no conocemos el límite superior:

En estos casos podemos calcular la media de manera aproximada haciendo alguna suposición razonable sobre el valor máximo del intervalo desconocido. A continuación explicamos el procedimiento a seguir si suponemos que el valor máximo del intervalo es 70:

- Partimos de un documento OpenOffice Calc en el que hemos creado una tabla de frecuencias absolutas y acumuladas como la que se muestra en la figura 22.
- Insertamos una nueva columna a la derecha de la columna *A*. Para ello situamos el cursor sobre la parte superior de la columna *B*, hacemos clic en el botón derecho del ratón y elegimos la opción *insertar columnas*. Una nueva columna *B* aparece y las columnas *B*, *C*, *D*, etc se desplazan hacia a la derecha (ver figura 23-arriba).
- En la primera casilla de la nueva columna escribimos *Edad media* y en las casillas inferiores escribimos las fórmulas que calculan los valores medios de los intervalos:  $=(18+20)/2$ , *Enter* (casilla *B2*);  $=(21+25)/2$ , *Enter* (casilla *B3*); etc. Finalmente, para la casilla *B9* *suponemos* que el valor máximo del intervalo es 70 y escribimos  $=(60+70)/2$ . La tabla resultante tiene la forma que se muestra en la figura 23-abajo.

	A	B	C	D	E
1	Edad		Frec. Absoluta	Frec. Acumulada	Porcentaje acumulado
2	18-20	19	155	155	4,47
3	21-25	23	543	698	20,12
4	26-30	28	653	1351	38,94
5	31-35	33	619	1970	56,79
6	36-40	38	515	2485	71,63
7	41-50	45,5	636	3121	89,97
8	51-60	55,5	248	3369	97,12
9	60-70	65	100	3469	100
10	..				

	A	B	C	D	E
1	Edad	Edad media	Frec. Absoluta	Frec. Acumulada	Porcentaje acumulado
2	18-20	19	155	155	4,47
3	21-25	23	543	698	20,12
4	26-30	28	653	1351	38,94
5	31-35	33	619	1970	56,79
6	36-40	38	515	2485	71,63
7	41-50	45,5	636	3121	89,97
8	51-60	55,5	248	3369	97,12
9	60-70	65	100	3469	100
10	..				

Figura 23: Tablas de frecuencias y porcentajes para el ejemplo 3. Arriba, inserción de una nueva columna. Abajo, datos insertados con los valores centrales de los intervalos

- El cálculo de la media se hace ahora de manera similar al ejemplo 2:
  - Situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla *A12*.

- 2) Escribimos la fórmula  
 $=\text{SUMA.PRODUCTO}(\text{B2:B9};\text{C2:C9})/\text{SUMA}(\text{C2:C9})$  y  
 pulsamos *Enter*. El valor de la media se escribe en la casilla *A12*. En este caso, media=35,43.

Como comentario final decir que otras suposiciones razonables sobre el valor máximo del intervalo 60 *y más* hubieran producido resultados similares. Por ejemplo, para la suposición 60 – 65 hubiéramos obtenido media=35,36; para 60 – 75, media=35,51; para 60 – 80, media=35,58, etc. Lo importante es no suponer valores absurdos (por ejemplo 60 – 150, pues es muy poco probable que haya personas de 150 años que cometan delitos).

### Ejemplo 5

Calcular los estadísticos de tendencia central asociados a la variable “tipo de infracción” a partir de los datos de la siguiente tabla:

Estadísticas judiciales 2005	
Estadística de lo Penal. Menores. Resultados autonómicos y provinciales	
Menores según infracción cometida	
Unidades:nº de menores	
	BALEARS (ILLES)
Homicidio	0
Lesiones	31
Contra la libertad sexual	12
Hurto	69
Robo	306
Contra la salud pública	10

Fuente:Instituto Nacional de Estadística

Se trata de una variable cualitativa por lo que el único estadístico que podemos calcular es la moda. Observando la tabla vemos que el valor máximo de frecuencia es 306, que corresponde al valor *Robo*. Por lo que concluimos que: moda=*Robo*.

### Ejemplo 6

Calcular los estadísticos de tendencia central asociados a la variable “autobuses matriculados por mes durante el año 2006” a partir de los datos de la siguiente tabla (fuente DGT):

MATRICULACIONES POR MES Y TIPO DE VEHÍCULO								
Meses	Total	Camiones MMA > 3.500 kg	Camiones MMA ≤ 3.500 kg y furgonetas	Autobuses	Turismos	Motocicletas	Tractores Industriales	Otros vehículos
Enero	158.712	1.579	25.601	170	115.490	14.402	1.013	457
Febrero	178.150	1.852	29.646	274	128.831	15.888	1.217	442
Marzo	243.927	2.274	39.265	359	176.075	23.389	1.791	774
Abril	187.706	2.003	29.014	427	131.631	21.985	1.936	710
Mayo	224.821	2.106	35.772	376	155.805	28.302	1.783	677
Junio	246.787	2.301	36.496	364	171.028	33.877	1.906	815
Julio	240.910	2.228	34.286	278	169.034	32.703	1.722	659
Agosto	158.200	1.746	25.308	158	105.190	24.002	1.377	419
Septiembre	158.949	1.478	24.497	634	107.510	21.908	2.578	344
Octubre	186.334	1.858	30.055	282	128.178	23.135	2.397	429
Noviembre	193.677	1.949	33.655	235	135.134	20.231	2.000	473
Diciembre	186.483	1.486	31.106	290	136.721	15.096	1.368	416

Los datos de la tabla son valores *en bruto*, por lo que aplicamos el procedimiento explicado en el ejemplo 1:

1. Escribimos los datos *brutos* en la primera columna de la hoja de cálculo (casillas A1 a A12). El resultado se muestra en la figura 24.

	A	I
1	170	
2	274	
3	359	
4	427	
5	376	
6	364	
7	278	
8	158	
9	634	
10	282	
11	235	
12	290	
13		

Figura 24: Datos brutos del ejemplo 6

2. A continuación escribimos las palabras “Moda”, “Media”, “Mediana”, “1er cuartil” y “3er cuartil” en las casillas C15, C16, C17, C18 y C19 de la hoja de cálculo (o en otras casillas cualesquiera)
3. **Moda.** Nos situamos en la casilla D15 y escribimos =Moda(A1:A12). Al pulsar *Enter* obtenemos el valor de la moda. En este caso obtenemos un mensaje de error pues todos los valores ocurren una única vez. Esto significa que el cálculo de la moda no tiene sentido en este problema.
4. **Media.** Nos situamos en la casilla D16 y escribimos =Promedio(A1:A12). Al pulsar *Enter* obtenemos el valor de la media.

5. **Mediana.** Nos situamos en la casilla  $D17$  y escribimos  $=\text{Mediana}(A1:A12)$ . Al pulsar *Enter* obtenemos el valor de la mediana.
6.  **$1^{er}$  cuartil.** Nos situamos en la casilla  $D18$  y escribimos  $=\text{Cuartil}(A1:A12;1)$ . Al pulsar *Enter* obtenemos el valor del primer cuartil.
7.  **$3^{er}$  cuartil.** Nos situamos en la casilla  $D19$  y escribimos  $=\text{Cuartil}(A1:A12;3)$ . Al pulsar *Enter* obtenemos el valor del tercer cuartil.

El resultado obtenido es:

Media	320,58
Mediana	286
$1^{er}$ cuartil	264,25
$3^{er}$ cuartil	367

**Nota:** si para el cálculo de la mediana y cuartiles hubiéramos calculado primero la tabla de frecuencias absolutas de cada valor y a continuación hubéramos hecho el cálculo con el procedimiento explicado en los ejemplos anteriores, los resultados hubieran sido ligeramente diferentes: mediana=282,  $1^{er}$  cuartil=235 y  $3^{er}$  cuartil=364. La razón es que Calc utiliza unas fórmulas diferentes a las nuestras para el cálculo de cuartiles para datos *brutos*.

### 3. Cálculo de medidas de dispersión

#### Ejemplo 1

Calcular el recorrido, recorrido intercuartílico, varianza, desviación estándar y coeficiente de variación para la variable “autobuses matriculados por mes durante el año 2006” a partir de los datos de la siguiente tabla (fuente DGT):

Meses	Total	MATRICULACIONES POR MES Y TIPO DE VEHÍCULO						
		Camiones MMA>3.500 kg	Camiones MMA ≤3.500 kg y furgonetas	Autobuses	Turismos	Motocicletas	Tractores Industriales	Otros vehículos
Enero	158.712	1.579	25.601	170	115.490	14.402	1.013	457
Febrero	178.150	1.852	29.646	274	128.831	15.888	1.217	442
Marzo	243.927	2.274	39.265	359	176.075	23.389	1.791	774
Abril	187.706	2.003	29.014	427	131.631	21.985	1.936	710
Mayo	224.821	2.106	35.772	376	155.805	28.302	1.783	677
Junio	246.787	2.301	36.496	364	171.028	33.877	1.906	815
Julio	240.910	2.228	34.286	278	169.034	32.703	1.722	659
Agosto	158.200	1.746	25.308	158	105.190	24.002	1.377	419
Septiembre	158.949	1.478	24.497	634	107.510	21.908	2.578	344
Octubre	186.334	1.858	30.055	282	128.178	23.135	2.397	429
Noviembre	193.677	1.949	33.655	235	135.134	20.231	2.000	473
Diciembre	186.483	1.486	31.106	290	136.721	15.096	1.368	416

Los datos de esta tabla ya se han utilizado en el tema anterior. Se trata de datos en bruto que deben escribirse en un documento de OpenOffice Calc, tal como muestra la figura 25

A	I
1	170
2	274
3	359
4	427
5	376
6	364
7	278
8	158
9	634
10	282
11	235
12	290
13	

Figura 25: Datos *brutos* del ejemplo 2

El procedimiento para el cálculo de la mediana y los cuartiles se ha explicado en el ejemplo 6 del tema anterior. Recordemos los resultados: mediana=286, 1<sup>er</sup> cuartil=264, 25 y 3<sup>er</sup> cuartil =367. El recorrido intercuartílico es por tanto:  $RIC = 367 - 264, 25 = 102, 75$ .

El cálculo de la varianza, la desviación típica, el coeficiente de variación y el recorrido es muy sencillo con Calc cuando se dispone de valores brutos. Para este ejemplo:

1. Escribimos las palabras “Varianza”, “Desv. típica”, “Coef. Variación”, “Mínimo”, “Máximo” y “Recorrido” en las casillas  $C20$  a  $C25$ , por ejemplo, de la hoja de cálculo.
2. **Varianza.** Debemos decidir primero si consideramos que los datos se refieren a una *población* o a una *muestra*. Dado que la variable bajo estudio es “autobuses matriculados por mes durante el año 2006” y disponemos de *todos* los datos de este año, podemos considerar que se trata de datos de población.  
Para el cálculo nos situamos en la casilla  $D20$  y escribimos  $=Varp(A1:A12)$ . Al pulsar *Enter* obtenemos el valor de la varianza poblacional. (Si hubiéramos querido calcular la varianza muestral la fórmula hubiera sido  $=Vara(A1:A12)$ ).
3. **Desviación estándar.** Nos situamos en la casilla  $D21$  y escribimos  $=Raíz(D20)$ . Al pulsar *Enter* obtenemos el valor de la desviación estándar.
4. **Coeficiente de variación.** Nos situamos en la casilla  $D22$  y escribimos  $=D21/(SUMA(A1:A12)/12)$ . Al pulsar *Enter* obtenemos el valor del coeficiente de variación.
5. **Mínimo.** Nos situamos en la casilla  $D23$  y escribimos  $=Mín(A1:A12)$ . Al pulsar *Enter* obtenemos el valor de mínimo de la variable.
6. **Máximo.** Nos situamos en la casilla  $D24$  y escribimos  $=Máx(A1:A12)$ . Al pulsar *Enter* obtenemos el valor de máximo de la variable.
7. **Recorrido.** Nos situamos en la casilla  $D25$  y escribimos  $=D23-D22$ . Al pulsar *Enter* obtenemos el recorrido de la variable.

La siguiente tabla resume los resultados obtenidos hasta el momento:

Mediana	286
1 <sup>er</sup> cuartil	264, 25
3 <sup>er</sup> cuartil	367
RIC	102, 75
Varianza	14902, 24
Desviación típica	122, 07
Coeficiente de variación	0, 38
Mínimo	158
Máximo	634
Rango	476

## Ejemplo 2

Se desea hacer un estudio sobre la obesidad en los institutos de secundaria de Baleares. Para ello se seleccionan al azar 300 alumnos de secundaria y se registra su peso. A partir de los datos de la siguiente tabla calcular la media, varianza y desviación típica de la variable *Peso*.

Peso (Kg)	Nº alumnos
60	6
63	10
65	20
67	25
68	15
70	35
72	44
75	50
77	37
79	22
80	15
83	10
89	7
90	4

1. En primer lugar creamos un documento OpenOffice Calc y escribimos estos datos en las casillas *A2* a *A15* (peso) y *B2* a *B15* (*nº* alumnos).
2. A continuación calculamos la media tal como se ha explicado en el tema anterior: nos situamos en una casilla cualquiera (por ejemplo la *A17*) y escribimos `=SUMA.PRODUCTO(A2:A15;B2:B15)/SUMA(B2:B15)`. Al pulsar *Enter* el resultado se escribe en la casilla *A17*. El valor es 73,18.
3. Antes de calcular la varianza debemos decidir si ésta es poblacional o muestral. Por el enunciado del problema se deduce que los datos se refieren a una muestra formada por 300 alumnos del total de estudiantes de secundaria de la Baleares. Calcularemos por tanto la varianza muestral.

El cálculo se hace en dos pasos:

- a) En la casilla *C2* escribimos `=(A2-$A$17)^2`. Extendemos el cálculo al resto de casillas de la columna C situando el cursor en la esquina inferior derecha de la casilla *C2* y, manteniendo el botón izquierdo del ratón pulsado, arrastrando el cursor hasta la casilla *C15*. De esta manera en la columna *C* tenemos todos los factores  $(x_i - \bar{x})^2$  de la fórmula de la varianza.

- b) A continuación situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla A18 y escribimos la fórmula  $=\text{SUMA}.\text{PRODUCTO}(\text{B2:B15};\text{C2:C15})/(\text{SUMA}(\text{B2:B15})-1)$ . Al pulsar *Enter* obtenemos el valor de la varianza muestral en la casilla A18 (ver figura 26). El resultado final es 37,41.

	A	B	C
1	Peso	Frecuencia	
2	60	6	173,62
3	63	10	103,56
4	65	20	66,86
5	67	25	38,15
6	68	15	26,8
7	70	35	10,09
8	72	44	1,38
9	75	50	3,32
10	77	37	14,62
11	79	22	33,91
12	80	15	46,56
13	83	10	96,5
14	89	7	250,38
15	90	4	283,02
16			
17	73,18		
18	37,41		
19			

Figura 26: Hoja de cálculo del ejemplo 3.

La varianza poblacional se habría calculado con la fórmula  $=\text{SUMA}.\text{PRODUCTO}(\text{B2:B15};\text{C2:C15})/\text{SUMA}(\text{B2:B15})$ .

4. La desviación típica se calcula como la raíz cuadrada de la varianza: nos colocamos por ejemplo en la casilla A19, escribimos la fórmula  $=\text{RAÍZ}(A18)$  y pulsamos *Enter*. El resultado es 6,12.

### Ejemplo 3

Calcular el recorrido intercuartílico para la variable “Edad de los condenados en Baleares en 2005” a partir de los datos de la siguiente tabla. Suponiendo que la edad máxima es de 70 años, calcular el recorrido, la varianza y la desviación estándar.

Estadísticas judiciales 2005	
Estadística de lo Penal. Condenados. Resultados autonómicos	
Condenados según edad y sexo	
Unidades: nº de condenados	
Ambos sexos	
Balears (Illes)	
De 18 a 20 años	155
De 21 a 25 años	543
De 26 a 30 años	653
De 31 a 35 años	619
De 36 a 40 años	515
De 41 a 50 años	636
De 51 a 60 años	248
De 60 y más	100

Fuente: Instituto Nacional de Estadística

Los valores de media, mediana y cuartiles primero y tercero para este problema ya se calcularon en el ejemplo 4 del tema anterior:

Media (suponiendo edad máxima=70)	35,43
Mediana	33,48
1 <sup>er</sup> cuartil	27,04
3 <sup>er</sup> cuartil	42,65

De aquí deducimos que el recorrido intercuartílico es  $RIC = 15,61$ . Por otra parte, el valor mínimo de la variable *Edad* es 18 y, según el enunciado, el máximo es 70. De manera que el recorrido es  $70 - 18 = 52$ .

Para calcular la varianza debemos decidir primero qué fórmula emplearemos (poblacional o muestral). En este caso, como disponemos de datos acerca de *todos* los condenados en Baleares en 2005 consideramos que los datos se refieren a toda una población. Procedemos del siguiente modo para hacer el cálculo:

1. Supongamos que el valor de la media (calculada en el ejemplo 4 del tema anterior) se ha escrito en la casilla A12.

2. Creamos una nueva columna con los valores medios de cada intervalo, tal como se ha explicado en el tema anterior.
3. Insertamos una nueva columna a la derecha de la columna *D*. Para ello situamos el cursor sobre la parte superior de la columna *E*, hacemos clic en el botón derecho del ratón y elegimos la opción *insertar columnas*. Una nueva columna *E* aparece desplazando las que tiene a su derecha.
4. En la casilla *E2* escribimos  $=(B2-\$A\$12)^2$ . Extendemos el cálculo al resto de casillas de la columna *E* situando el cursor en la esquina inferior derecha de la casilla *E2* y, manteniendo el botón izquierdo del ratón pulsado, arrastrando el cursor hasta la casilla *E9*. De esta manera en la columna *E* tenemos todos los factores  $(x_i - \bar{x})^2$  de la fórmula de la varianza.
5. Finalmente, situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla *A13* y escribimos la fórmula  
 $=SUMA.PRODUCTO(B2:B9;E2:E9)/SUMA(B2:B9).$

Al pulsar *Enter* obtenemos el valor de la varianza poblacional en la casilla *A13* (ver figura 27). El resultado final es 121,27.

	A	B	C	D	E
1	Edad	Edad media	Frec. Absoluta	Frec. Acumulada	
2	18-20		19	155	155
3	21-25		23	543	698
4	26-30		28	653	1351
5	31-35		33	619	1970
6	36-40		38	515	2485
7	41-50		45,5	636	3121
8	51-60		55,5	248	3369
9	60-70		65	100	3469
10					
11					
12		35,43			
13		121,27			
14		11,01			

Figura 27: Hoja de cálculo del ejemplo 4.

En caso de tener que calcular la varianza muestral hubiéramos utilizado la siguiente fórmula:

$$=SUMA.PRODUCTO(B2:B9;E2:E9)/(SUMA(B2:B9)-1).$$

Finalmente calculamos la desviación típica como la raíz cuadrada de la varianza: nos colocamos por ejemplo en la casilla *A14*, escribimos la fórmula  $=RAÍZ(A13)$  y pulsamos *Enter*. El resultado es 11,01.

## 4. Cálculo de medidas de simetría y apuntamiento

### Ejemplo 1

Calcular la media, la varianza, la desviación típica y las medidas de simetría y apuntamiento para la siguiente variable que representa el total de personal dedicado a investigación en las diferentes comunidades autónomas en el 2007, según el INE:

Estadística de I+D 2007	
Resultados por Comunidades Autónomas	
Total sectores. Gastos internos totales y personal en I+D por comunidades autónomas	
Unidades:especificadas en las variables	
	Investigadores en EJC: Total personal
Andalucía	13232,5
Aragón	4548,5
Asturias (Principado de)	2013,4
Baleares (Illes)	1094,7
Canarias	3256
Cantabria	1207,1
Castilla y León	6227,2
Castilla - La Mancha	1649
Cataluña	25063
Comunitat Valenciana	10702,1
Extremadura	1261,5
Galicia	5413,7
Madrid (Comunidad de)	29497,1
Murcia (Región de)	3978,6
Navarra (Comunidad Foral de)	2983
País Vasco	9816
Rioja (La)	627,1
Ceuta	21,9
Melilla	31,8

Notas:

1.- EJC: equivalencia a jornada completa

En primer lugar creamos un documento OpenOffice Calc y escribimos estos datos en las casillas A2 a A20 (comunidad autónoma) y B2 a B20 (número total de personal).

A continuación calculamos la media de la variable tal como se ha explicado en temas anteriores: nos situamos en una casilla cualquiera (por ejemplo la B25) y escribimos =Promedio(B2:B20). Al pulsar *Enter* el resultado se escribe en la casilla B25. El valor es 10584,63.

El cálculo de la varianza, la desviación típica, el índice de simetría y la curtosis es muy sencillo con Calc cuando se dispone de valores en “bruto”. Antes de calcular la varianza debemos decidir si ésta es poblacional o muestral. Por el enunciado del problema se deduce que los datos se refieren a todos los trabajadores de todas las comunidades autónomas. Calcularemos por tanto la varianza poblacional. Debemos seguir los siguientes pasos:

1. Escribimos las palabras “Varianza”, “Desv. típica”, “Coef. Simetría” y “Curtosis” en las casillas de la A26 a la A29, por ejemplo, de la hoja de cálculo.
2. **Varianza.** Para realizar el cálculo nos situamos en la casilla B26 y escribimos =VARP(B2:B20). Al pulsar *Enter* obtenemos el valor de la varianza poblacional. (Si hubiéramos querido calcular la varianza muestral la fórmula hubiera sido =VARA(B2:B20)).
3. **Desviación estándar.** Nos situamos en la casilla B27 y calculamos la raíz cuadrada de la varianza, escribiendo =RAÍZ(B26). Al pulsar *Enter* obtenemos el valor de la desviación estándar.
4. **Índice de simetría** En el caso de tener los datos en bruto, tenemos una función que nos calcula directamente el valor del índice de simetría visto en clase. Así para calcularlo nos situaremos en una nueva casilla, por ejemplo en B28 y escribimos =COEFICIENTE.ASIMETRÍA(B2:B20). Al pulsar *Enter* obtenemos el resultado que es 2.
5. **Coeficiente de apuntamiento.** Finalmente, para calcular el coeficiente de apuntamiento, debemos aplicar una de las funciones de Calc, la función curtosis. Para ello nos situaremos en la casilla B28 y escribimos =CURTOSIS(B2:B20) obteniendo el resultado del coeficiente de apuntamiento.

En la Figura 28 podemos observar como nos quedaría la hoja de cálculo una vez realizadas las diferentes operaciones.

## Ejemplo 2

Una cadena de distribución en grandes superficies compra frutos secos en bolsas de diez kilogramos y los envasa y comercializa en recipientes de cien gramos. El peso real en gramos de veinte de las bolsas que compra la cadena son:

9834, 9657, 9978, 10122, 9654, 9845, 9932, 9846, 9952, 9934, 9912, 9734, 9852, 9935, 9899, 9898, 9945, 9911, 9923, 9834

Se pide calcular el índice de simetría y la curtosis de los datos e interpretar los resultados.

En primer lugar creamos un documento OpenOffice Calc y escribimos estos datos en las casillas A2 a A21.

	A	B
1		TOTAL
2	Andalucía	22102,6
3	Aragón	6521,7
4	Asturias (Principado de)	3152,4
5	Baleares (Illes)	1557,2
6	Canarias	4513,7
7	Cantabria	1816,7
8	Castilla - La Mancha	2899
9	Castilla y León	9763,3
10	Cataluña	43037
11	Ceuta	22,4
12	Comunitat Valenciana	17810,8
13	Extremadura	1864,2
14	Galicia	8658,8
15	Madrid (Comunidad de)	49972,8
16	Mejilla	35,1
17	Murcia (Región de)	5755,1
18	Navarra (Comunidad Foral de)	4880,6
19	País Vasco	15570,6
20	Rioja (La)	1174
21		
22		
23		
24		
25	Media	10584,63
26	Varianza	188845779,64
27	Desv. Tipica	13742,12
28	Coef. Simetría	2
29	Curtosis	3,45
30		

Figura 28: Hoja de cálculo del ejemplo 1.

Como tenemos los datos en “bruto” podemos calcular directamente los dos coeficientes que nos piden usando las funciones detalladas en el ejercicio anterior. Así, el **Índice de simetría** lo podríamos calcular situándonos en la casilla *A24* y escribiendo =COEFICIENTE.ASIMETRÍA(A2:A21). Al pulsar *Enter* obtenemos el resultado que es  $-0,45$ . Este dato nos informa que la distribución es asimétrica por la izquierda

El **Coeficiente de apuntamiento** lo calcularemos situándonos por ejemplo en la casilla *A25* y escribiendo =CURTOSIS(B2:B20) obteniendo el resultado del coeficiente de apuntamiento, que en este caso vale  $1,37$ . En este caso tenemos que la distribución es puntiaguda o leptocúrtica.

### Ejemplo 3

Un empresario desea repartir unas bonificaciones entre sus empleados en base a la categoría y productividad de los mismos. Dicha distribución quedó de la siguiente forma:

Bonificaciones (Cientos Euros)	N. Empleados
10 - 15	3
15 - 25	8
25 - 28	12
28 - 32	15
32 - 40	7
40 - 55	5

Debemos calcular:

- Bonificación media por trabajador
- Bonificación más frecuente
- Bonificación tal que la mitad de las restantes sea inferior a ella
- La varianza
- El coeficiente de variación.
- El coeficiente de asimetría de Pearson y significado.
- Dibujar un gráfico para verificar el resultado obtenido en el apartado anterior.

Se trata de datos agrupados en forma de intervalos, de manera que calcularemos los estadísticos siguiendo el procedimiento explicado en clase:

1. El primero paso es la introducción de los datos y el cálculo de la tabla de frecuencias y de la marca de clase, tal y como se ha explicado en el apartado 1. La hoja de cálculo después de este procedimiento quedaría como se muestra en la Figura 29

	A	B	C	D	E	F
1	Bonificaciones	Marca de Clase	Empleados	Frec. Abs. Ac.	Frec. Rel.	Frec. Rel. Ac.
2	10 – 15	12,5	3	3	0,06	0,06
3	15 – 25	20	8	11	0,16	0,22
4	25 – 28	26,5	12	23	0,24	0,46
5	28 – 32	30	15	38	0,3	0,76
6	32 – 40	36	7	45	0,14	0,9
7	40 – 55	47,5	5	50	0,1	1
8			50			
9						
10						

Figura 29: Hoja de cálculo del ejemplo 3, después de la introducción de los datos y el cálculo de la tabla de frecuencias.

2. **Media.** Como se ha explicado en el apartado 2, el cálculo de la media en el caso de datos agrupado es muy sencilla. Únicamente debemos situarnos en una casilla cualquiera, por ejemplo, en la casilla A10 y escribir la fórmula `=SUMA.PRODUCTO(B2:B7;C2:C7)/SUMA(C2:C7)`. Así obtendremos que la bonificación media es de 29,1.
3. **Moda.** Observando la tabla vemos que el valor máximo se da en el intervalo 28 – 32. Al no tener todos los intervalos de igual amplitud, la moda se calcula como el valor medio del intervalo, es decir:  $\text{Moda} = \frac{28+32}{2} = 30$ .

4. **Mediana.** Como se ha explicado anteriormente, debemos encontrar el intervalo que tenga frecuencia relativa acumulada igual o superior a 0,5. En este caso, la mediana se encuentra en el intervalo 28 – 32 (ver figura 29).

Para calcular de modo más preciso la mediana utilizamos la fórmula dada en clase:

$$\text{mediana} = L_i + \frac{50\% \cdot n - N_{i-1}}{n_i} \cdot (L_{i+1} - L_i)$$

donde  $L_i$  y  $L_{i+1}$  denotan los límites inferior y superior del intervalo,  $n_i$  es la frecuencia del intervalo,  $N_{i-1}$  es la frecuencia acumulada en el intervalo anterior y  $n$  es la suma de todas las frecuencias absolutas.

En nuestro caso:  $L_i = 28$ ,  $L_{i+1} = 32$ ,  $n_i = 15$ ,  $N_{i-1} = 23$  y  $n = 50$ , por tanto

$$\begin{aligned}\text{mediana} &= 28 + \frac{50\% \cdot 50 - 23}{15} \cdot (32 - 28) = \\ &= 28 + \frac{25 - 23}{15} \cdot 4 = 28,53\end{aligned}$$

Por tanto, la bonificación tal que la mitad de las restantes sea inferior a ella es 28,53.

5. **Varianza.** Para calcular la varianza debemos decidir primero qué fórmula emplearemos (poblacional o muestral). En este caso, como disponemos de datos acerca de *todos* los empleados de la empresa consideramos que los datos se refieren a toda una población. Procedemos del siguiente modo para hacer el cálculo:

- a) Supongamos que el valor de la media se ha escrito en la casilla A10.
  - b) Insertamos una nueva columna a la derecha de la columna B. Para ello situamos el cursor sobre la parte superior de la columna C, hacemos clic en el botón derecho del ratón y elegimos la opción *insertar columnas*. Una nueva columna C aparece desplazando las que tiene a su derecha.
  - c) En la casilla C2 escribimos  $=(B2-\$A\$10)^2$ . Extendemos el cálculo al resto de casillas de la columna C situando el cursor en la esquina inferior derecha de la casilla C2 y, manteniendo el botón izquierdo del ratón pulsado, arrastrando el cursor hasta la casilla C7. De esta manera en la columna C tenemos todos los factores  $(x_i - \bar{x})^2$  de la fórmula de la varianza.
  - d) Finalmente, situamos el cursor en una casilla vacía cualquiera, por ejemplo la casilla A11 y escribimos la fórmula  
 $=SUMA.PRODUCTO(B2:B7;C2:C7)/SUMA(E2:E7)$ .
- Al pulsar *Enter* obtenemos el valor de la varianza poblacional en la casilla A11. El resultado final es 461,99.

En caso de tener que calcular la varianza muestral hubiéramos utilizado la siguiente fórmula:

=SUMA.PRODUCTO(B2:B7;C2:C7)/(SUMA(E2:E7)-1).

6. **Coeficiente de variación.** Para calcular este coeficiente debemos aplicar la fórmula vista en clase:

$$CV = \frac{s}{\bar{x}}.$$

Por tanto, nos situaremos en la casilla *A12*, por ejemplo, suponiendo que hemos escrito la media en la casilla *A10* y la varianza en la casilla *A11*, escribiremos =RAIZ(A11)/A10. Entonces al pulsar *Enter* obtendremos que el valor del coeficiente de variación es 0,74.

7. **Coeficiente de asimetría de Pearson.** Al tener los datos agrupados en intervalos no podemos aplicar las fórmulas usadas en los dos ejemplos anteriores. Para hacer el cálculo del coeficiente deberemos utilizar la fórmula vista en clase:

$$g_1 = \frac{m_3}{s^3} \quad \text{donde} \quad m_3 = \frac{1}{n} \sum_{j=1}^J n_j (X_j - \bar{x})^3.$$

Para ello procederemos como en el paso de la varianza y añadiremos una nueva columna *D* en la que calcularemos los valores de  $(X_j - \bar{x})^3$ . Para ello nos situaremos en la casilla *D2* y pondremos =(B2-\$A\$10)^3, suponiendo que tenemos la media escrita en la casilla *A10*. Posteriormente extendemos el cálculo al resto de las casillas de *D* situando el cursor en la esquina inferior derecha de la casilla *D2* y, manteniendo el botón izquierdo del ratón pulsado, arrastrando el cursor hasta la casilla *D7*. De esta manera en la columna *D* tenemos todos los factores  $(X_j - \bar{x})^3$ . Finalmente, nos situaremos en una casilla cualquiera, por ejemplo, en la casilla *A13* y escribiremos =(SUMA.PRODUCTO(B2:B7; D2:D7)/SUMA(E2:E7))/(RAÍZ(A11)^3), suponiendo que tenemos el valor de la varianza en la casilla *A11*. Así obtendremos que el valor del coeficiente de asimetría es 0,47, es decir, nos encontramos ante una distribución asimétrica por la derecha.

Finalmente debemos dibujar un gráfico representativo de los datos que tenemos. Al tener los datos agrupados en intervalos deberíamos dibujar un histograma, pero como ya hemos dicho OpenOffice no dispone ninguna herramienta para dibujarlos. Por tanto, procederemos como en el ejemplo 3 de la sección 1 y calcularemos un diagrama de barras, con la altura que le correspondería tener a los bloques en el histograma. Como, además, los diferentes intervalos no tienen la misma amplitud, deberemos calcular esta altura caso por caso. Para ello nos situaremos en una nueva columna y haremos el cálculo de las frecuencias absolutas (situadas en la columna *E*) dividido por la amplitud de cada intervalo (extremo derecho - extremo izquierdo). Los resultados son los que se observan en la columna *I* de la Figura 30.

	A	B	C	D	E	F	G	H	I
1	<u>Bonificaciones</u>	<u>Marca de Clase</u>			<u>Empleados</u>	<u>Frec. Abs. Ac.</u>	<u>Frec. Rel.</u>	<u>Frec. Rel. Ac.</u>	
2	10 - 15	12,5	275,56	-4574,3	3	3	0,06	0,06	0,6
3	15 - 25	20	82,81	-753,57	8	11	0,16	0,22	0,8
4	25 - 28	26,5	6,76	-17,58	12	23	0,24	0,46	4
5	28 - 32	30	0,81	0,73	15	38	0,3	0,76	3,75
6	32 - 40	36	47,61	328,51	7	45	0,14	0,9	0,88
7	40 - 55	47,5	338,56	6229,5	5	50	0,1	1	0,33
8					50				

Figura 30: Hoja de cálculo del ejemplo 3, después del cálculo de la altura de los bloques del histograma.

Ahora el histograma se puede calcular como un diagrama de barras. Seguimos el procedimiento descrito anteriormente, seleccionando las casillas  $A2$  a  $A7$  y  $I2$  a  $I7$ . Finalmente “unimos” las barras según el procedimiento explicado en el ejemplo 3 del apartado 1. El resultado sería el gráfico que se puede observar en la Figura 31.

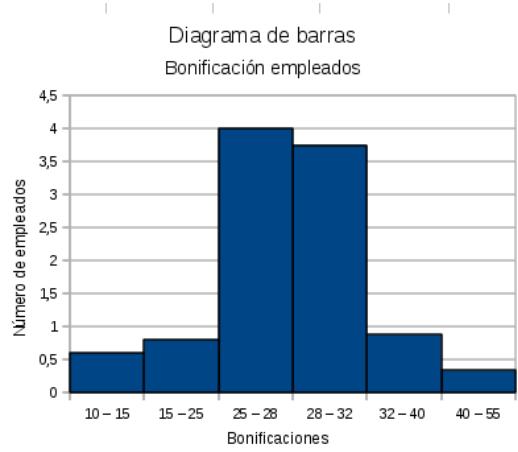


Figura 31: Histograma del ejemplo 3.

## 5. Análisis bivariante

### Ejemplo 1

Deseamos saber si existe alguna relación entre la reincidencia en los delitos y el sexo de los delincuentes, para ello vamos a calcular el coeficiente de correlación de Pearson para las variables “Sexo” y “Reincidencia” de los condenados en el año 2006 a partir de los siguientes datos.

Estadísticas judiciales 2006			
Estadística de lo Penal. Condenados. Resultados nacionales			
Condenados según tipo de delito, reincidencia y sexo			
Unidades: nº de condenados			
	Reincidente	No reincidente	
	Varón	Mujer	Varón
Total	26.771	1.352	85.230
			8.625
Notas:			
1) Reincidencia= Sujeto que ha sido condenado con anterioridad			
Fuente: Instituto Nacional de Estadística			

Utilizamos la aplicación OpenOffice Calc para resolver el ejercicio, siguiendo los siguientes pasos:

1. Abrimos la aplicación y escribimos los datos formando una tabla de contingencia.

	A	B	C
1		Varón	Mujer
2	Reincidente	26771	1352
3	No reincidente	85230	8625
4			

2. Para aplicar la fórmula de chi-cuadrado hemos de calcular primero las frecuencias absolutas parciales de cada variable. Las de la variable ‘Reincidencia’ se escriben en la columna D y las de ‘Sexo’ en la fila 4.

Los valores de la columna D se calculan en dos pasos:

- a) nos situamos en la casilla *D2* y escribimos =SUMA(B2:C2). Al pulsar *Enter* obtenemos el valor  $n_{1\bullet} = 28123$ .
- b) el cálculo para las demás casillas de la columna se hace automáticamente situándonos con el cursor en la esquina inferior derecha de la casilla *D2*, pulsando el botón izquierdo del ratón y arrastrando el cursor hasta la casilla *D3*. Obtenemos:  $n_{2\bullet} = 93855$ .

De manera similar se calculan los valores de la fila 4:

- nos situamos en la casilla  $B4$  y escribimos  $=SUMA(B2:B3)$ . Al pulsar *Enter* obtenemos el valor  $n_{\bullet 1} = 112001$ .
- el cálculo para las demás casillas de la fila se hace automáticamente situándonos con el cursor en la esquina inferior derecha de la casilla  $B4$ , pulsando el botón izquierdo del ratón y arrastrando el cursor hasta la casilla  $C4$ . Obtenemos:  $n_{\bullet 2} = 9977$ .

La suma de todos los valores de la tabla se calcula escribiendo la fórmula  $=SUMA(B2:C3)$  en la casilla  $D4$ . Obtenemos  $N = 121798$ .

La siguiente figura muestra el estado de la hoja de cálculo al finalizar este paso:

	A	B	C	D
1		Varón	Mujer	Suma
2	Reincidente	26771	1352	28123
3	No reincidente	85230	8625	93855
4	Suma	112001	9977	121978
5				

- Para calcular las frecuencias teóricas de la fórmula de chi cuadrado hacemos lo siguiente:

- escribimos la fórmula  $=B\$4*\$D2/\$D\$4$  en la casilla  $B6$
- a partir de la esquina inferior derecha de  $B6$  extendemos el cálculo a  $C6$
- seleccionamos simultáneamente  $B6$  y  $C6$  y a partir de la esquina inferior derecha de  $C6$  extendemos el cálculo a  $B7$  y  $C7$

Al final de este paso la hoja de cálculo muestra los siguientes valores:

	A	B	C	D
1		Varón	Mujer	Suma
2	Reincidente	26771	1352	28123
3	No reincidente	85230	8625	93855
4	Suma	112001	9977	121978
5				
6		25822,72	2300,28	
7		86178,28	7676,72	
8				

- A continuación, si llamamos  $e_{ij}$  a las frecuencias teóricas, debemos calcular los cocientes  $\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$ . Procedemos de la siguiente forma:

- escribimos la fórmula  $=(B2-B6)^2/B6$  en la casilla  $B9$

- b) a partir de la esquina inferior derecha de *B9* extendemos el cálculo a *C9*
- c) seleccionamos simultáneamente *B9* y *C9* y a partir de la esquina inferior derecha de *C9* extendemos el cálculo a *B10* y *C10*
5. Finalmente calculamos chi-cuadrado y el coeficiente de correlación de Pearson:
- Chi-cuadrado se calcula sumando los valores obtenidos en el paso anterior: nos situamos en la casilla *C12*, escribimos =SUMA(*B9:C10*) y al pulsamos *Enter*. Obtenemos  $\chi^2 = 553,32$ .
  - El coeficiente C de contingencia se calcula aplicando la fórmula vista en clase: escribimos =RAÍZ(*C12/(D4+C12)*) en *C13*, pulsamos *Enter* y obtenemos  $C_P = 0,07$ .

La hoja de cálculo final muestra el siguiente aspecto:

	A	B	C	D	E
1		Varón	Mujer	Suma	
2	Reincidente	26771	1352	28123	
3	No reincidente	85230	8625	93855	
4	Suma	112001	9977	121978	
5					
6		25822,72	2300,28		
7		86178,28	7676,72		
8					
9		34,82	390,92		
10		10,43	117,14		
11					
12		Chi cuadrado	553,32		
13		C conting.	0,07		
14		C max	0,71		
15		%C	9,5		
16					

### Comentario.

El valor de  $C_P$  obtenido, 0,07, indica que las variables ‘Reincidencia’ y ‘Sexo’ del delincuente son prácticamente independientes: la proporción de reincidentes no es muy diferente en el caso de hombres que en el caso de mujeres.

### Ejemplo 2

Hallar la covarianza y el coeficiente de correlación para las variables ‘Cantidad de precipitaciones’ y ‘Número de incendios’ en Mallorca a partir de los datos de la siguiente tabla (fuentes: Conselleria de Medi Ambient y Instituto Nacional de Meteorología).

Año	Precipitaciones (mm)	Número de incendios
1993	423,6	134
1994	526,1	110
1995	296,7	86
1996	605,1	58
1997	446,6	83
1998	455,8	77
1999	306,5	104
2000	225,7	113
2001	397,1	83
2002	702,2	40
2003	472,2	66
2004	403,5	100
2005	294,6	94

Con OpenOffice Calc es muy sencillo calcular la covarianza y el coeficiente de correlación a partir de datos brutos:

1. Abrimos la aplicación y escribimos los datos de precipitación y número de incendios en las columnas A y B de la tabla, respectivamente:

	A	B	
1	Precipitaciones	Número incendios	
2	423,6	134	
3	526,1	110	
4	296,7	86	
5	605,1	58	
6	446,6	83	
7	455,8	77	
8	306,5	104	
9	225,7	113	
10	397,1	83	
11	702,2	40	
12	472,2	66	
13	403,5	100	
14	294,6	94	
15			

2. En este ejemplo consideramos que los datos proporcionados corresponden a una población y no a una muestra por lo que calcularemos covarianza y correlación poblacionales. Para ello procedemos del siguiente modo:

- a) la covarianza se calcula situándonos en una casilla cualquiera, por ejemplo *D2*, escribiendo la fórmula `=COVAR(A2:A14;B2:B14)` y pulsando *Enter*. El resultado es  $-1966,63$ . La covarianza muestral se calcularía multiplicando este valor por  $\frac{N}{N-1}$ .

- b) el coeficiente de correlación se calcula situándonos en una casilla cualquiera, por ejemplo *D3*, escribiendo la fórmula  
 $=COEF.DE.CORREL(A2:A14;B2:B14)$  y pulsando *Enter*.  
 El resultado es  $-0,64$ .

La hoja de cálculo final muestra el siguiente aspecto:

	A	B	C	D	
1	Precipitaciones	Número incendios			
2	423,6	134		-1966,63	
3	526,1	110		-0,64	
4	296,7	86			
5	605,1	58			
6	446,6	83			
7	455,8	77			
8	306,5	104			
9	225,7	113			
10	397,1	83			
11	702,2	40			
12	472,2	66			
13	403,5	100			
14	294,6	94			
15					

### Comentario.

Este resultado indica una cierta correlación lineal negativa entre las variables: a un mayor nivel de precipitaciones corresponde un menor número de incendios.

### Ejemplo 3

Calcular la recta de regresión lineal para los datos del ejercicio anterior y predecir a partir de ella el número de incendios que tendremos un año en que las precipitaciones sean de 550 mm. Dibujar el diagrama de dispersión y representar sobre él la recta de regresión.

Calculamos la recta de regresión con la fórmula vista en clase. Para utilizar la fórmula debemos calcular:

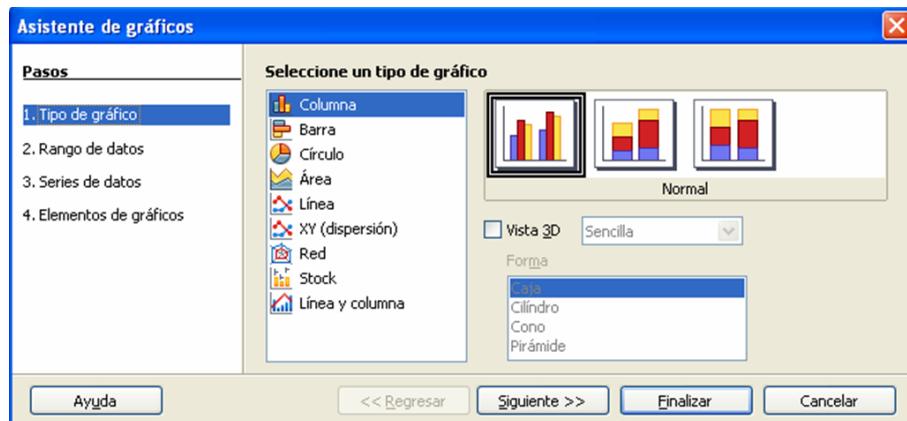
1. la covarianza ( $-1966,63$ , calculada en el ejemplo anterior),
2. la varianza de la primera variable (fórmula  $=VARP(A2:A14)$ , resultado  $16272,15$ ),
3. las medias de cada variable (fórmulas  
 $=PROMEDIO(A2:A14)$  y  $=PROMEDIO(B2:B14)$ , respectivamente, resultados  $427,36$  y  $88,31$ )

- calculamos los parámetros  $a$  y  $b$  de la recta. Si los valores de covarianza, varianza y medias están en las casillas  $D2$ ,  $D4$ ,  $D5$  y  $D6$ , respectivamente y el valor de  $a$  se escribe en la casilla  $D7$ :  $=D2/D4$  y  $=D6-D7*D5$ . Los resultados son  $a = -0,12$  y  $b = 139,96$ .

La ecuación de la recta de regresión es por tanto:  $\hat{Y} = -0,12X + 139,96$ . De manera que el valor estimado para  $x = 550$  será:  $\hat{Y} = -0,12 \cdot 550 + 139,96 = 73,96$ .

El diagrama de dispersión se dibuja fácilmente con Calc:

- Partimos de la hoja de cálculo final del ejemplo anterior.
- Hacemos clic sobre el icono  del menú *Insertar* y a continuación sobre una casilla cualquiera para insertar el gráfico en esa posición. Aparece el siguiente cuadro de diálogo:



- Seleccionamos la opción  y la opción *Sólo puntos*.
- En el rango de datos escribimos  $A1:B14$  y pulsamos el botón *Siguiente*.
- En el diálogo *Series de datos* hacemos clic sobre *Valores X* y escribimos  $A2:A14$  en *Rango para valores X*. Repetimos el proceso para los valores Y, cuyo rango es  $B2 : B14$ , y pulsamos *Siguiente*.
- En el último diálogo desactivamos la opción *Mostrar leyenda* y escribimos **Precipitaciones** e **Número incendios**, respectivamente, en las opciones *Título del Eje X* y *Título del Eje Y*. También desactivamos la opción *Eje Y*.
- Pulsamos la tecla *Finalizar* y el diagrama aparece en la posición seleccionada. Ahora podemos reescalarlo con el cursor a un tamaño mayor.
- Si deseamos dibujar la recta de regresión procedemos del siguiente modo:

- a) Nos situamos sobre el diagrama y hacemos clic sobre cualquiera de los puntos dibujados. Todos los puntos quedarán marcados.
- b) Hacemos clic con el botón derecho del ratón sobre cualquiera de los puntos y aparecerá un menu desplegable en el que seleccionamos la opción *Insertar Línea de Tendencia* ...
- c) Dentro de las opciones de *Línea de tendencia* seleccionamos el ícono



(Lineal) y aceptamos. La recta de regresión se dibuja sobre el diagrama de dispersión.

El resultado final del proceso anterior se muestra en la siguiente figura:

