

ESTADÍSTICA APLICADA

R. Alberich

J.L. Lisani

Índice general

Prefacio: Estadística en Seguridad y Ciencias Policiales

El enfoque racional de los problemas en cualquier área de las Ciencias implica un paso previo de recopilación de datos y análisis de los mismos que permite conocer en profundidad el problema y eventualmente solucionarlo.

Ya sea para descubrir cómo se mueven los planetas, cómo se desintegran los átomos, cómo distribuir el presupuesto municipal o cómo asignar los recursos policiales para reducir la criminalidad, en el origen de la solución está la recopilación de datos sobre el problema. Las Ciencias Sociales, en las que se enmarcan las Ciencias Policiales y de Seguridad, no son ajenas a este método de trabajo.

Sin embargo, los datos por sí solos poco aportan a la solución de los problemas. Es su organización lo que permite descubrir tendencias, singularidades, etc. que conducirán a la solución. La **Estadística Descriptiva** es la rama de las Matemáticas que proporciona las técnicas necesarias para recopilar, organizar, representar, analizar e interpretar los datos.

En la mayoría de ocasiones, y por razones prácticas, el análisis estadístico se hace sobre un conjunto de datos inferior al total disponible. El ejemplo típico son las encuestas sobre intención de voto, realizadas sobre unos pocos miles de personas que *representan* al total de la población. Los resultados obtenidos se generalizan después a toda la población. La **Estadística Inferencial** permite conocer el grado de fiabilidad de estas generalizaciones.

En este curso estudiaremos las técnicas básicas de la Estadística Descriptiva e Inferencial, de modo que aprenderemos a organizar y analizar conjuntos de datos y a conocer el grado de fiabilidad de las generalizaciones realizadas a partir de ellos.

El curso se organiza en tres módulos:

1. El módulo I da una introducción a la notación habitual de la estadística y presenta las técnicas básicas de la Estadística Descriptiva.
2. En el módulo II se avanza en el estudio de la Estadística Descriptiva y se dan los fundamentos de Probabilidad necesarios para la Estadística Inferencial.
3. El módulo III se dedica al estudio de la Estadística Inferencial

Módulo I

El módulo I define los conceptos básicos de la Estadística Descriptiva y presenta las técnicas básicas para la organización, representación y análisis de los datos estadísticos.

Las definiciones teóricas se acompañan de múltiples ejemplos para facilitar la comprensión. Asimismo, se explica cómo resolver los problemas con la ayuda de aplicaciones informáticas y cómo acceder a bases de datos estadísticos oficiales.

Capítulo 1

El lenguaje de la estadística

En múltiples ámbitos de las Ciencias o la Gestión es necesario el estudio de temas relacionados con cada disciplina. Por ejemplo:

- un experto en Seguridad puede estar interesado en estudiar la criminalidad en una ciudad;
- un Economista, el nivel de paro de un país;
- un Ecologista, la deforestación de la selva amazónica;
- un Físico, la desintegración de elementos radioactivos en una central nuclear;
- un Ingeniero, la calidad de las piezas producidas en una fábrica, etc.

Sea cual sea el tema a estudiar, la información acerca del mismo se obtiene realizando múltiples **observaciones** de algunas de sus características. Estas observaciones proporcionan una serie de **datos** que son organizados, analizados e interpretados usando unas técnicas estadísticas comunes.

Antes de comenzar a explicar cualquier técnica estadística es necesario definir algunos conceptos básicos y un vocabulario estadístico elemental. Entre otros aspectos, en este tema estudiaremos las nociones de población y muestra, la diferencia entre estadística descriptiva e inferencial y las distinción entre datos y variables.

1.1. Población y muestra

El término **población** hace referencia al conjunto total de *elementos* objeto del estudio estadístico. Por ejemplo:

- el número total de personas víctimas de un delito, en un estudio sobre la criminalidad;
- el número de personas en paro, en un estudio sobre el desempleo;
- el número de árboles talados en el último año, en un estudio sobre la deforestación;
- el número de partículas subatómicas generadas por un reactor nuclear, en un estudio sobre la desintegración de elementos radioactivos;
- el número de piezas defectuosas fabricadas, en un estudio sobre la calidad de la producción.

En la mayoría de ocasiones es demasiado difícil, caro o imposible obtener información de todos los elementos de la población, por lo que el estudio estadístico se realiza sobre un subconjunto de la población. Este subconjunto se denomina **muestra**.

Como veremos en el apartado siguiente, es frecuente que se deseen extraer conclusiones relativas a la población total a partir de los datos de la muestra. Para que estas conclusiones sean fiables es necesario que la muestra sea **representativa** de la población. Se considera que si los elementos de la muestra han sido elegidos al azar entre toda la población (**muestra aleatoria simple**), se obtiene una muestra representativa, aunque hay otras formas de tomar muestras. Otro factor a tener en cuenta es el tamaño de la muestra, cuanto mayor sea mejor representará al conjunto de la población.

1.2. Tipos de estadísticas: descriptiva e inferencial

Tradicionalmente la disciplina de Estadística se divide en dos ramas: descriptiva e inferencial.

La Estadística Descriptiva tiene por objeto la descripción de los datos recopilados, para ello proporciona técnicas que permiten:

- la **organización** de los datos mediante tablas y representaciones gráficas;
- el **análisis** de los datos mediante el cálculo de valores representativos como son las medidas de tendencia central y de dispersión.

La Estadística Inferencial parte de los datos obtenidos a partir de una muestra e intenta extraer conclusiones de las características generales de toda la población. También proporciona métodos para medir la fiabilidad de las conclusiones obtenidas y relacionar esta fiabilidad con el tamaño de la muestra estudiada.

Por ejemplo, en el caso del experto en Seguridad que hace un estudio sobre la criminalidad en una ciudad, la manera de hacer el estudio puede ser la siguiente:

1. En lugar de recopilar los datos relativos a los delitos cometidos en toda la ciudad, recoge los datos de unos pocos barrios (elegidos al azar o siguiendo algún criterio determinado);
2. Organiza y analiza los datos recogidos utilizando técnicas de estadística descriptiva;
3. Utiliza técnicas de estadística inferencial para generalizar los resultados obtenidos a toda la ciudad y estimar el grado de fiabilidad de esta generalización.

1.3. Datos y variables

Los **datos** se definen como las unidades de información recopiladas al hacer un estudio estadístico. Una **variable** es una característica o atributo que permite clasificar en diferentes categorías los elementos de la muestra o población en función de los datos recopilados. Por ejemplo, el conjunto de personas víctimas de un crimen se puede clasificar en función del sexo de las personas (hombre/mujer), la edad (menores de 20 años, entre 20 y 50, mayores de 50), etc. Cada uno de estos atributos (sexo, edad, etc.) es una variable.

Existen distintos tipos de variables que se pueden clasificar siguiendo tres criterios:

- Tipo de dato
 - Nominales (cualitativas o de atributos): cuando los datos no son numéricos y la comparación entre sus valores sólo puede ser de igualdad o desigualdad.
Por ejemplo: sexo, color de los ojos, afiliación política, lugar de residencia, etc,...
 - Ordinales: cuando los datos no son numéricos pero la comparación entre ellos establece un orden.
Por ejemplo: estado de ánimo (valores posibles: depresivo, normal y eufórico), estudios (valores posibles: ninguno, primarios, secundarios, superiores), etc...
 - Cuantitativas: cuando los datos son numéricos. Dentro de las variables cuantitativas hay dos tipos más
 - Discretas: cuando entre dos posibles valores no hay otro. Por ejemplo: número de hijos de una familia, número de letras de una palabra en un texto, etc,...
 - Continuas: cuando entre dos posibles valores, siempre podemos encontrar otro valor posible. Por ejemplo: altura, intereses de una cuenta bancaria, etc,...
- Dimensión
 - Unidimensionales: si sólo se considera una única característica.
Ejemplos: altura, edad, etc,...
 - Multidimensionales: si se consideran conjuntamente varias características.
Ejemplos: edad y altura, altura y peso, edad, altura y sexo, etc,...
- Tiempo
 - Atemporales: cuando los datos no están referidos, o no se considera, el momento de tiempo en el que fueron obtenidos.
Ejemplos: color de los ojos de cierto conjunto de individuos, peso de los estudiantes de la clase de la clase de hoy, etc,...
 - Temporales o series cronológicas: en caso contrario.
Ejemplos: P.I.B. anual de España durante el periodo 1980 hasta 2004, número de turistas llegados al aeropuerto de Palma durante los años 1970 al 2004, etc,...

1.4. Ejercicios propuestos

Ejercicio 1

Identificar la población y la muestra estudiados en los siguientes casos:

- a) En un estudio sobre el consumo de drogas en un instituto se hace una encuesta al 30 % por ciento de los alumnos de 2º de bachillerato.
- b) En un estudio sobre el consumo de drogas entre los jóvenes de Ciutadella (menores de 35 años) se entrevista al 10 % de los clientes de los principales locales de copas.
- c) En un estudio a nivel nacional sobre la influencia del alcohol en los accidentes de tráfico se realizan 10.000 controles de alcoholemia en diferentes carreteras del país.

Ejercicio 2

Decidir si para estudiar los siguientes casos se utilizan herramientas de estadística descriptiva o inferencial:

- a) Un profesor de universidad debe proporcionar a su jefe de Departamento un informe sobre el número de alumnos matriculados y sus calificaciones en el periodo 2005-2007. Para ello utilizará estadística
- b) Una empresa desea conocer los hábitos de trabajo de sus trabajadores. Para ello les hace llenar una encuesta sobre sus horas de llegada y salida, tiempo dedicado a responder el teléfono o el correo electrónico, tiempo dedicado a reuniones de trabajo con los jefes u otros compañeros, etc. Los datos obtenidos se organizarán y analizarán usando estadística
- c) La Dirección General de Tráfico desea evaluar la eficiencia a nivel nacional de la última campaña de prevención de accidentes a partir de los datos en una serie de municipios. Para ello utilizará herramientas de la estadística

Ejercicio 3

Identificar al menos tres variables que pueden aparecer en los siguientes estudios estadísticos:

- a) Consumo de drogas en una ciudad.
- b) Satisfacción laboral de los empleados de una empresa.
- c) Notas obtenidas por los alumnos de una asignatura.

Ejercicio 4

Clasificar las siguientes variables según su tipo, dimensión y nivel temporal:

- a) Número de personas que han sufrido un accidente de tráfico en los últimos 5 años.
- b) Nivel profesional de un militar (por ejemplo: soldado, cabo, sargento, etc.).
- c) Número de goles conseguidos por un jugador de fútbol en la liga 2006-07.
- d) Religión de un individuo (por ejemplo: católico, musulmán, budista, etc.).
- e) El peso y la altura de las participantes en un desfile de moda.
- f) Cantidad de dinero gastada por una Administración en obras públicas el último año.

1.5. Soluciones de los ejercicios

Ejercicio 1

- a) Población: el total de los alumnos del instituto. Muestra: el 30 % por ciento de los alumnos de 2º de bachillerato
- b) Población: los habitantes de Ciutadella menores de 35 años. Muestra: el 10 % de los clientes de los principales locales de copas.

c) Población: todos los conductores del país. Muestra: 10.000 conductores.

Ejercicio 2

- a) Descriptiva.
- b) Descriptiva.
- c) Inferencial.

Ejercicio 3

- a) Sexo, edad, nivel de estudios, nivel de ingresos, etc.
- b) Sexo, edad, antigüedad en la empresa, categoría profesional, salario, etc.
- c) Tiempo dedicado al estudio, compaginación de estudios y trabajo, número de veces que se ha cursado la asignatura, nota en los exámenes de acceso a la universidad, etc.

Ejercicio 4

- a) Cuantitativa discreta, unidimensional, temporal.
- b) Ordinal, unidimensional, atemporal.
- c) Cuantitativa discreta, unidimensional, temporal.
- d) Nominal, unidimensional, atemporal.
- e) Cuantitativa continua, multidimensional (bidimensional), atemporal.
- f) Cuantitativa continua, unidimensional, temporal.

Capítulo 2

Organización y descripción de datos estadísticos

El primer paso en la realización de un estudio estadístico consiste en la recopilación de los datos que caracterizan el tema bajo estudio. Estos datos se obtienen de múltiples observaciones de los elementos de una población o una muestra.

En ocasiones, la recopilación de datos está controlada por el equipo de personas responsable del estudio. Por ejemplo, en un estudio sobre la satisfacción de los consumidores de un determinado producto de limpieza se pueden hacer encuestas a un cierto número de clientes de grandes centros comerciales, o solicitar a los compradores del producto el envío de un cuestionario de satisfacción.

En algunas ocasiones, no obstante el estudio puede hacerse a partir de datos recopilados por organismos o instituciones independientes. Tal sería el caso de un estudio sobre siniestralidad laboral, donde los datos podrían proceder del Ministerio de Trabajo.

Una fuente importante de datos estadísticos a nivel nacional la ofrece el Instituto Nacional de Estadística. A través de su página web (www.ine.es) se pueden obtener datos oficiales sobre economía, sociedad, medio ambiente, etc. En la sección ?? del tema explicamos como acceder a estos datos. También la Dirección General de Tráfico en su página web (www.dgt.es) ofrece datos estadísticos sobre seguridad vial. En Baleares, una fuente importante de datos oficiales es el Institut Balear de Estadística (IBAE, <http://www.caib.es/ibae/ibae.htm>).

Existen protocolos de recogida de datos que aseguran que los datos recopilados a partir de una muestra representan de manera fiel a toda la población bajo estudio. En este curso no estudiaremos tales protocolos sino que consideraremos que ya disponemos de una serie de datos y veremos cómo organizarlos e interpretarlos.

En el presente tema explicaremos diferentes maneras de organizar y representar los datos obtenidos de una manera visualmente agradable y que facilite su posterior interpretación. Existen dos formas básicas de organizar los datos: mediante tablas y mediante gráficas. En estas tablas y gráficas se muestran una serie de parámetros calculados a partir de los datos originales (denominados también **datos brutos** o **datos en bruto**, *raw data* en inglés). A continuación definimos los parámetros más habituales y explicamos como representarlos mediante tablas y gráficas.

Tabla 2.1: Inmigrantes según su nacionalidad. Datos brutos.

Persona	Nacionalidad
1	Colombia
2	Rumanía
3	Colombia
4	Senegal
5	Perú
6	Colombia
7	Ecuador
8	Ecuador
9	Marruecos
10	Perú
:	:
998	Colombia
999	Perú
1000	Rumanía

2.1. Frecuencias y porcentajes

El primer parámetro que definiremos es la **frecuencia absoluta**. Consideremos el siguiente ejemplo: deseamos hacer un estudio sobre la inmigración y tomamos una muestra formada por 1000 inmigrantes a los que preguntamos por su nacionalidad. Supongamos que algunos de los resultados obtenidos son los que se muestran en la tabla ??.

La tabla ?? consta de 1000 datos, de los cuales se muestran los 10 primeros y los 3 últimos. Leer los datos de una tabla tan grande es complicado. Una manera de mostrar los datos de una forma más resumida consiste en contar cuántas personas hay de cada nacionalidad. Supongamos por ejemplo que tuviéramos: 350 personas de Colombia, 250 de Ecuador, 120 de Perú, 100 de Argentina, 80 de Rumanía, 70 de Marruecos y 30 de Senegal. Las cantidades 350, 250, 120, 100, 80, 70 y 30, que representan la cantidad de personas de cada nacionalidad, se denominan **frecuencias absolutas**. Las frecuencias absolutas resumen los datos brutos. En este ejemplo, los 7 valores de las frecuencias absolutas describen el comportamiento de 1000 valores de datos brutos.

En general, dada una variable de un estudio estadístico (en nuestro ejemplo la variable es *Nacionalidad*), la frecuencia absoluta de los valores de la variable es el número de veces que aparece cada valor en la muestra considerada.

Otro ejemplo sería el siguiente: se desea estudiar la edad de los estudiantes de la UIB (variable: *Edad*) y se encuentran los siguientes valores (ya agrupados en frecuencias absolutas) para una muestra de 626 estudiantes: 120 personas de 18 años, 150 de 19 años, 90 de 20 años, 70 de 21 años, 65 de 22 años, 50 de 23 años, 30 de 24 años, 20 de 25 años, 10 de 26 años, 7 de 27 años, 8 de 28 años, 2 de 29 años, 1 de 30 años, 1 de 34 años, 1 de 35 años y 1 de 40 años. Los valores que toma la variable *Edad* en este caso son: 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35 y 40. Y sus frecuencias absolutas respectivas son: 120, 150, 90, 70, 65, 50, 30, 20, 10, 7, 8, 2, 1, 1, 1.

En general se denota como n_i la frecuencia absoluta del valor i -ésimo de una variable. Por ejemplo, en el estudio sobre la inmigración, tenemos que $n_{\text{Argentina}} = 100$; y en el estudio sobre los estudiantes de la UIB $n_{25} = 20$. Observemos que la suma de todas las frecuencias absolutas es igual al tamaño total de la muestra (que denotamos como N): $\sum_i n_i = N$.

En ocasiones es habitual agrupar distintos valores de una variable en un mismo **intervalo**. Por ejemplo, en el estudio sobre la edad en la UIB, podríamos agrupar los resultados por grupos de edad: de 18 a 21 años hay 430 estudiantes, de 22 a 25 años 165 y mayores de 25 31 estudiantes. En este caso diríamos que 430, 165 y 31 son las frecuencias absolutas de los intervalos 18 – 21, 22 – 25 y *mayor de* 25. Los valores de frecuencia variarán con una elección distinta de los intervalos. En la siguiente sección se muestran otras dos maneras de agrupar los valores para este mismo ejemplo.

A partir de la frecuencia absoluta se pueden definir otros parámetros que caracterizan la muestra:

- **Frecuencia relativa o proporción.**

$$f_i = \frac{n_i}{N}$$

En el ejemplo sobre la inmigración: $f_{\text{Colombia}} = \frac{350}{1000} = 0,35$, $f_{\text{Ecuador}} = \frac{250}{1000} = 0,25$, etc.

En el ejemplo sobre la edad en la UIB: $f_{18-21} = \frac{430}{626} = 0,6869$, $f_{22-25} = \frac{165}{626} = 0,2636$ y $f_{\text{mayor } 25} = \frac{31}{626} = 0,0399$.

Observemos que la suma de todos los valores f_i es siempre igual a 1.

- **Porcentaje.** Es la frecuencia relativa multiplicada por 100.

$$p_i = \frac{n_i}{N} \times 100 \%$$

En el ejemplo sobre la inmigración: $p_{\text{Colombia}} = 35\%$, $p_{\text{Ecuador}} = 25\%$, etc.

En el ejemplo sobre la edad en la UIB: $p_{18-21} = 68,69\%$, $p_{22-25} = 26,36\%$ y $p_{\text{mayor } 25} = 3,99\%$.

Observemos que la suma de todos los valores p_i es siempre igual a 100.

- **Frecuencia absoluta acumulada.** En el caso de variables ordinales o cuantitativas es posible ordenar los valores siguiendo algún criterio. En este caso, si n_1 representa la frecuencia absoluta del primer valor, n_2 la del segundo, etc, la frecuencia absoluta acumulada para el valor i se define como

$$N_i = \sum_{j \leq i} n_j$$

No tiene sentido aplicar esta definición al caso del ejemplo sobre la inmigración, pues la variable *Nacionalidad* es una variable cualitativa.

En el ejemplo sobre la edad en la UIB, si el orden de los intervalos es 18 – 21, 22 – 25 y *mayor* 25: $N_{18-21} = 430$, $N_{22-25} = 430 + 165 = 595$ y $N_{\text{mayor } 25} = 430 + 165 + 31 = 626$.

Observemos que la frecuencia absoluta acumulada del último valor de la variable es siempre igual al tamaño de la muestra N .

- **Frecuencia relativa (proporción) acumulada.** De manera similar al estadístico anterior definimos la frecuencia relativa acumulada como:

$$F_i = \frac{N_i}{N}$$

En el ejemplo sobre la edad en la UIB, considerando el mismo orden que en el estadístico anterior: $F_{18-21} = \frac{430}{626} = 0,6869$, $F_{22-25} = \frac{595}{626} = 0,9505$ y $F_{\text{mayor } 25} = \frac{626}{626} = 1$.

Se puede comprobar fácilmente que la frecuencia relativa acumulada para un valor i es igual a la suma de las frecuencias relativas de todos los valores anteriores (incluido i): $F_i = \sum_{j \leq i} f_j$.

Además, la frecuencia relativa acumulada del último valor de la variable es siempre igual 1.

- **Porcentajes acumulado.** Es la frecuencia relativa acumulada multiplicada por 100.

$$P_i = \frac{N_i}{N} \times 100\%$$

En el ejemplo sobre la edad en la UIB, considerando el mismo orden que en el estadístico anterior: $P_{18-21} = 68,69\%$, $P_{22-25} = 95,05\%$ y $P_{\text{mayor } 25} = 100\%$.

El porcentaje acumulado del último valor de la variable es siempre igual a 100%.

2.2. Tablas de frecuencia

Los parámetros anteriores se pueden representar mediante tablas. En la primera columna de la tabla se enumeran los distintos valores de la variable considerada y en las columnas siguientes se muestran las frecuencias o porcentajes asociados a cada valor de la variable.

Para los ejemplos utilizados en la sección anterior obtendríamos las tablas de frecuencias y porcentajes ?? y ??, respectivamente.

Tabla 2.2: Inmigración por nacionalidades

Nacionalidad	Frecuencia absoluta	Frecuencia relativa	Porcentaje
Colombia	350	0,35	35,00%
Ecuador	250	0,25	25,00%
Perú	120	0,12	12,00%
Argentina	100	0,10	10,00%
Rumanía	80	0,08	8,00%
Marruecos	70	0,07	7,00%
Senegal	30	0,03	3,00%
Suma	1000	1	100,00%

Ya se ha comentado en la sección anterior que para el ejemplo sobre la edad de los estudiantes de la UIB podrían haberse utilizado otras maneras de agrupar los valores de edad en intervalos. La tablas ?? y ?? muestran las frecuencias y porcentajes que obtendríamos con dos agrupaciones diferentes. En el primer caso se muestran los parámetros obtenidos para cada valor de la variable

Tabla 2.3: Estudiantes UIB por edad

<i>Edad</i>	Frecuencia absoluta	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
18-21	430	0,69	68,69%	430	0,69	68,69%
22-25	165	0,26	26,36%	595	0,95	95,05%
Mayor 25	31	0,05	4,95%	626	1	100,00%
Suma	626	1	100,00%			

Tabla 2.4: Estudiantes UIB por edad

<i>Edad</i>	Frecuencia absoluta	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
18	120	0,19	19,17%	120	0,19	19,17%
19	150	0,24	23,96%	270	0,43	43,13%
20	90	0,14	14,38%	360	0,58	57,51%
21	70	0,11	11,18%	430	0,69	68,69%
22	65	0,10	10,38%	495	0,79	79,07%
23	50	0,08	7,99%	545	0,87	87,06%
24	30	0,05	4,79%	575	0,92	91,85%
25	20	0,03	3,19%	595	0,95	95,05%
26	10	0,02	1,60%	605	0,97	96,65%
27	7	0,01	1,12%	612	0,98	97,76%
28	8	0,01	1,28%	620	0,99	99,04%
29	2	0,00	0,32%	622	0,99	99,36%
30	1	0,00	0,16%	623	1	99,52%
34	1	0,00	0,16%	624	1	99,68%
35	1	0,00	0,16%	625	1	99,84%
40	1	0,00	0,16%	626	1	100,00%
Suma	626	1	100,00%			

Tabla 2.5: Estudiantes UIB por edad

<i>Edad</i>	Frecuencia absoluta	Frecuencia relativa	Porcentaje	Frecuencia absoluta acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
18-19	270	0,43	43,13%	270	0,43	43,13%
20-21	160	0,26	25,56%	430	0,69	68,69%
22-23	115	0,18	18,37%	545	0,87	87,06%
24-25	50	0,08	7,99%	595	0,95	95,05%
26-27	17	0,03	2,72%	612	0,98	97,76%
28-29	10	0,02	1,60%	622	0,99	99,36%
Mayor 29	4	0,01	0,64%	626	1	100,00%
Suma	626	1	100,00%			

Edad sin agrupar en intervalos. En el segundo caso se agrupan las edades en intervalos de 2 años, salvo los últimos valores que se agrupan como *mayores de 29 años*.

Observamos como los valores de la tabla ?? son más difíciles de leer que los de las tablas ?? y ??, pues hay muchos más datos. En cambio en las tablas ?? y ?? se pierde el detalle de que hay más estudiantes de 19 años que de 18. En general debe buscarse un tamaño de los intervalos que permita un compromiso entre la claridad de la representación y los detalles que ofrece.

2.3. Representaciones gráficas

La frecuencias y porcentajes de las variables estadísticas pueden representarse gráficamente de diversas maneras:

- **Diagramas de barras.** En este tipo de gráfica los valores de la variable se representan sobre el eje horizontal y a cada valor se le asocia una barra vertical cuya altura es proporcional a la frecuencia (absoluta o relativa) o porcentaje del valor. Las dos siguientes figuras muestran los diagramas de barras correspondientes a los ejemplos mostrados en la sección anterior.

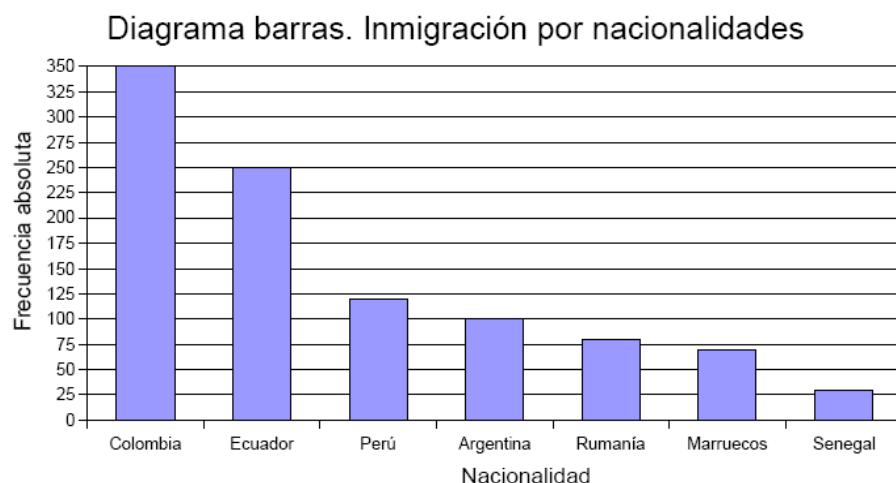


Figura 2.1: Diagrama de barras de frecuencias absolutas de la tabla ??

Para el ejemplo sobre los estudiantes de la UIB se muestran también (figuras ?? y ??) los diagramas correspondientes a las agrupaciones de valores de las tablas ?? y ???. Como ya se ha comentado en la sección anterior, el uso de intervalos grandes permite observar mejor la distribución de los valores pero impide apreciar los detalles.

Una variante del diagrama de barras es el **diagrama de Pareto**. En este caso las frecuencias o porcentajes están ordenadas de mayor a menor y además se dibuja una línea indicativa de la frecuencia o porcentaje acumulados. El diagrama de Pareto correspondiente a los porcentajes del ejemplo sobre la edad de los estudiantes de la UIB se muestra en la figura ??.

- **Diagramas de barras dobles** Las tablas mostradas en la sección anterior mostraban valores de frecuencia relativos a una única variable. En ocasiones se desea mostrar de



Figura 2.2: Diagrama de barras de frecuencias absolutas de la tabla ??

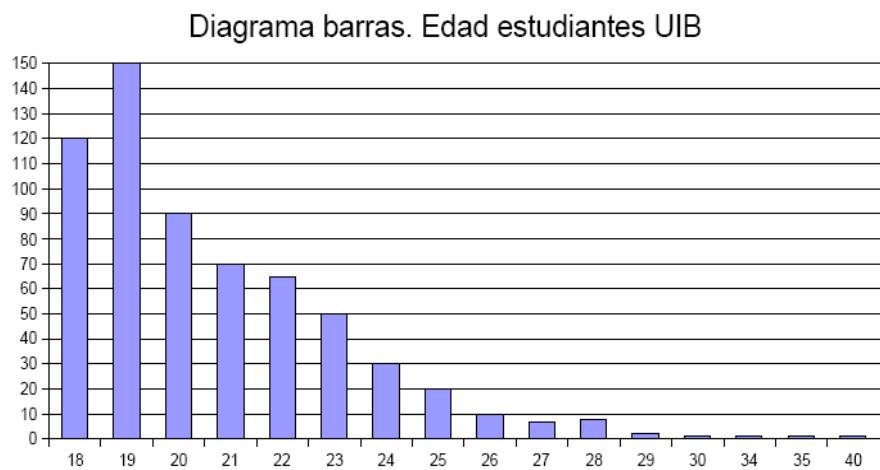


Figura 2.3: Diagrama de barras de frecuencias absolutas de la tabla ??

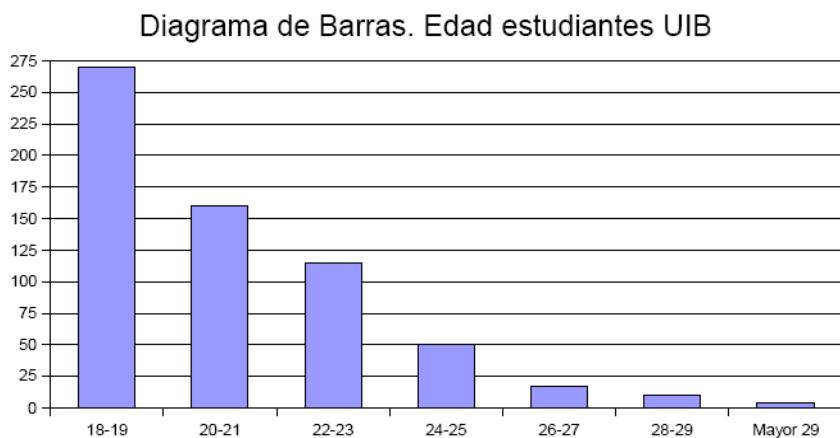


Figura 2.4: Diagrama de barras de frecuencias absolutas de la tabla ??

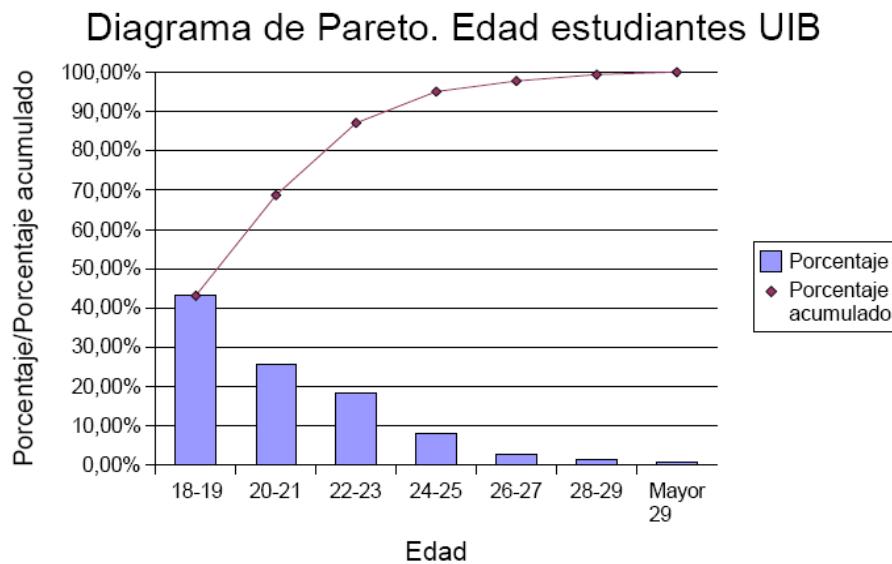


Figura 2.5: Diagrama de Pareto de porcentajes y porcentajes acumulados de la tabla ??

manera conjunta los datos de dos variables, para ello se emplean los diagramas de barras dobles. En el ejemplo 2 de la sección ?? (figura ??) se muestra un ejemplo de este tipo de diagramas.

- **Histogramas.** Un histograma es un gráfico que describe las frecuencias absolutas de una variable cuantitativa mediante barras contiguas de área proporcional a la frecuencia representada. Para el ejemplo de la edad de los estudiantes de la UIB (tabla ??) el histograma correspondiente se muestra en la figura ??.

El eje horizontal del histograma representa los distintos intervalos de valores considerados. En el ejemplo las edades están agrupadas en períodos de 4 años (salvo el último intervalo “mayores de 25”), los dos siguientes histogramas (figuras ?? y ??) muestran el resultado agrupando en períodos de 1 y 2 años, respectivamente (tablas ?? y ??).

El valor que se muestra en el interior de las barras de los histogramas de las figuras ?? y ?? es la frecuencia absoluta de cada intervalo y es igual a la anchura del intervalo multiplicada por la altura de la barra. Por ejemplo, para la segunda barra del histograma de la figura ??, el valor de frecuencia es 160 y la anchura del intervalo [20, 21] es 2, por este motivo la altura de la barra es 80 ($160 = 2 \times 80$).

La comparación de estos histogramas con los diagramas de barras mostrados en las figuras ??, ?? y ?? nos permiten observar las principales diferencias entre ambas representaciones:

1. En los histogramas, el área y no la altura de las barras es proporcional a la frecuencia representada. Esto significa que si dos intervalos tienen la misma frecuencia pero uno de ellos es mayor que el otro entonces su altura será inferior.
2. La barras no están separadas por un espacio en blanco en los histogramas, al contrario que en los diagramas de barras.
3. Todos los valores entre el mínimo y el máximo de la variable están representados en el histograma, pero no así en el diagrama de barras. Esto se aprecia comparando

Histograma. Edad estudiantes UIB

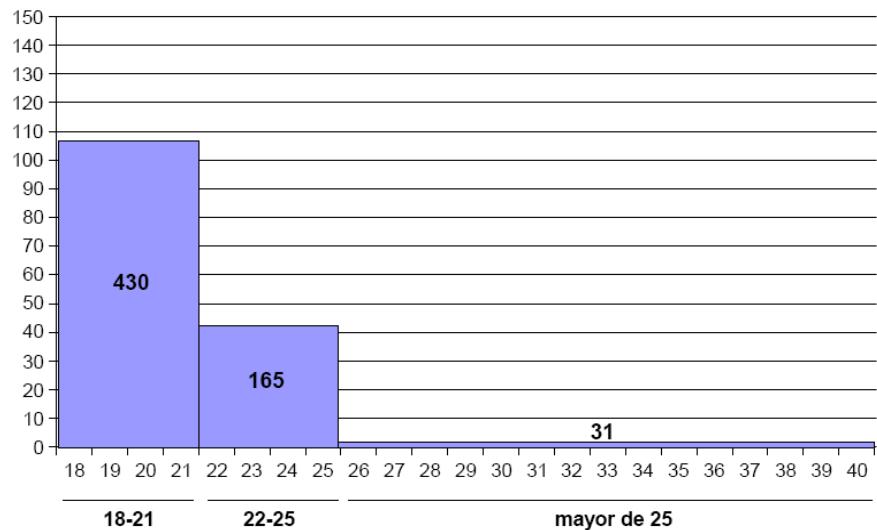


Figura 2.6: Histograma de frecuencias absolutas de la tabla ??

Histograma. Edad estudiantes UIB

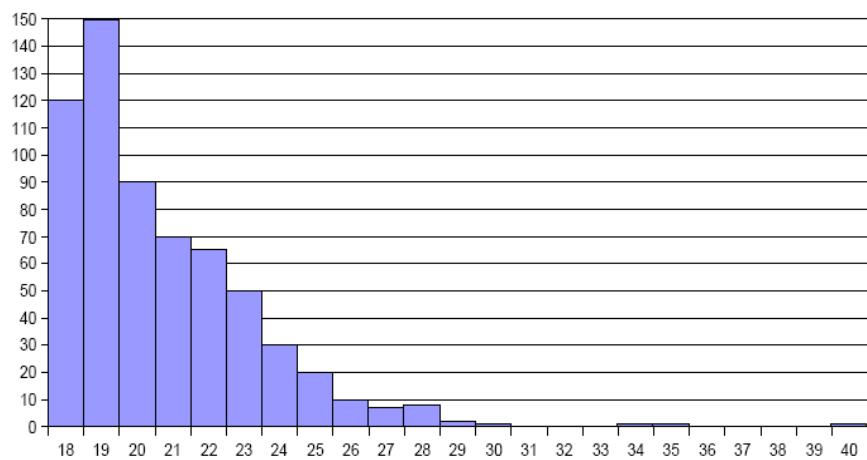


Figura 2.7: Histograma de frecuencias absolutas de la tabla ??

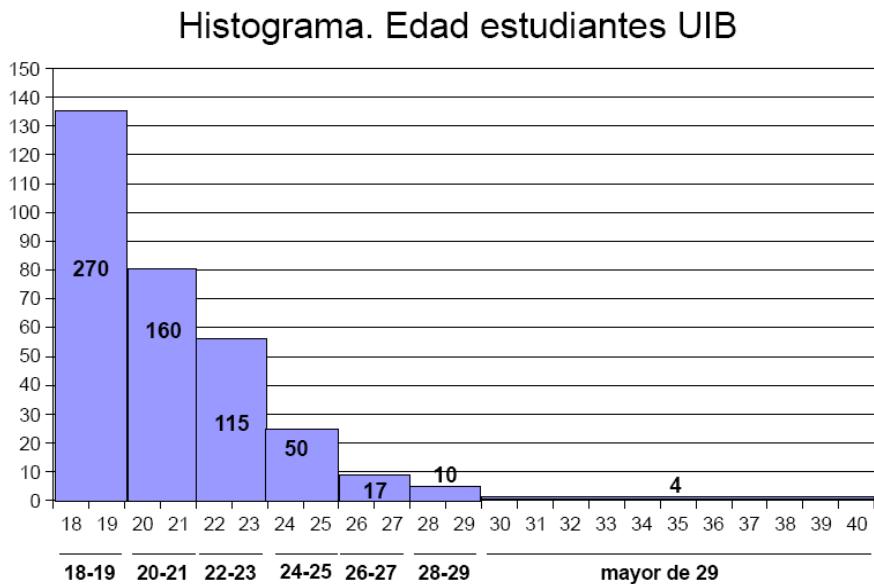


Figura 2.8: Histograma de frecuencias absolutas de la tabla ??

las figuras ?? y ?. En este caso los valores de edad 31, 32, 33, 36, 37, 38 y 39 se representan en el histograma (con una barra de altura cero, que no se dibuja), en cambio no aparecen en el diagrama de barras.

Los intervalos de valores representados en el histograma pueden ser de anchuras diferentes, aunque es habitual que todos sean iguales (salvo para los valores extremos que suelen representarse con intervalos de tipo “mayor de” o “menor de”).

- **Diagramas de tarta o pictogramas.** En un diagrama de tarta se representan las proporciones o porcentajes mediante sectores circulares de tamaño proporcional al valor representado. Se suelen utilizar para representar variables nominales. Para el ejemplo sobre la inmigración por nacionalidades obtendríamos el diagrama de tarta que se muestra en la figura ??.

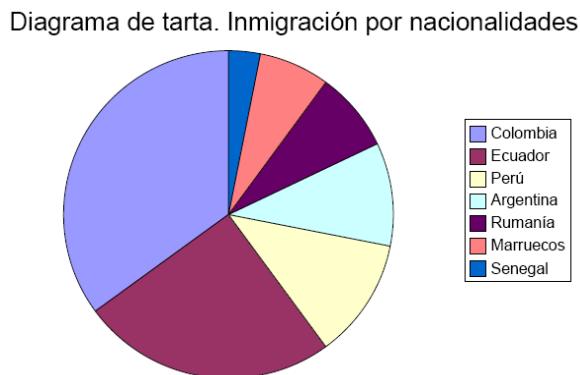


Figura 2.9: Diagrama de tarta de porcentajes de la tabla ??

- **Diagramas lineales.** En estos diagramas se unen mediante líneas una serie de puntos cuya coordenada horizontal representa el valor de la variable y la vertical la frecuencia o porcentaje asociados al valor. Se utilizan para la descripción de variables cuantitativas y son ideales para apreciar las tendencias de los datos. Además, usando líneas de distintos colores o puntos de distintas formas permiten la representación conjunta de datos de varias variables. Cuando se emplean para representar frecuencias (absolutas, relativas o acumuladas) se denominan **polígonos de frecuencia** y cuando en el eje horizontal se representan valores temporales (meses, años, etc.) se denominan **cronogramas**.

En la figura ?? se muestran los polígonos de frecuencias absolutas y acumuladas para el ejemplo sobre la edad de los estudiantes de la UIB de la tabla ???. En la figura ?? se muestra un ejemplo de cronograma.

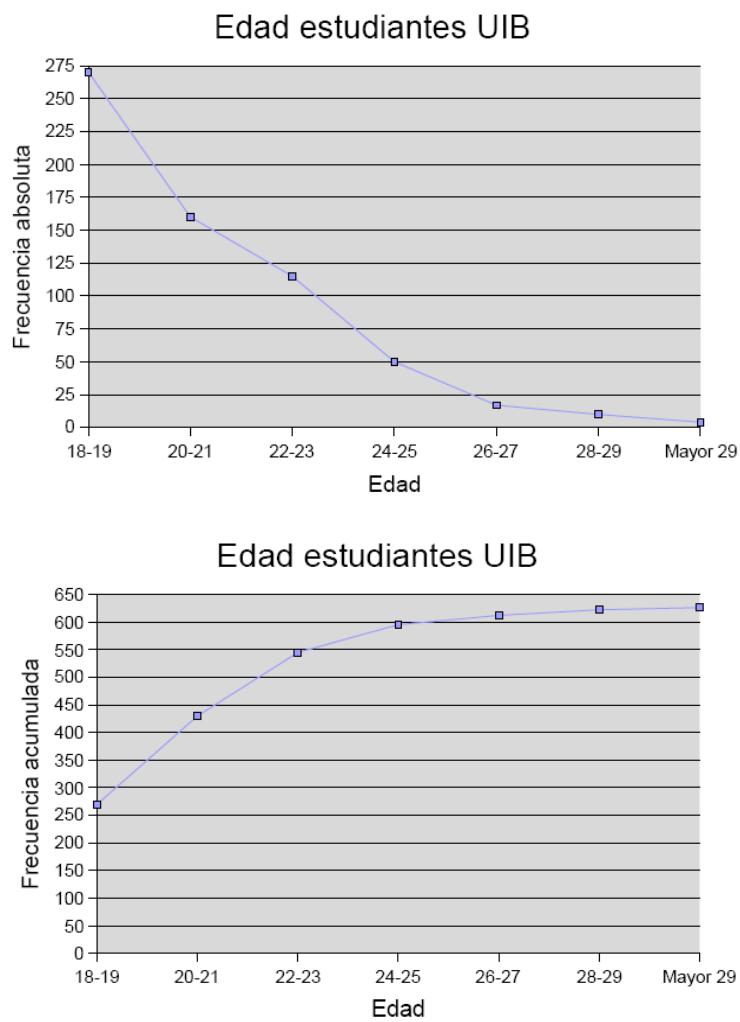


Figura 2.10: Ejemplo de la tabla ???. Arriba: polígono de frecuencias absolutas. Abajo: polígono de frecuencias acumuladas.

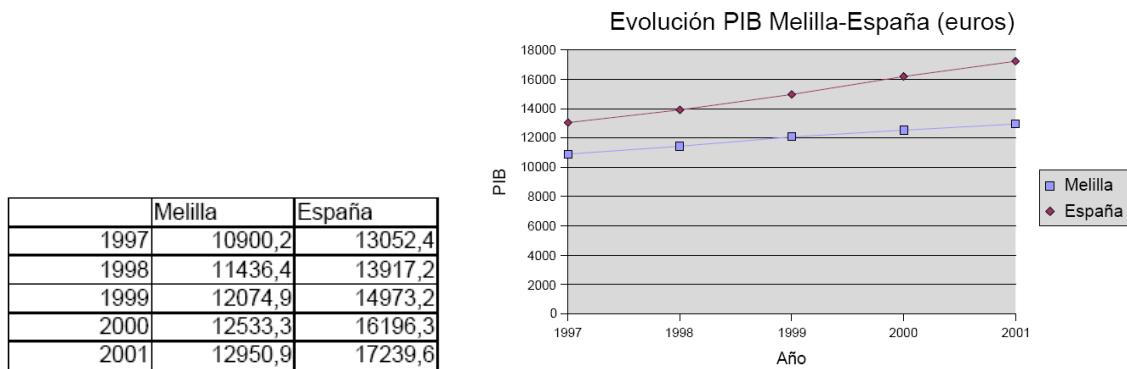


Figura 2.11: Izquierda: tabla de valores del PIB para Melilla y el conjunto de España entre 1997 y 2001 (fuente INE). Derecha: cronograma conjunto de los PIBs de Melilla y España

2.4. Tablas y gráficas estadísticas con ordenador

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 el uso de 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 2.2.

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 desde el menú de inicio de Windows:

Inicio: Todos los programas:OpenOffice.org:OpenOffice.org Calc

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

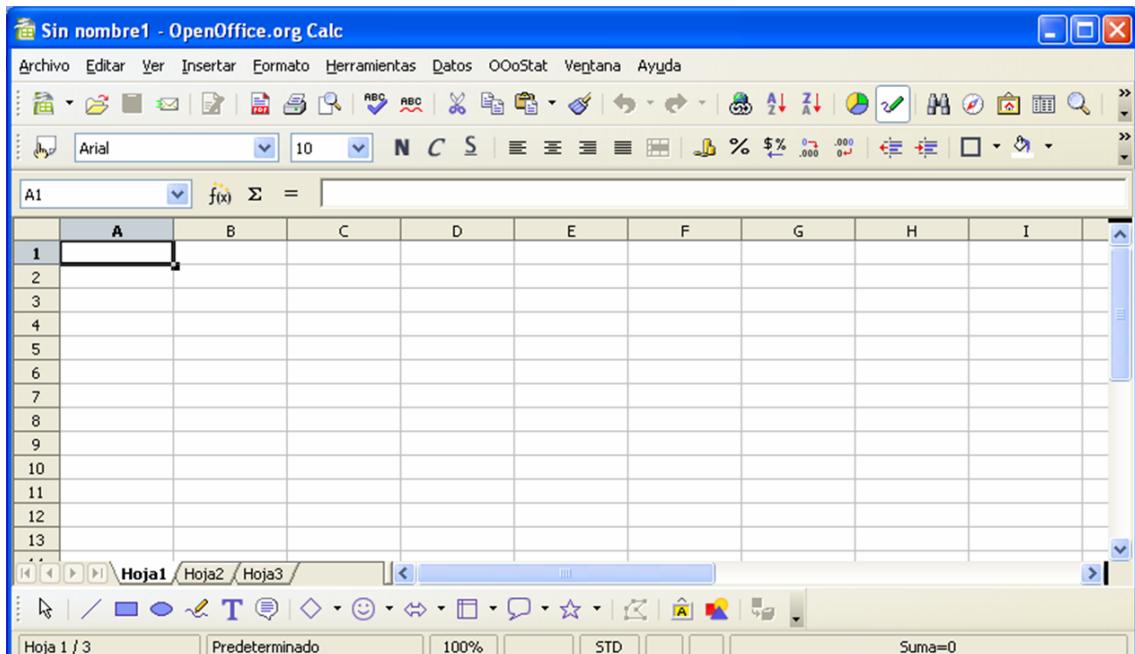


Figura 2.12: Ventana de inicio de OpenOffice Calc

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.

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

3. Los valores de la columna B (*nº 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

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

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

4. 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).
5. 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 ??), 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 ??).
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 ??).
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 ??).
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 ??.

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

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 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 *Bordes* y haremos clic sobre el ícono .

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

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 2.15: 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 **Diagrama...**.
- Se abrirá una nueva ventana titulada *Formatead o automático diagrama*. Aquí seleccionaremos las casillas de datos a representar. Para ello seleccionaremos las casillas *B3* a *B11* manteniendo el botón izquierdo del ratón apretado. A continuación, mientras mantenemos la tecla *Shift* pulsada, seleccionaremos del mismo modo las casillas *C3* a *C11*. Los datos de la primera columna seleccionada se representarán sobre el eje horizontal y los de la segunda sobre el eje vertical. Los valores de la ventana se muestran en la figura ???. Haciendo clic sobre el botón *Siguiente* pasamos a la siguiente ventana.
- A continuación debemos seleccionar el tipo de diagrama, en nuestro caso un diagrama de barras con texto, por lo que hacemos clic sobre el ícono , seleccionamos *Representación de texto en previsualización* y pulsamos *Avanzar*.

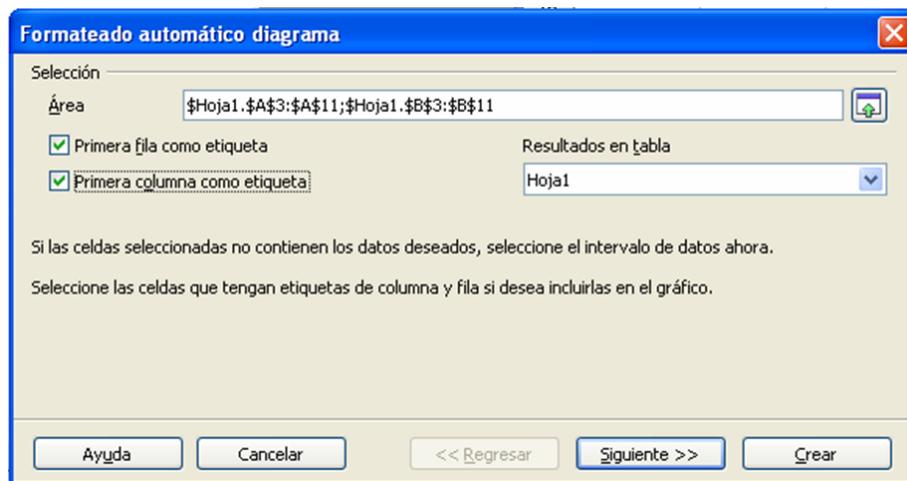


Figura 2.16: Inserción de un diagrama de barras. Ventana inicial del proceso de formateado automático

En la siguiente ventana se elige una variante del diagrama de barras, nosotros nos quedamos con la opción por defecto por lo que volvemos a pulsar *Avanzar*.

- d) 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 *Leyenda*. Los datos de la ventana se muestran en la figura ???. Finalmente pulsamos el botón *Crear*.

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 ??.

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.

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 de *Formato automático diagrama* debemos seleccionar las casillas *A3* hasta *A11* y *D3* hasta *D11*.
- b) Cuando seleccionamos el tipo de diagrama debemos hacer clic sobre el icono .
- c) Antes de pulsar el botón *Crear* debemos activar la opción *Leyenda* para que se muestre el significado de los colores del diagrama.
- d) 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



Figura 2.17: Inserción de un diagrama de barras. Ventana final del proceso de formateado automático

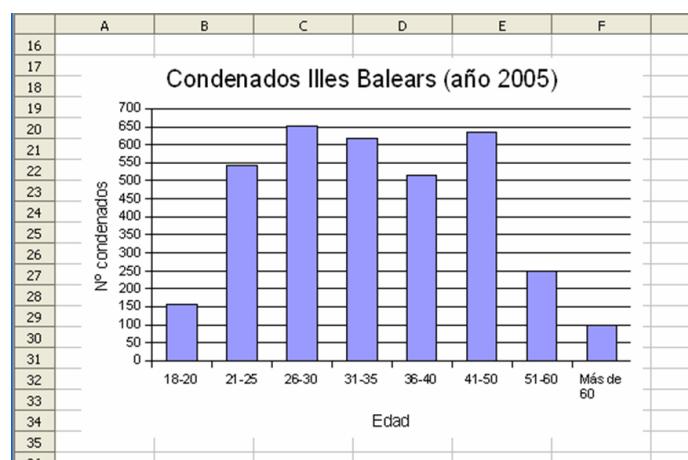


Figura 2.18: Diagrama de barras de frecuencias absolutas del ejemplo 1

resultado final se muestra en la figura ??.

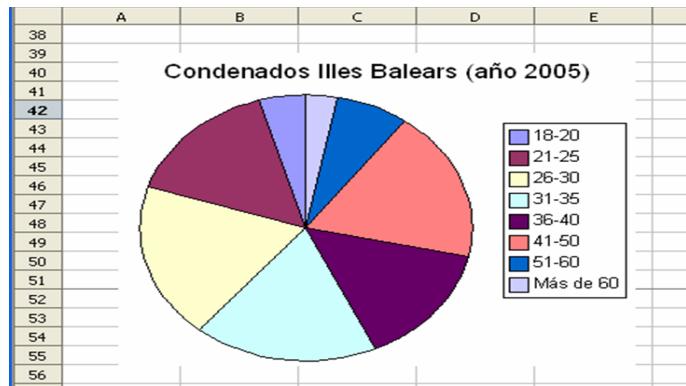


Figura 2.19: Diagrama de tarta de porcentajes del ejemplo 1

- e) Este diagrama puede ser imprimido o insertado en otro documento al igual que el diagrama de barras.
3. **Creación de un diagrama de Pareto.** Explicamos a continuación cómo crear un diagrama de Pareto de frecuencias relativas y frecuencias relativas acumuladas. Los pasos a seguir son muy similares a los de los diagramas anteriores, con las siguientes modificaciones:
- a) En la ventana inicial de *Formateado automático diagrama* debemos seleccionar las casillas *A3 hasta A11*, *C3 hasta C11* y *F3 hasta F11*.
 - b) El tipo de diagrama a seleccionar es el de barras pero la variante a escoger es la representada con el icono
 - c) En la última ventana para el formateado del diagrama activamos la opción *Leyenda* y dejamos sin título el eje Y. A continuación creamos el diagrama. Mediante el cursor podemos cambiar de tamaño y posición el gráfico creado y podemos aumentar la talla del texto de la leyenda tal como se ha explicado para el diagrama de tarta. El resultado final se muestra en la figura ??.

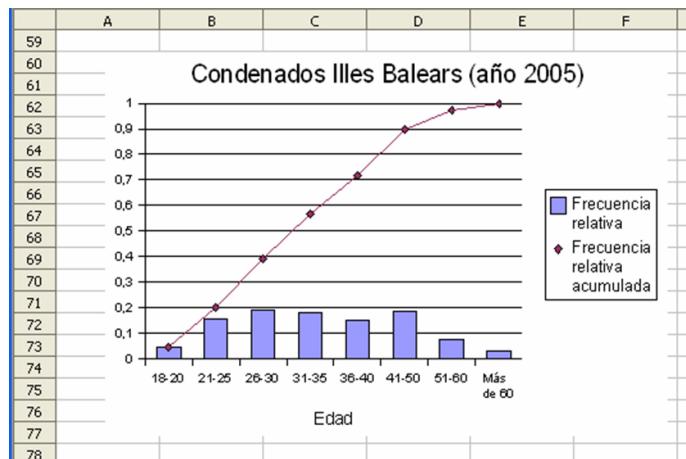


Figura 2.20: Diagrama de Pareto de frecuencias relativas del ejemplo 1

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.
2. Seguimos los pasos explicados para la creación de diagramas de barras en el ejemplo 1 pero seleccionando ahora las tres columnas de datos. 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 *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 recluyos”.

Al pulsar sobre el botón *Crear* obtenemos el resultado que se muestra en la figura ??.

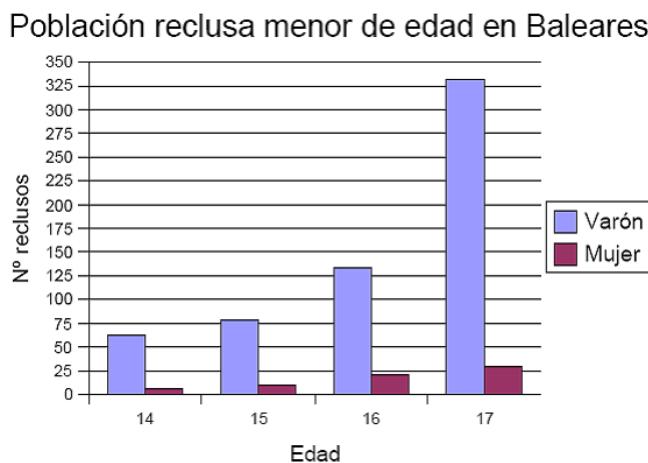


Figura 2.21: Diagrama de barras dobles del ejemplo 2

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*.
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 ??).

	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 2.22: 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 *A4* a *A14* y *C4* a *C14*. Los títulos del diagrama y del eje X son, respectivamente, “Histograma pesos” y “Pesos (Kg)” y la opción *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 ???. 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 ??.

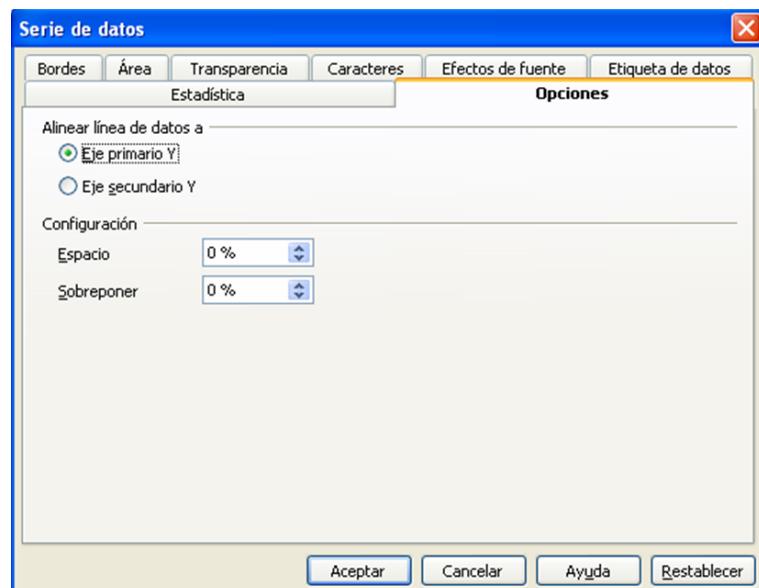


Figura 2.23: Ventana de diálogo para ajustar la anchura de las barras del diagrama de barras

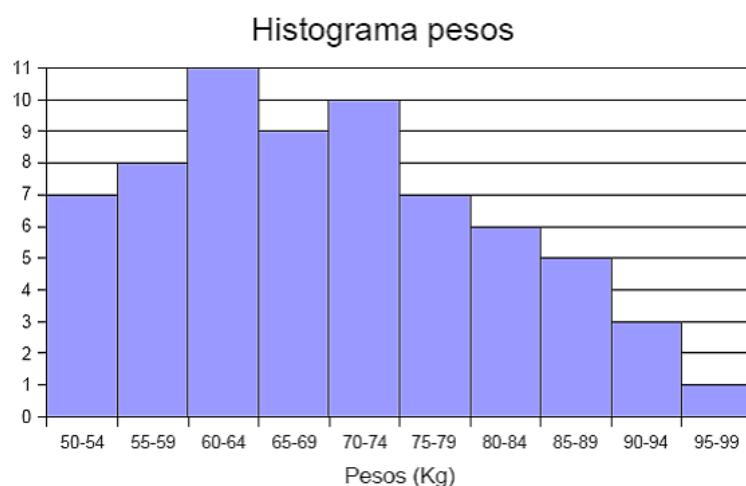


Figura 2.24: Histograma del ejemplo 3

Ejemplo 4

En este ejemplo mostramos como 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	Año	PIB
1992	612	2000	582
1993	513	2001	609
1994	515	2002	688
1995	597	2003	885
1996	622	2004	1045
1997	573	2005	1131
1998	601	2006	1231
1999	618	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*.
2. Creamos un diagrama siguiendo el procedimiento explicado en anteriores ejemplos. Las casillas de datos a seleccionar son de *A4* a *A19* y *B4* a *B19*.

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

No seleccionamos la opción de *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 ??.

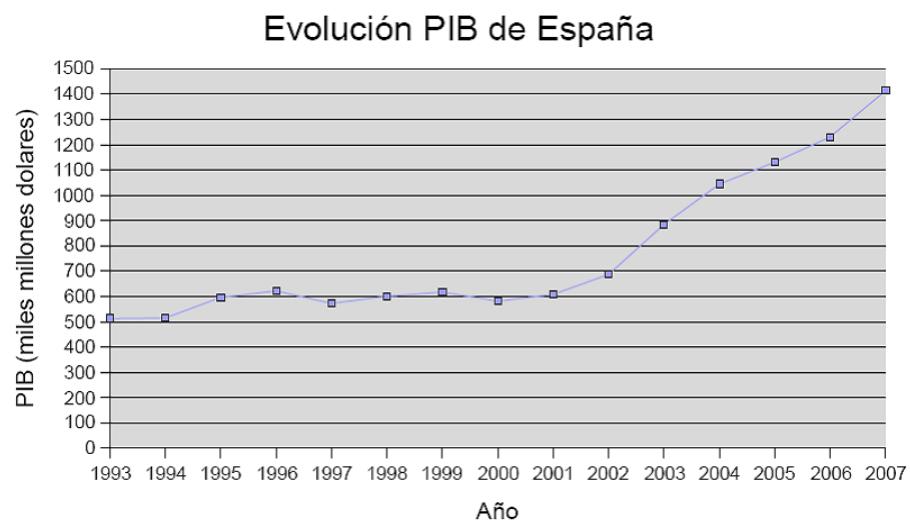
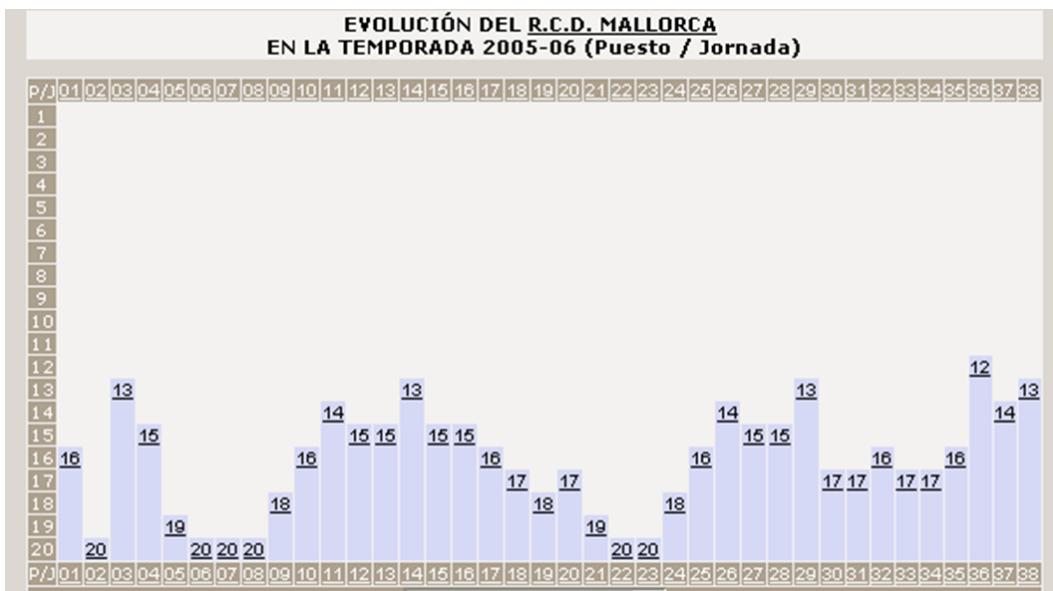


Figura 2.25: Diagrama lineal del ejemplo 4

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):



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

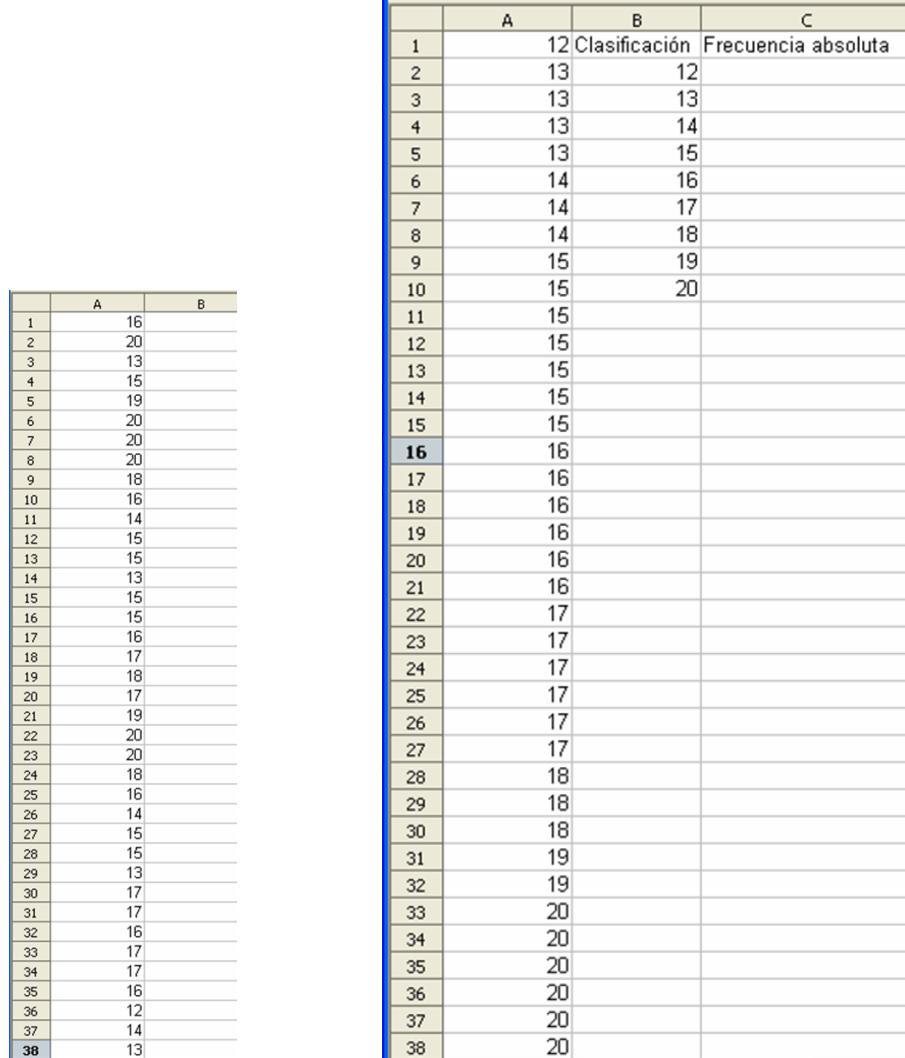
Para facilitar la tarea de escribir en orden creciente los distintos valores de la variable podemos **ordenar** los valores *brutos* del siguiente modo:

- 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 ??-derecha



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

Figura 2.26: Izquierda: 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)

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$38; "="&B2)$ ¹ 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 ??-izquierda.

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

Figura 2.27: Izquierda: frecuencias absolutas del ejemplo 5 (paso 3). Derecha: tabla final.

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

¹La 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

2.5. Utilización de la base de datos del INE

El Instituto Nacional de Estadística (INE) ofrece a través de su página web 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. 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 ??)

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

Figura 2.28: Opciones de la base de datos del INE

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 (ver Figura ??). Supongamos que deseamos conocer los datos sobre *Gasto público en educación*, para ello haremos clic sobre el icono  correspondiente a este concepto

Operaciones estadísticas relacionadas					
Elaboradas por el INE	Tablas INEbase	Contenido	Elaboradas por otros organismos	Tablas INEbase	Contenido
Estadística de enseñanza universitaria			Enseñanzas anteriores a la universidad		
Avance de la enseñanza universitaria			Becas y ayudas al estudio		
Pruebas de acceso a la universidad			Gasto público en educación		
Encuesta de transición educativo-formativa e inserción laboral			Tecnología de la información en la enseñanza no universitaria		
Encuesta de financiación y gastos de la enseñanza privada					
Módulo especial 2003 (EPA): Cursos de educación/formatación recibidos en los últimos doce meses					
Módulo especial 2000 (EPA): Transición de la educación al mercado laboral					

Figura 2.29: Datos sobre educación en la base de datos del INE

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 ??). 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*

Becas y ayudas. Curso 2004-2005	
Todas las enseñanzas	
	1.1 Becas, ayudas, becarios, beneficiarios e importe de las mismas por CCAA de destino, entidad que las concede, número y tipo de enseñanza.
	1.2 Becas, ayudas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de beca o ayuda (1).
Enseñanzas obligatorias, educación infantil y educación especial	
	2.1 Ayudas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de ayuda.
	2.2 Ayudas e importe de las mismas por CCAA de destino, número, entidad que las concede y tipo de ayuda.
	2.3 Ayudas, beneficiarios e importe de las mismas concedidas por CCAA de destino, número, entidad que las concede y nivel educativo del beneficiario.
Enseñanzas postobligatorias no universitarias	
	3.1 Becas e importe de las mismas concedidas por administración educativa financiadora, número y tipo de beca.
	3.2 Becas e importe de las mismas por CCAA de destino, número, entidad que las concede y tipo de beca.
	3.3 Becas, becarios e importe de las mismas por CCAA de destino, número, entidad que las concede y nivel educativo del becario.
Enseñanzas universitarias	
	4.1 Becas e importe de las mismas por administración educativa financiadora, número y tipo de beca.
	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.
	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.

Figura 2.30: Datos sobre gasto público en educación en la base de datos del INE

8. la nueva página que se abre nos permite escoger los datos a mostrar y la manera de mostrarlos mediante una serie de menús (ver Figura ??). 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.

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 ??

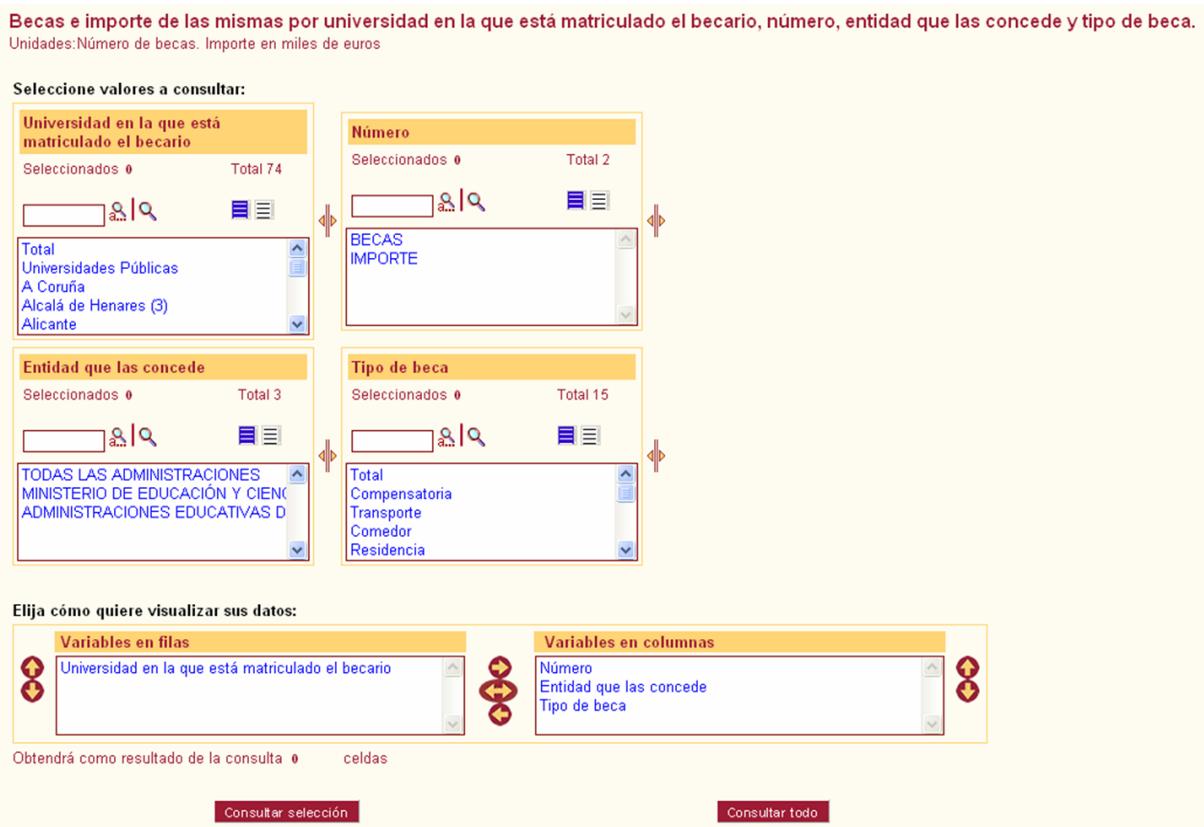


Figura 2.31: Menus para la selección de datos sobre número de becas e importe de las mismas en centros universitarios

9. 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

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

leerse desde la aplicación Excel de Microsoft y también desde la aplicación Calc de OpenOffice, que es la que utilizamos en nuestros ejemplos.

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.

2.6. Ejercicios propuestos

Ejercicio 1

A partir de los datos de la siguiente tabla calcular las frecuencias relativas y porcentajes.

- ¿Es posible calcular frecuencias y porcentajes acumulados? Calcularlos en caso de respuesta afirmativa.
- Dibujar un diagrama de tarta que muestre los porcentajes.



Fuente: Instituto Nacional de Estadística

Ejercicio 2

A partir de los datos de la siguiente tabla calcular las frecuencias relativas y porcentajes:

- ¿Es posible calcular frecuencias y porcentajes acumulados? Calcularlos en caso de respuesta afirmativa.
- Dibujar un diagrama de barras que muestre las frecuencias absolutas.
- Agrupar los datos en periodos de 2 años y dibujar el histograma de porcentajes.



Fuente: Instituto Nacional de Estadística

Ejercicio 3

Acceder a la base de datos del INE y dibujar un diagrama de barras doble con los datos de la encuesta de migraciones del año 2003, en valor absoluto, para los grupos de edad “de 0 a 9” hasta “de 65 y más”. Representar en una barra los datos correspondientes a varones y en la otra los correspondientes a mujeres.

Ejercicio 4

Acceder a la base de datos del INE y dibujar un diagrama de Pareto de porcentajes con los datos de divorcios en España en el año 2005 clasificados según la duración del matrimonio para cualquier edad del esposo.

Ejercicio 5

Acceder al banco de series temporales del INE (Tempus) y representar mediante un diagrama de línea los datos de población de ambos sexos en Baleares desde 1996.

Capítulo 3

Medidas de tendencia central

3.1. Introducción

En el tema anterior hemos definido el concepto de frecuencia (absoluta o relativa), en relación a una colección de datos estadísticos.

El conjunto de valores de frecuencia asociados a una variable estadística se conoce como **distribución de frecuencias** de la variable.

Por ejemplo, en el tema anterior hemos utilizado las siguientes distribuciones de frecuencias:

Nacionalidad	Frecuencia absoluta
Colombia	350
Ecuador	250
Perú	120
Argentina	100
Rumanía	80
Marruecos	70
Senegal	30

Inmigración por nacionalidades

Edad	Frecuencia absoluta
18	120
19	150
20	90
21	70
22	65
23	50
24	30
25	20
26	10
27	7
28	8
29	2
30	1
34	1
35	1
40	1

Edad estudiantes UIB

En un caso la variable es *Nacionalidad* (variable cualitativa) y la distribución de frecuencias muestra el número de inmigrantes de cada nacionalidad. En el otro caso la variable es *Edad* (variable cuantitativa) y la distribución de frecuencias representa el número de estudiantes de la UIB (de una muestra de 1000) según su edad.

En el primer caso la distribución de frecuencias consta de 7 valores. Estos valores describen los datos brutos relativos a 1000 personas (tabla ?? del tema anterior). Por otra parte, en el ejemplo sobre los estudiantes de la UIB, los 16 valores de la distribución de frecuencias resumen

626 datos brutos. Resulta evidente que leer una tabla formada por 7 valores resulta mucho más sencillo que leer una formada por 1000 valores (del mismo modo es más fácil leer 16 valores que 626), pero ¿es posible describir estas distribuciones de una manera aún más resumida, con dos o tres valores?. La respuesta a esta pregunta la damos en el presente tema.

3.2. Moda, mediana y media

En este tema veremos cómo describir las distribuciones de frecuencia de una variable mediante unos pocos parámetros (llamados **estadísticos**) que resumen su comportamiento (o *tendencia*) global. Estos estadísticos reciben el nombre de estadísticos de tendencia central. En el siguiente tema estudiaremos otro tipo de estadísticos (llamados de *dispersión*) que completan la descripción de la distribución de frecuencia.

Definimos tres estadísticos de tendencia central:

- La **moda** es el valor (o valores) con la máxima frecuencia. La moda se define tanto para variables cualitativas como ordinales o cuantitativas.

Para el ejemplo sobre la inmigración la moda es igual a *Colombia*, pues tiene una frecuencia de 350, mayor que la del resto de valores de la variable *Nacionalidad*.

Para el ejemplo sobre los estudiantes de la UIB la moda es igual a 19 pues su frecuencia es 150, mayor que el resto de frecuencias.

Si el valor máximo de la distribución se produce para dos valores distintos de la variable decimos que la distribución es de tipo **bimodal**. Si se produce para un único valor (como en los ejemplos comentados) se dice que la distribución es **unimodal**.

Un ejemplo de distribución bimodal es el siguiente:

Apellidos más frecuentes en Illes Balears (fuente INE)

<i>Primer apellido</i>	Frecuencia relativa (por 1000 habitantes)
GARCIA	15,0
PONS	15,0
TORRES	11,9
MARTINEZ	11,0
FERRER	9,5
FERNANDEZ	9,2
LOPEZ	9,0
SANCHEZ	8,9
SERRA	8,7
RIERA	8,2

Observamos como los apellidos “García” y “Pons” presentan el mismo máximo de frecuencia, por lo que la moda de la distribución es doble.

- La **mediana** es aquel valor que, cuando consideramos todos los valores de la muestra ordenados, ocupa el lugar central. Es decir, la mitad de las frecuencias están por encima de la mediana y la otra mitad por debajo.

En la práctica la mediana se calcula utilizando la siguiente propiedad: la mediana es el menor valor de la variable cuya *frecuencia acumulada* es mayor o igual a la mitad de la suma de todas las frecuencias.

Recordemos que la frecuencia acumulada se definió en el tema anterior para variables que admiten una ordenación de sus valores (variables ordinales o cuantitativas). En este caso, la frecuencia acumulada del valor que ocupa la posición i es igual a la suma de su frecuencia y las de los valores anteriores: $N_i = n_1 + n_2 + \dots + n_i$.

Por lo tanto la mediana es el primer valor que ocupa la posición i tal que $N_i \geq 0.5 \cdot N$, o lo que es lo mismo $N_i \geq 50\% \cdot N$, donde N es la suma de todas las frecuencias. La mediana sólo puede definirse para variables ordinales o cuantitativas, ya que para las variables cualitativas no tiene sentido calcular frecuencias acumuladas.

Por ejemplo, en el caso de la tabla de edades de los estudiantes de la UIB, tenemos las siguientes frecuencias acumuladas:

Edad	Frecuencia absoluta	Frecuencia acumulada
18	120	120
19	150	270
20	90	360
21	70	430
22	65	495
23	50	545
24	30	575
25	20	595
26	10	605
27	7	612
28	8	620
29	2	622
30	1	623
34	1	624
35	1	625
40	1	626

La suma de todas las frecuencias es 626 por lo que $50\% \cdot 626 = 313$. Vemos que a partir del 3^{er} valor las frecuencias acumuladas son mayores o iguales a 313, por lo que la mediana es el valor que ocupa la tercera posición en la tabla: mediana=20. Observemos que por encima de la mediana la suma de las frecuencias (incluyendo la frecuencia de la mediana) es $120 + 150 + 90 = 360$, mientras que por debajo es $90 + 70 + 65 + 50 + 30 + 20 + 10 + 7 + 8 + 2 + 1 + 1 + 1 = 356$, es decir, aproximadamente el mismo valor.

Un concepto relacionado con el de mediana es el de **percentil**. Si el 50 % de las frecuencias están por encima de la mediana, el percentil P se define de manera que el $P\%$ de las frecuencias están por encima suyo. El percentil P es el primer valor que ocupa la posición i tal que $N_i \geq P\% \cdot N$.

Por ejemplo, el percentil 30 de la tabla anterior es 19, que ocupa la posición 2 de la tabla, ya que $30\% \cdot 626 = 187,8$, $N_1 = 120 < 187,8$, $N_2 = 270 > 187,8$.

El percentil 80 es 23, que ocupa la posición 6 de la tabla, ya que $80\% \cdot 626 = 500,8$, $N_5 = 495 < 500,8$, $N_6 = 545 > 500,8$.

Los percentiles 25, 50 y 75 se denominan, respectivamente, primer, segundo y tercer **cuartiles**. Por ejemplo, el tercer cuartil de la tabla anterior es 22, que ocupa la posición 5 de la tabla, ya que $75\% \cdot 626 = 469.5$, $N_4 = 430 < 469.5$, $N_5 = 495 > 469.5$.

- La **media** (o media aritmética, denotada \bar{x}) es el valor medio de los valores de la variable, calculado en función de las frecuencias de la distribución según la siguiente fórmula:

$$\bar{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \cdots + n_k \cdot x_k}{N}$$

donde x_1, x_2, \dots, x_k son los distintos valores de la variable, n_1, n_2, \dots, n_k sus frecuencias, k el número de valores distintos que toma la variable y $N = n_1 + n_2 + \cdots + n_k$.

La media sólo puede calcularse para variables cuantitativas. Para el ejemplo sobre los estudiantes de la UIB tenemos que:

$$\bar{x} = \frac{120 \cdot 18 + 150 \cdot 19 + 90 \cdot 20 + 70 \cdot 21 + \cdots + 1 \cdot 34 + 1 \cdot 35 + 1 \cdot 40}{120 + 150 + 90 + 70 + \cdots + 1 + 1 + 1} = 20,69$$

3.3. Cálculo de moda, mediana y media para variables descritas por intervalos

Consideremos la siguiente distribución de frecuencias para el ejemplo de la edad de los estudiantes de la UIB.

Edad	Frecuencia absoluta
18-19	270
20-21	160
22-23	115
24-25	50
26-27	17
28-29	10
30-40	4

En este caso las edades se han agrupado en distintos intervalos. ¿Cómo podemos calcular la moda, mediana y media de este caso?

El cálculo es un poco diferente al de la sección anterior:

- **Moda.** Hallamos primero el intervalo con el valor máximo de frecuencia. En nuestro caso es el 18 – 19. La moda se puede definir como el punto medio de este intervalo (aunque otras definiciones son también posibles): $\frac{18 + 19}{2} = 18,5$.

Observamos que el resultado es similar, aunque no el mismo, al obtenido cuando no se agrupaban las edades en intervalos (el valor de la moda era 19). En general, al agrupar los datos en intervalos se pierde precisión en el cálculo de los estadísticos.

- **Mediana.** Calculamos primero las frecuencias acumuladas para cada intervalo. El resultado para nuestro ejemplo es:

Edad	Frecuencia absoluta	Frecuencia acumulada
18-19	270	270
20-21	160	430
22-23	115	545
24-25	50	595
26-27	17	612
28-29	10	622
30-40	4	626

Hallamos a continuación el primer intervalo para el que la frecuencia acumulada es mayor o igual a la mitad de la suma de todas las frecuencias. Si este intervalo no es ni el primero ni el último de la distribución podemos utilizar la siguiente fórmula para afinar el cálculo de la mediana:

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

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 ejemplo: $L_i = 20$, $L_{i+1} = 21$, $n_i = 160$, $N_{i-1} = 270$ y $N = 626$, por tanto

$$\text{mediana} = 20 + \frac{50\% \cdot 626 - 270}{160} \cdot (21 - 20) = 20 + \frac{313 - 270}{160} \cdot 1 = 20.27$$

De nuevo observamos que este valor es similar, aunque diferente, al obtenido al hacer el cálculo sin agrupar en intervalos (en ese caso la mediana valía 20).

Siguiendo un procedimiento muy similar al del cálculo de la mediana podemos calcular los percentiles para datos agrupados en intervalos. El **percentil P** se halla del siguiente modo:

1. calculamos las frecuencias acumuladas para cada intervalo
2. hallamos el primer intervalo para el que la frecuencia acumulada es mayor o igual al $P\%$ de la suma de todas las frecuencias
3. si el intervalo hallado no es ni el primero ni el último de la distribución utilizamos la siguiente fórmula para afinar el cálculo del percentil P :

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

Por ejemplo, el percentil 75 % (*tercer cuartil*) del ejemplo anterior es 22,34, ya que:

- el percentil 75 % se encuentra para el intervalo 22–23, ya que $75\% \cdot 626 = 0,75 \cdot 626 = 469,5$ y $N_{22-23} = 545 > 469,5$
- tenemos que $L_i = 22$, $L_{i+1} = 23$, $n_i = 115$, $N_{i-1} = 430$ y $N = 626$, por tanto el tercer cuartil es

$$22 + \frac{75\% \cdot 626 - 430}{115} \cdot (23 - 22) = 22 + \frac{469,5 - 430}{115} \cdot 1 = 22,34$$

- **Media.** Calculamos en primer lugar los puntos medios de los intervalos. Si los límites de un intervalo son L_i y L_{i+1} , el punto medio es $m_i = \frac{L_i + L_{i+1}}{2}$. A estos puntos se les denomina **marcas de clase**.

Para nuestro ejemplo:

Edad (intervalo)	Edad (punto medio)	Frecuencia absoluta
18-19	18,5	270
20-21	20,5	160
22-23	22,5	115
24-25	24,5	50
26-27	26,5	17
28-29	28,5	10
30-40	35	4

La media se calcula con una fórmula muy parecida a la de la sección anterior:

$$\bar{x} = \frac{n_1 \cdot m_1 + n_2 \cdot m_2 + \cdots + n_k \cdot m_k}{N}$$

donde k es el número total de intervalos, m_i el punto medio del intervalo i , n_i su frecuencia y N la suma de todas las frecuencias.

En nuestro ejemplo:

$$\bar{x} = \frac{270 \cdot 18,5 + 160 \cdot 20,5 + 115 \cdot 22,5 + \cdots + 10 \cdot 28,5 + 4 \cdot 35}{626} = 20,7$$

Nuevamente comprobamos como este valor es similar al obtenido cuando los valores no están agrupados en intervalos (la media era 20,69), aunque no es exactamente el mismo. En general, cuanto más anchos sean los intervalos mayor es la pérdida de precisión en el cálculo de los estadísticos.

3.4. Cálculo de mediana y media a partir de datos brutos

En las secciones anteriores hemos explicado como calcular la mediana y la media a partir de datos dados en forma de tablas de frecuencias. Sin embargo, el cálculo de estos estadísticos puede hacerse directamente y de manera muy sencilla a partir de los datos brutos.

Consideremos por ejemplo los datos de la siguiente tabla, en la que se muestran las notas de varios alumnos de la asignatura de Economía:

	Nota
Alumno 1	7
Alumno 2	6,5
Alumno 3	8
Alumno 4	4,5
Alumno 5	9
Alumno 6	3,5
Alumno 7	8
Alumno 8	7
Alumno 9	4,5
Alumno 10	5

Se trata de datos brutos puesto que no están organizados en una tabla de frecuencias. En este caso la mediana y la media pueden calcularse fácilmente de la siguiente forma:

- **Mediana.** Ordenamos los datos de menor a mayor. La mediana es el valor que se encuentra en la mitad de la tabla. En nuestro ejemplo:

	Nota
Alumno 6	3,5
Alumno 4	4,5
Alumno 9	4,5
Alumno 10	5
Alumno 2	6,5
Alumno 1	7
Alumno 8	7
Alumno 3	8
Alumno 7	8
Alumno 5	9

Como hay 10 alumnos, la mitad de la tabla ordenada corresponde a la posición $\frac{10}{2} = 5$, por lo que la mediana es el valor 6,5.

Si el número de valores fuera impar se tomaría como posición media de la tabla el primer valor entero superior a $N/2$ (donde N es el número total de valores). Para el ejemplo sobre las notas, si hubiera 11 alumnos la mitad sería $\frac{11}{2} = 5,5$ por lo que la mediana se buscaría en la posición 6.

- **Media.** Se calcula como la suma de todos los valores dividida por el número total de valores:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Para nuestro ejemplo:

$$\bar{x} = \frac{7 + 6,5 + 8 + 4,5 + 9 + 3,5 + 8 + 7 + 4,5 + 5}{10} = 6,3$$

3.5. Cálculo de moda, mediana y media según el tipo de variable

Para variables cualitativas sólo es posible el cálculo de la moda. Las variables ordinales permiten calcular moda y mediana, mientras que los tres estadísticos se pueden calcular para variables cuantitativas.

Por tanto, para variables no cualitativas es posible calcular varios estadísticos, ¿cuál de ellos describe mejor la distribución de frecuencias? Para responder a esta pregunta debemos tener en cuenta las siguientes consideraciones:

- la moda es un descriptor poco global ya que sólo da información sobre el valor más frecuente. Por ejemplo, si nos dicen que la moda de una distribución es 20 sólo sabemos que este es el valor más frecuente, pero no si hay muchos o pocos valores por encima o por debajo de 20.

- la mediana proporciona más información que la moda pues nos dice cuál es el valor central de la distribución. Por ejemplo, si nos dicen que la mediana de una distribución es 15, sabemos que aproximadamente la mitad de los valores son superiores a 15 y la otra mitad son inferiores a 15.
- la media es la medida más usada pues su cálculo implica el uso de información de toda la distribución de frecuencias. Si nos dicen que la media es 10 sabemos que éste es el promedio de los valores de la distribución.

Sin embargo, la media es muy sensible a valores extremos de la distribución. Por ejemplo, la media de la secuencia de valores 10, 12, 14, 16, 18 es 14 (suponemos en este caso que todos los valores tienen frecuencia igual a 1). Sin embargo, la media de 10, 12, 14, 16, 18, 80 es 25, a pesar de que ambos conjuntos de números son muy parecidos. La mediana en ambos casos es, no obstante, 14. Es por esto que la mediana es preferible a la media en los casos en que existen valores de la variable muy diferentes al resto de valores.

3.6. Moda, media y mediana y simetría de las distribuciones de frecuencias

Consideremos las distribuciones de frecuencias representadas en las gráficas de la figura ??, para las que se indican los valores de moda, mediana y media

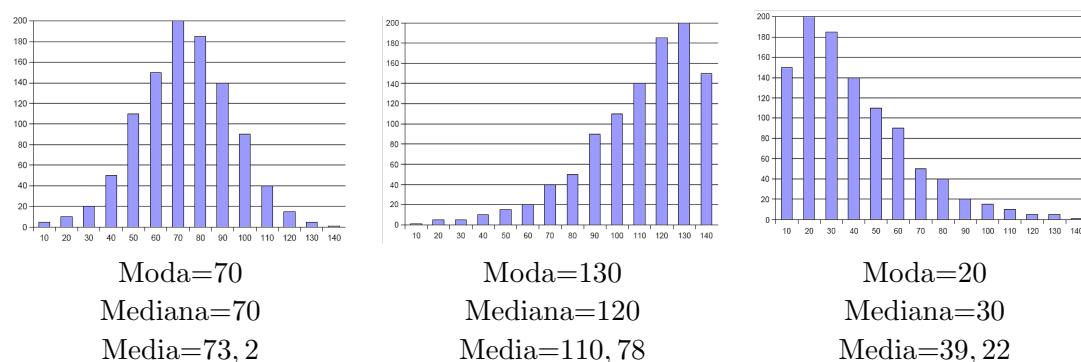


Figura 3.1: Relación entre la simetría y la moda, mediana y media de una distribución unimodal

En los tres casos el valor de la moda es único (distribuciones unimodales). Podemos observar como en la primera gráfica los valores de los tres estadísticos son muy similares y que la gráfica es simétrica respecto al valor central.

La segunda es asimétrica y se cumple que $\text{media} < \text{mediana} < \text{moda}$. Este tipo de asimetría recibe el nombre de asimetría negativa o por la izquierda.

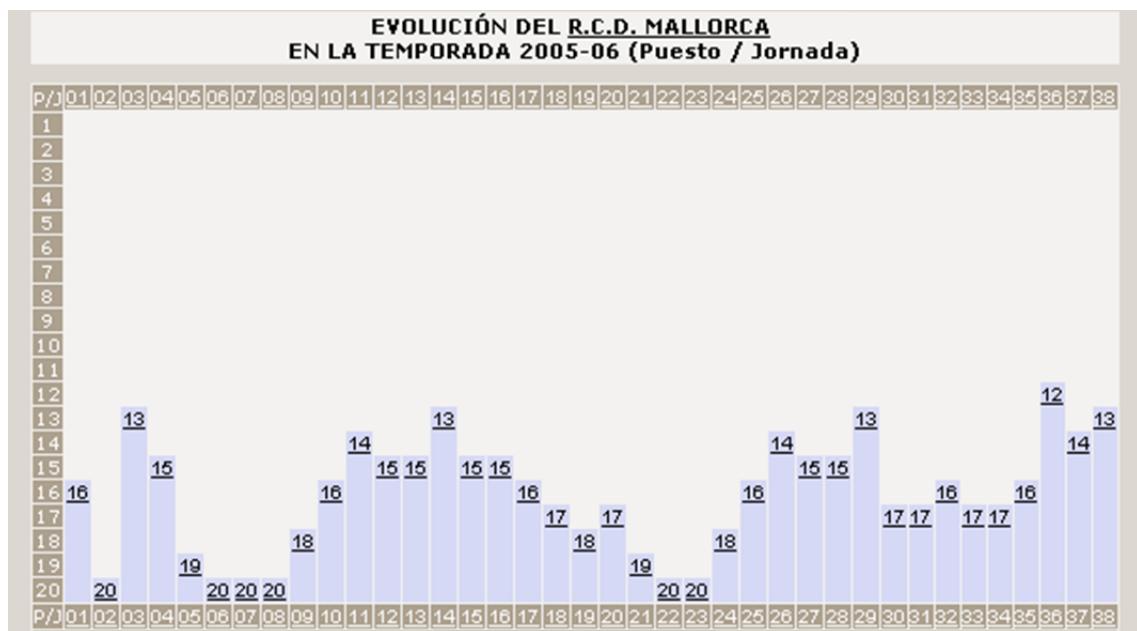
Finalmente, la tercera gráfica también es asimétrica, cumpliéndose que $\text{moda} < \text{mediana} < \text{media}$. Esta asimetría se conoce como asimetría positiva o por la derecha.

3.7. Cálculo de moda, mediana, media y percentiles con ordenador

Las operaciones matemáticas que deben realizarse para calcular los estadísticos explicados en este tema 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 de una distribución. 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 ??-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^{er} cuartil.** Nos situamos en la casilla *D18* y escribimos `=Cuartil(A1:A38;1)`. Al pulsar *Enter* obtenemos el valor del 1^{er} cuartil.
7. **3^{er} cuartil.** Nos situamos en la casilla *D19* y escribimos `=Cuartil(A1:A38;3)`. Al pulsar *Enter* obtenemos el valor del 3^{er} cuartil.

El resultado obtenido es:

Moda	15
Mediana	16
Media	16,34
1 ^{er} cuartil	15
3 ^{er} 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 una 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 acumuladas, tal como se ha explicado en el tema anterior. La tabla que obtenemos es como la de la figura ??.
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).

	A	B	C	
1	Clasificación	Frecuencia absoluta	Frecuencia acumulada	
2	12	1	1	
3	13	4	5	
4	14	3	8	
5	15	7	15	
6	16	6	21	
7	17	6	27	
8	18	3	30	
9	19	2	32	
10	20	6	38	
11				

Figura 3.2: OpenOffice Calc con los del ejemplo 2 y las frecuencias acumuladas calculadas

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 *Enter*.
- 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 ???. 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 3.3: Ilustración del cálculo de la moda en el ejemplo 2

3. **Mediana.** Para facilitar la localización del valor de la mediana podemos seguir el siguiente procedimiento:

- Suponemos que los datos de frecuencia acumulada se encuentran en las casillas $C2$ a $C10$, tal como se muestran en la figura ??.
 - Nos colocamos en la casilla $D2$, escribimos $=SI(C2$>=$(C$10)*0,5;1;0)$ y pulsamos *Enter*. Esta fórmula escribe un “1” en la casilla si la frecuencia acumulada en $C2$ es mayor o igual que la mitad de frecuencia acumulada total ($C10$); en caso contrario escribe “0”.
 - Extendemos el cálculo al resto de casillas de la columna D del siguiente modo: situamos el cursor en la esquina inferior derecha de la casilla $D2$ y, manteniendo el botón izquierdo del ratón pulsado, arrastramos el cursor hasta la casilla $D10$. Al soltar el botón se mostrarán los valores calculados.
- El resultado se muestra en la figura ??.

	A	B	C	D	E	F
1	Clasificación	Frecuencia absoluta	Frecuencia acumulada			
2		12	1	0	0	0
3		13	4	0	0	0
4		14	3	0	0	0
5		15	7	15	1	0
6		16	6	21	1	0
7		17	6	27	1	0
8		18	3	30	1	1
9		19	2	32	1	1
10		20	6	38	1	1
11						

Figura 3.4: Ilustración del cálculo de la mediana y el primer y tercer cuartiles en el ejemplo 2

- La mediana es el primer valor de la variable para el cual tenemos un “1” en la columna D . En nuestro caso: mediana=16 (casilla $A6$), ya que el primer “1” de la columna D está en la casilla $D6$.
- 1^{er} **cuartil.** El cálculo es idéntico al de la mediana con la sola diferencia de que la fórmula que debemos escribir ahora es $=SI(C2$>=$(C$10)*0,25;1;0)$. Podemos hacer los cálculos en una nueva columna (por ejemplo columna E), los resultados se muestran en la figura ???. El primer cuartil vale 15.
 - 3^{er} **cuartil.** Igual que en el caso anterior. La fórmula ahora es $=SI(C2$>=$(C$10)*0,75;1;0)$. los resultados se muestran en la figura ?? (columna F). El tercer cuartil vale 18.
 - Media.** La media en este caso se calcula de manera muy sencilla del siguiente modo:
 - Situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla $F2$.
 - Escribimos la fórmula $=SUMA.PRODUCTO(A2:A10;B2:B10)/SUMA(B2:B10)$ ¹ y pulsamos *Enter*. El valor de la media se escribe en la casilla $F2$. En este caso, media=16,34.

¹Esta fórmula calcula la media empleando la fórmula dada en la sección 3.2: 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 $B10$

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
Matrícula 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 las frecuencias acumuladas. 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 ??.

	A	B	C	D
1	Qualificació	Frec. Absoluta	Frec. acumulada	
2	Suspens	17	17	0
3	Aprovat	16	33	1
4	Notable	12	45	1
5	Excel·lent	1	46	1
6	Matrícula H.	3	49	1
7				

Figura 3.5: Izquierda: frecuencias absolutas y acumuladas para el ejemplo 3 y columna adicional para el cálculo de la mediana.

A continuación seguimos el procedimiento explicado en el ejemplo 2 para el cálculo de la mediana:

1. Nos colocamos en la casilla *D2*, escribimos =SI(C2\$>=\$C\$10)/2;1;0 y pulsamos *Enter*. Esta fórmula escribe un “1” en la casilla si la frecuencia acumulada en *C2* es mayor o igual que la mitad de frecuencia acumulada total (*C10*); en caso contrario escribe “0”.
2. Extendemos el cálculo al resto de casillas de la columna *D* tal como se ha explicado en el ejemplo anterior. El resultado se muestra en la figura ??.

3. La mediana es el primer valor de la variable para el cual tenemos un “1” en la columna D . En nuestro caso: mediana=*Aprovat*.

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 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, ¿como 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 la sección ??:

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: 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 ??).

Para calcular de modo más preciso la mediana utilizamos la fórmula de la sección ??:

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

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

	A	B	C	D	E	F
1	Edad	Frec. Absoluta	Frec. Acumulada			
2	18-20	155	155	0	0	0
3	21-25	543	698	0	0	0
4	26-30	653	1351	0	1	0
5	31-35	619	1970	1	1	0
6	36-40	515	2485	1	1	0
7	41-50	636	3121	1	1	1
8	51-60	248	3369	1	1	1
9	60 y más	100	3469	1	1	1
10						

Figura 3.6: Tabla de frecuencias absolutas y acumuladas para el ejemplo 3. Las tres últimas columnas de la tabla facilitan el cálculo de la mediana y los cuartiles primero y tercero.

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

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

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 (ver figura ??). A continuación aplicamos la fórmula explicada en la sección ??:

- **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:

- a) 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 ??.
- b) 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 click 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 ??-arriba).

- c) 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 ??-abajo.

	A	B	C	D	E
1	Edad		Frec. Absoluta	Frec. Acumulada	
2	18-20		155	155	0
3	21-25		543	698	0
4	26-30		653	1351	0
5	31-35		619	1970	1
6	36-40		515	2485	1
7	41-50		636	3121	1
8	51-60		248	3369	1
9	60 y más		100	3469	1
10					
11					

	A	B	C	D	E
1	Edad	Edad media	Frec. Absoluta	Frec. Acumulada	
2	18-20	19	155	155	0
3	21-25	23	543	698	0
4	26-30	28	653	1351	0
5	31-35	33	619	1970	1
6	36-40	38	515	2485	1
7	41-50	45,5	636	3121	1
8	51-60	55,5	248	3369	1
9	60 y más	65	100	3469	1
10					
11					

Figura 3.7: Tablas de frecuencias absolutas y acumuladas para el ejemplo 3. Arriba, inserción de una nueva columna. Abajo, datos insertados con los valores centrales de los intervalos

- d) El cálculo de la media se hace ahora de manera similar al ejemplo 2:
- 1) Situamos en cursor en una casilla vacía cualquiera, por ejemplo la casilla *A12*.
 - 2) Escribimos la fórmula $=\text{SUMA}.\text{PRODUCTO}(B2:B9;C2:B9)/\text{SUMA}(C2:C9)$ y pulsamos *Enter*. El valor de la media se escribe en la casilla *A12*. En este caso, $\text{media}=35,43$.

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 $\text{media}=35,36$; para *60 – 75*, $\text{media}=35,51$; para *60 – 80*, $\text{media}=35,58$, etc. Lo importante es no suponer valores absurdos (por ejemplo *60 – 150*, pues es muy poco probable que haya personas de 150 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 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 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 ??.
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.

	A	
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 3.8: Datos brutos del ejemplo 6

4. **Media.** Nos situamos en la casilla $D16$ y escribimos $=\text{Promedio}(A1:A12)$. 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 de la mediana.
7. **3^{er} cuartil.** Nos situamos en la casilla $D18$ y escribimos $\text{Cuartil}(A1:A12;3)$. Al pulsar *Enter* obtenemos el valor de la mediana.

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*.

3.8. Ejercicios propuestos

Ejercicio 1

A partir de los datos de la siguiente tabla calcular los estadísticos de tendencia central para la variable “Edad de víctimas de accidentes en 2006”. Calcular también el primer y el tercer cuartiles.

Edad (años)	<i>Nº</i> víctimas
0 a 4	343
5 a 14	1172
15 a 17	333
18 a 24	918
25 a 64	5026
65 y más	2947

Edad de víctimas de accidentes en 2006 (fuente DGT)

Ejercicio 2

Calcular los estadísticos de tendencia central y los cuartiles 1º y 3º asociados a la variable “turismos matriculados durante el año 2006” a partir de los datos de la tabla del ejemplo 6.

Ejercicio 3

Calcular los estadísticos de tendencia central asociados a la variable “tipo de alojamiento” a partir de los datos de la siguiente tabla (fuente INE):

Encuesta de ocupación en alojamientos de turismo rural 2006	
Datos nacionales por modalidades. Oferta	
Grado de ocupación por plazas en fin de semana por modalidad y meses.	
Unidades: %	
	Total
Hotel Rural	40,50
Apartamento Rural	30,19
Casa Rural	32,49
Albergue Rural	24,25

Ejercicio 4

Calcular los estadísticos de tendencia central y los cuartiles 1º y 3º de la variable “terminación del cupón de la ONCE en el período 30 de octubre a 27 de noviembre de 2007” a partir de los datos de la siguiente tabla (fuente ONCE). ¿Qué tipo de simetría presenta la distribución? Calcular los cuartiles primero y tercero.

Terminación	Frecuencia
0	0
1	4
2	0
3	2
4	2
5	3
6	3
7	6
8	3
9	1

Ejercicio 5

Calcular los estadísticos de tendencia central del grado de satisfacción de los clientes de un determinado establecimiento a partir de los datos siguientes:

Grado satisfacción	<i>Nº</i> clientes
Muy satisfecho	50
Bastante satisfecho	80
Satisfecho	100
Poco satisfecho	40
Nada satisfecho	10

Capítulo 4

Medidas de dispersión

4.1. Introducción

En el tema anterior hemos visto como los estadísticos de tendencia central resumen en unos pocos valores el comportamiento global de una distribución de frecuencias.

Sin embargo, estos estadísticos no bastan para describir de manera suficientemente precisa la distribución. Consideremos el ejemplo de la figura ??.

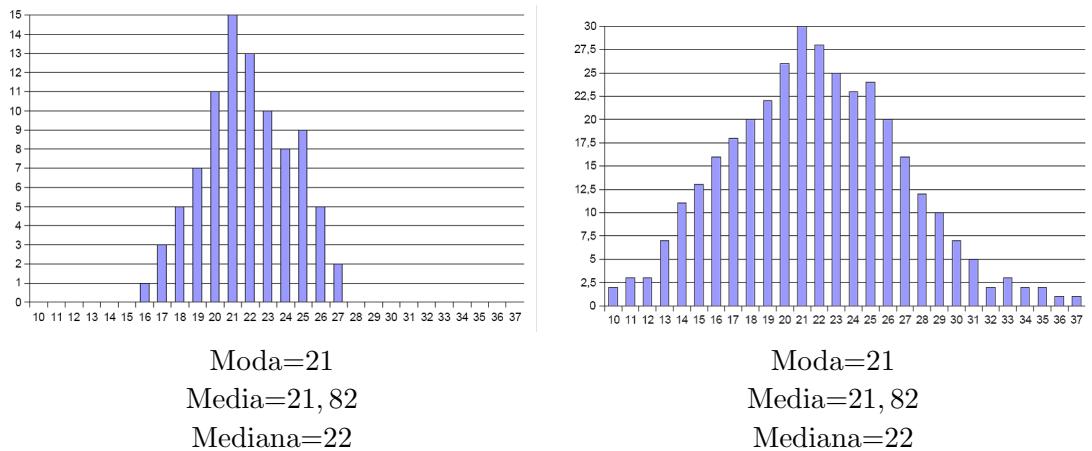


Figura 4.1: Dos distribuciones diferentes con los mismos estadísticos de tendencia central

Las dos distribuciones de la figura ?? tienen los mismos valores de moda, media y mediana. Sin embargo, las distribuciones son bastante diferentes entre sí . En la primera, los valores aparecen concentrados en torno al valor central; en cambio en la segunda los valores están más dispersos.

En este tema definiremos unos nuevos estadísticos cuyos valores dan cuenta de la heterogeneidad de los datos de la distribución, por este motivo se llaman **medidas de variabilidad o de dispersión**.

4.2. Medidas de dispersión

Las medidas de dispersión más habituales son la varianza y la desviación típica. Sin embargo hay muchas más: rango, rango intercuantílico, el *ratio* de variación, el coeficiente de variación,

etc. A continuación definimos cada uno de estos conceptos.

- **Ratio de variación.** Es una medida de dispersión basada en la moda. Puede utilizarse para cualquier tipo de variables (cualitativas, ordinales y cuantitativas) con distribuciones unimodales. Es la única medida de dispersión que puede calcularse para variables cualitativas. Se define como:

$$RV = 1 - \frac{n_{\text{moda}}}{N}$$

donde n_{moda} es la máxima frecuencia absoluta de la distribución (por tanto la frecuencia de la moda) y N es la frecuencia total (la suma de todas las frecuencias absolutas).

Por ejemplo, consideremos la distribución de la siguiente tabla:

<i>Nacionalidad</i>	Frecuencia absoluta
Colombia	350
Ecuador	250
Perú	120
Argentina	100
Rumanía	80
Marruecos	70
Senegal	30

La moda de la variable *Nacionalidad* es *Colombia*, y su frecuencia absoluta es 350. La suma de todas las frecuencias absolutas es 1000, de manera que el ratio de variación de la variable es

$$RV = 1 - \frac{350}{1000} = 0.65$$

Interpretación del ratio de variación: este valor mide el grado de concentración de los datos en torno a la moda. Valores cercanos a cero significan que casi todos los valores de la variable están concentrados en torno al valor de la moda; mientras que valores próximos a 1 implican que la moda tiene una frecuencia baja y que los valores están dispersos. En resumen:

$$\text{(concentración)} \quad 0 \leq VR < 1 \quad \text{(dispersión)}$$

- **Rango.** El rango puede definirse tanto para variables cuantitativas. Se define como la diferencia entre el valor máximo y el mínimo de la variable:

$$\text{Rango} = \text{máx} - \text{mín}$$

A continuación se muestran los rangos de dos distribuciones de datos diferentes, correspondientes a ejemplos del tema anterior:

<i>Terminación décimo ONCE</i>	Frecuencia
0	0
1	4
2	0
3	2
4	2
5	3
6	3
7	6
8	3
9	1

Min=1
Max=9
Rango=8

<i>Edad</i>	Frecuencia absoluta
18-19	270
20-21	160
22-23	115
24-25	50
26-27	17
28-29	10
30-40	4

Min=18
Max=40
Rango=22

Remarcar que para la primera tabla el valor mínimo en la muestra no es cero pues su frecuencia es cero. El valor mínimo es el primer valor cuya frecuencia es distinta de cero. En cuanto a la segunda tabla, hemos tomado como valores máximo y mínimo el límite inferior del primer intervalo y el superior del último. Otra opción igualmente válida hubiera sido tomar los valores medios de estos intervalos.

Interpretación del rango: en general podemos afirmar que para valores similares de media, moda o mediana, valores grandes del rango implican mayor dispersión que valores más pequeños. Sin embargo esto no siempre es cierto, como veremos a continuación.

- **Rango intercuantílico (RIC).** Para entender el significado de este estadístico consideremos primero los siguientes ejemplos:

Edad	Frecuencia
18	5
19	17
20	12
21	9
22	9
23	8
24	4

Moda=19
 Mediana=19
 Media=20,62
 Min=18
 Max=24
 Rango=6

Edad	Frecuencia
18	5
19	17
20	12
21	9
22	9
23	8
24	4
35	1

Moda=19
 Mediana=19
 Media=20,84
 Min=18
 Max=35
 Rango=17

Observamos como en la segunda tabla, un solo valor extremo (35, con frecuencia 1) provoca que el rango de ambas distribuciones sea muy distinto, cuando en realidad ambas contienen prácticamente los mismos valores.

Este ejemplo muestra que el rango es muy sensible a los valores extremos de la distribución. Una medida de dispersión más robusta es el rango intercuantílico, que se define como la diferencia entre el tercer cuartil y el primer cuartil:

$$\text{RIC} = 3^{\text{er}} \text{ cuartil} - 1^{\text{er}} \text{ cuartil}$$

Para las tablas anteriores tendríamos: $1^{\text{er}} \text{ cuartil}=19$ (para ambas tablas), $3^{\text{er}} \text{ cuartil}=22$ (para ambas tablas). De manera que en ambos casos tenemos $\text{RIC}=22 - 19 = 3$.

Interpretación del rango intercuartílico. La interpretación es la misma que la del rango: un valor grande implica mayor dispersión que uno pequeño. Sin embargo, el rango intercuantílico presenta la ventaja de ser robusto ante valores extremos poco frecuentes, por eso se utiliza más que el rango. Además, el 50% de los valores centrales de la muestra se encuentran entre los límites del RIC.

- **Varianza.** La varianza es una medida de dispersión que sólo puede utilizarse para variables cuantitativas. Su definición depende de si se aplica a datos de una población o una muestra:

$$\text{Var} = \frac{\sum_{i=1}^k n_i \cdot (x_i - \bar{x})^2}{N} = \frac{n_1 \cdot (x_1 - \bar{x})^2 + n_2 \cdot (x_2 - \bar{x})^2 + \dots + n_k \cdot (x_k - \bar{x})^2}{N} \quad (\text{varianza poblacional})$$

$$\text{Var} = \frac{\sum_{i=1}^k n_i \cdot (x_i - \bar{x})^2}{N-1} = \frac{n_1 \cdot (x_1 - \bar{x})^2 + n_2 \cdot (x_2 - \bar{x})^2 + \dots + n_k \cdot (x_k - \bar{x})^2}{N-1} \quad (\text{varianza muestral})$$

donde x_i representa los distintos valores de la variable, n_i sus frecuencias, k es el número de valores diferentes que toma la variable, $N = n_1 + n_2 + \dots + n_k$ y \bar{x} es la media de los valores.

Como ya vimos en el tema 1, el término *población* hace referencia al total de individuos sobre los que se evalúa una variable estadística, mientras que una *muestra* es una

pequeña porción de la población total. Cuando la varianza se calcula sobre el total de datos posibles de una variable empleamos la fórmula de la varianza poblacional. Cuando los datos disponibles representan una pequeña porción del total de datos posibles empleamos la fórmula de la varianza muestral. El por qué de esta distinción entre ambos tipos de varianza se explicará en el módulo III.

Tomemos por ejemplo los datos de la siguiente tabla, ya utilizada en el tema anterior (ver ejemplo 2, sección ??):

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

El objeto del estudio es analizar la clasificación de un equipo de fútbol (RCD Mallorca) durante la temporada 2005-2006. Los datos de la tabla muestran *todos* los valores de la clasificación durante ese año, por lo que podemos considerar que son datos poblacionales. De manera que utilizaremos la primera versión de la fórmula de la varianza:

$$\text{Var} = \frac{1 \cdot (12 - 16,34)^2 + 4 \cdot (13 - 16,34)^2 + \cdots + 6 \cdot (20 - 16,34)^2}{1 + 4 + \cdots + 6} = 5,23$$

donde se ha utilizado que $\bar{x} = 16,34$, que es el valor de la media obtenido en el tema anterior.

Como segundo ejemplo consideremos los datos de la tabla ??.

Estos datos corresponden a un estudio sobre la edad de los estudiantes de la UIB, pero en lugar de tener datos de **todos** los estudiantes de la UIB sólo disponemos de datos de 626 estudiantes. Debemos calcular por tanto la varianza muestral:

$$\text{Var} = \frac{120 \cdot (18 - 20,69)^2 + 150 \cdot (19 - 20,69)^2 + \cdots + 40 \cdot (40 - 20,69)^2}{(120 + 150 + \cdots + 1) - 1} = 7,03$$

donde se ha utilizado que $\bar{x} = 20,69$, que es el valor de la media obtenido en el tema anterior.

Cálculo de la varianza para variables descritas por intervalos

El procedimiento es muy similar al utilizado para la media en el tema anterior (sección ??). Calculamos en primer lugar los puntos medios de los intervalos. Si los límites de un intervalo son L_i y L_{i+1} , el punto medio es $m_i = \frac{L_i + L_{i+1}}{2}$.

Para el ejemplo de los estudiantes de la UIB:

Tabla 4.1: Edad estudiantes UIB

<i>Edad</i>	Frecuencia absoluta
18	120
19	150
20	90
21	70
22	65
23	50
24	30
25	20
26	10
27	7
28	8
29	2
30	1
34	1
35	1
40	1

<i>Edad</i> (intervalo)	<i>Edad</i> (punto medio)	Frecuencia absoluta
18-19	18,5	270
20-21	20,5	160
22-23	22,5	115
24-25	24,5	50
26-27	26,5	17
28-29	28,5	10
30-40	35	4

La varianza se calcula con una fórmula muy parecida a la de la sección anterior:

$$\text{Var} = \frac{\sum_{i=1}^k n_i \cdot (m_i - \bar{x})^2}{N} = \frac{n_1 \cdot (m_1 - \bar{x})^2 + n_2 \cdot (m_2 - \bar{x})^2 + \dots + n_k \cdot (m_k - \bar{x})^2}{N} \quad (\text{varianza poblacional})$$

$$\text{Var} = \frac{\sum_{i=1}^k n_i \cdot (m_i - \bar{x})^2}{N-1} = \frac{n_1 \cdot (m_1 - \bar{x})^2 + n_2 \cdot (m_2 - \bar{x})^2 + \dots + n_k \cdot (m_k - \bar{x})^2}{N-1} \quad (\text{varianza muestral})$$

donde k es el número total de intervalos, m_i el punto medio del intervalo i , n_i su frecuencia, N la suma de todas las frecuencias y \bar{x} la media.

En nuestro caso calculamos la varianza muestral, por las mismas razones que en el ejemplo anterior:

$$\text{Var} = \frac{270 \cdot (18,5 - 20,7)^2 + 160 \cdot (20,5 - 20,7)^2 + \dots + 4 \cdot (35 - 20,7)^2}{(270 + 160 + \dots + 4) - 1} = 7,05$$

donde se ha utilizado que $\bar{x} = 20,7$, que es el valor de la media obtenido en el tema anterior.

Cálculo de la varianza a partir de datos brutos Al igual que ocurría con el cálculo de la media, el cálculo de la varianza a partir de datos brutos puede hacerse de una manera muy sencilla utilizando las siguientes fórmulas:

$$\text{Var} = \frac{\sum_{i=1}^N x_i^2}{N} - \bar{x}^2 = \frac{x_1^2 + x_2^2 + \dots + x_N^2}{N} - \bar{x}^2 \quad (\text{varianza poblacional})$$

$$\text{Var} = \frac{\sum_{i=1}^N x_i^2}{N-1} - \frac{N}{N-1} \bar{x}^2 = \frac{x_1^2 + x_2^2 + \dots + x_N^2}{N-1} - \frac{N}{N-1} \bar{x}^2 \quad (\text{varianza muestral})$$

donde x_i representa los distintos valores de la variable, N es el número total de valores y \bar{x} su media.

Consideremos por ejemplo los datos de la siguiente tabla, en la que se muestran las notas de varios alumnos de la asignatura de Economía:

	Nota
Alumno 1	7
Alumno 2	6,5
Alumno 3	8
Alumno 4	4,5
Alumno 5	9
Alumno 6	3,5
Alumno 7	8
Alumno 8	7
Alumno 9	4,5
Alumno 10	5

Queremos analizar las notas de estos 10 alumnos pero no deseamos extrapolar nuestros resultados al análisis de un grupo mayor, por lo que calcularemos la varianza poblacional.

En el tema anterior se ha calculado la media de estos datos ($\bar{x} = 6,3$). La varianza se calcula de la siguiente forma:

$$\text{Var} = \frac{7^2 + 6,5^2 + 8^2 + 4,5^2 + 9^2 + 3,5^2 + 8^2 + 7^2 + 4,5^2 + 5^2}{10} - 6,3^2 = 3,01$$

Interpretación de la varianza: la varianza mide la dispersión media de los valores de la variable. Valores grandes de la varianza significan gran dispersión respecto al valor medio mientras que valores pequeños implican una dispersión menor.

- **Desviación típica.** También llamada **desviación estándar**, se calcula simplemente como la raíz cuadrada de la varianza (sea esta poblacional o muestral):

$$\text{desviación estándar} = s = \sqrt{\text{Var}}$$

Al igual que la varianza la desviación típica sólo puede calcularse para variables cuantitativas. Para los ejemplos de la sección anterior:

$$\text{Clasificación RCD Mallorca 2005-2006: } \sqrt{5,23} = 2,29$$

$$\text{Edad estudiantes UIB (sin agrupar en intervalos): } \sqrt{7,03} = 2,651$$

$$\text{Edad estudiantes UIB (agrupando en intervalos): } \sqrt{7,05} = 2,655$$

Interpretación de la desviación estándar: al igual que la varianza la desviación típica mide la dispersión media de los valores de la variable. Cuanto mayor sea el valor mayor será la dispersión respecto al valor medio.

Una propiedad útil que relaciona los valores de media y desviación típica con la manera en que se distribuyen los datos de la variable es la **desigualdad de Chebychev**. Según esta desigualdad, la proporción de valores de la variable en el intervalo $[\bar{x} - k \cdot s, \bar{x} + k \cdot s]$ es, como mínimo, $(1 - \frac{1}{k^2}) \times 100\%$, donde \bar{x} y s son, respectivamente, la media y la desviación típica de la variable.

Aplicando esta propiedad a distintos valores de k tenemos, por ejemplo:

- al menos el 75 % de los datos están entre $[\bar{x} - 2 \cdot s \text{ y } \bar{x} + 2 \cdot s]$
- al menos el 89 % de los datos están entre $[\bar{x} - 3 \cdot s \text{ y } \bar{x} + 3 \cdot s]$
- al menos el 93.75 % de los datos están entre $[\bar{x} - 4 \cdot s \text{ y } \bar{x} + 4 \cdot s]$

Por ejemplo, para los datos sobre la edad de los estudiantes de la UIB (sin agrupación en intervalos) teníamos $\bar{x} = 20,69$ y $s = 2,65$. De acuerdo con la propiedad anterior debería cumplirse que al menos el 75 % de los datos está entre $20,69 - 2 \cdot 2,65 = 15,39$ y $20,69 + 2 \cdot 2,65 = 25,99$. Los valores de la variable *Edad* en este intervalo son: 18, 19, 20, 21, 22, 23, 24 y 25. Y la suma de las frecuencias de estos valores es: $120 + 150 + 90 + 70 + 65 + 50 + 30 + 20 = 595$. Esta cantidad representa un $\frac{595}{626} \times 100\% = 95\%$ del total de datos, por lo que, en efecto, se cumple la propiedad.

- **Coeficiente de variación.** Es el cociente entre la desviación estándar y la media de una variable. Sólo se define para variables cuantitativas positivas.

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

Interpretación del coeficiente de variación. Este coeficiente nos dice cuántas medias caben en s . Normalmente las variables cuyo valor medio es mayor presentan una varianza también mayor. Por ello, no tiene mucho sentido comparar las varianzas (o desviaciones típicas) de dos variables con valores de media muy diferentes. En estos casos es mejor comparar sus coeficientes de variación. Un mayor valor del coeficiente implica una mayor dispersión de los datos.

Para los ejemplos de las secciones anteriores:

$$\text{Clasificación RCD Mallorca 2005-2006: } CV = \frac{2,29}{16,34} = 0,14$$

$$\text{Edad estudiantes UIB (sin agrupar en intervalos): } CV = \frac{2,65}{20,69} = 0,128$$

4.3. Tipificación de variables estadísticas (*z-scores*)

En ocasiones debemos comparar valores de una misma variable obtenidos a partir de distintos datos estadísticos. Un ejemplo típico es el siguiente:

Una asignatura se imparte en dos grupos distintos por diferentes profesores, cada profesor pone un examen diferente, más difícil en el grupo 1 que en el grupo 2. Al estudiante con mejor nota se le concede una matrícula gratuita para el curso siguiente. El mejor alumno del primer grupo tiene una nota de 8, mientras que el del segundo grupo de 9. ¿Es justo concederle la matrícula gratuita al alumno del segundo grupo sabiendo que su examen ha sido más fácil? Si conocemos la media y la desviación típica (o varianza) de las notas de ambos grupos es posible decidir qué alumno es realmente mejor con ayuda de la tipificación.

La tipificación *normaliza* o *estandariza* los valores de las variables obtenidos a partir de conjuntos de datos diferentes. El valor tipificado (también llamado *z-score*) se calcula del siguiente modo:

$$z = \frac{v - \bar{x}}{s}$$

donde \bar{x} y s son, respectivamente, la media y la varianza del conjunto de datos al que pertenece el valor v .

Los datos tipificados pueden ya compararse de manera directa. Para nuestro ejemplo, si las notas medias de ambos grupos son $\bar{x}_1 = 6$ y $\bar{x}_2 = 7$, y las desviaciones típicas $s_1 = 1,5$ y $s_2 = 2$, respectivamente, los *z-scores* de los mejores alumnos serán:

$$z_1 = \frac{8 - 6}{1,5} = 1,33 \quad z_2 = \frac{9 - 7}{2} = 1$$

como $z_1 > z_2$ probablemente el mejor alumno es el del primer grupo, aunque su nota sea inferior.

4.4. Diagramas de caja

Los diagramas de caja son representaciones gráficas que permiten visualizar de manera rápida la dispersión de los valores de una variable. En estos diagramas de muestran la mediana, el rango intercuartílico, los valores “atípicos” y los valores “extremos”.

La figura ?? muestra de forma esquemática un diagrama de caja y los valores que lo definen. Calcularemos por ejemplo el diagrama de caja correspondiente a la tabla ??.

El diagrama de caja presenta las siguientes características:

- El eje vertical representa los valores de la variable. En el eje horizontal se indica la frecuencia acumulada total y el diagrama es simétrico respecto a este valor.
- El rango intercuantílico se representa mediante un rectángulo. El borde inferior del rectángulo representa el primer cuartil y el superior el tercer cuartil. La altura del rectángulo es por tanto igual al rango intercuartílico (RIC). La anchura del rectángulo no tiene ningún significado.

Dentro del rectángulo una línea más gruesa indica el valor de la mediana. Al menos el 50 % de los valores de la variable se encuentran dentro de este rectángulo.

En nuestro ejemplo los límites del rectángulo son: 1^{er} cuartil=19 y 3^{er} cuartil=22 (ver figura ??).

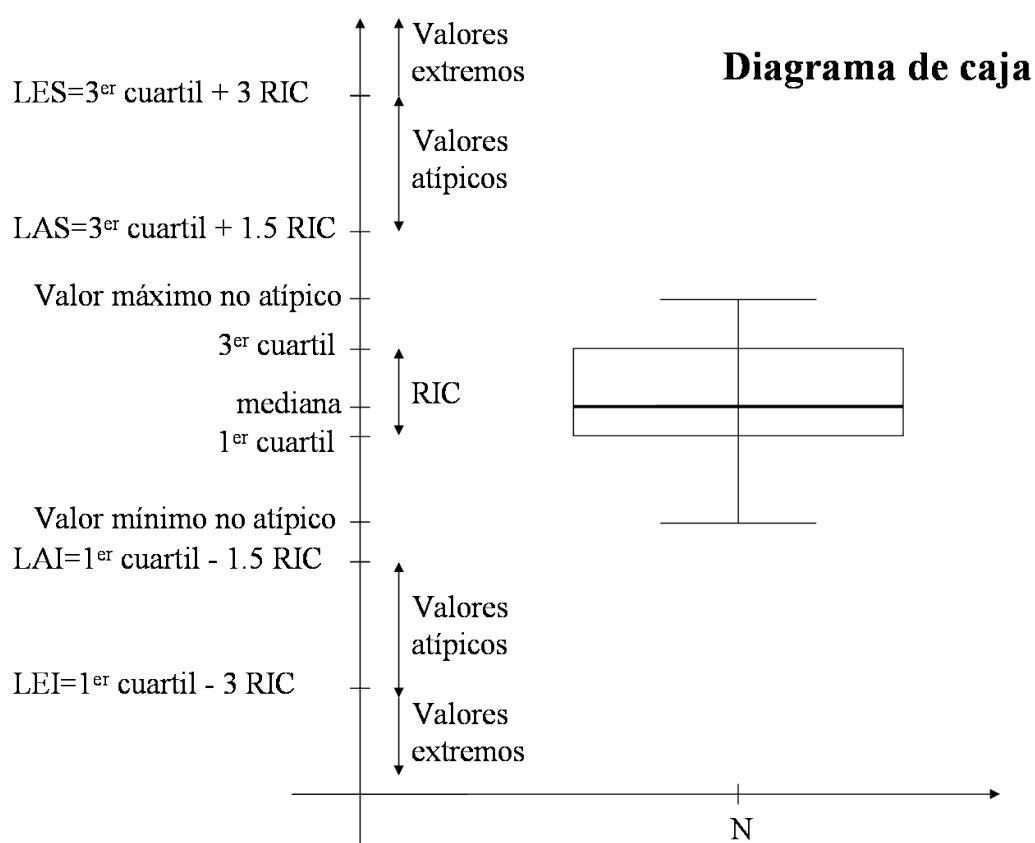


Figura 4.2: Representación esquemática de un diagrama de caja

- Para definir los valores “atípicos” y “extremos” se utilizan los siguientes límites (aunque en algunos autores utilizan otros):

$$\text{límite atípico inferior (LAI)} = 1^{\text{er}} \text{ cuartil} - 1,5 \times RIC$$

$$\text{límite atípico superior (LAS)} = 3^{\text{er}} \text{ cuartil} + 1,5 \times RIC$$

$$\text{límite extremo inferior (LEI)} = 1^{\text{er}} \text{ cuartil} - 3 \times RIC$$

$$\text{límite extremo superior (LES)} = 3^{\text{er}} \text{ cuartil} + 3 \times RIC$$

Se denominan **valores atípicos** aquellos valores que se encuentran entre LAI y LEI o bien entre LAS y LES. Son **valores extremos** aquellos inferiores a LEI o superiores a LES (ver figura ??).

En nuestro ejemplo $RIC = 3$, por lo que estos límites son:

$$LAI = 19 - 1,5 \cdot 3 = 14,5$$

$$LAS = 22 + 1,5 \cdot 3 = 26,5$$

$$LEI = 19 - 3 \cdot 3 = 10$$

$$LES = 22 + 3 \cdot 3 = 31$$

En el diagrama de caja se marcan con \circ los valores atípicos y con \bullet los extremos. En nuestro ejemplo (ver figura ??) son valores atípicos aquellos que están entre 10 y 14,5 (no incluido) y entre 26,5 (no incluido) y 31: 27, 28, 29, 30. Son valores extremos los inferiores a 10 y los superiores a 31: 34, 35, 40.

- Por último, en el diagrama se marcan con dos líneas horizontales los valores máximo y mínimo no atípicos (ver figura ??). En nuestro ejemplo el máximo valor no atípico de la variable es 26 (inferior a LAS); mientras que el mínimo valor no atípico es 18 (superior a LAI). Estos dos valores se conectan con el rectángulo mediante líneas verticales (ver figura ?? y ??).

En la figura ?? se muestra el diagrama de caja correspondiente a los datos de la tabla ??.

4.5. Cálculo de medidas de dispersión con el ordenador

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

Ejemplo 1

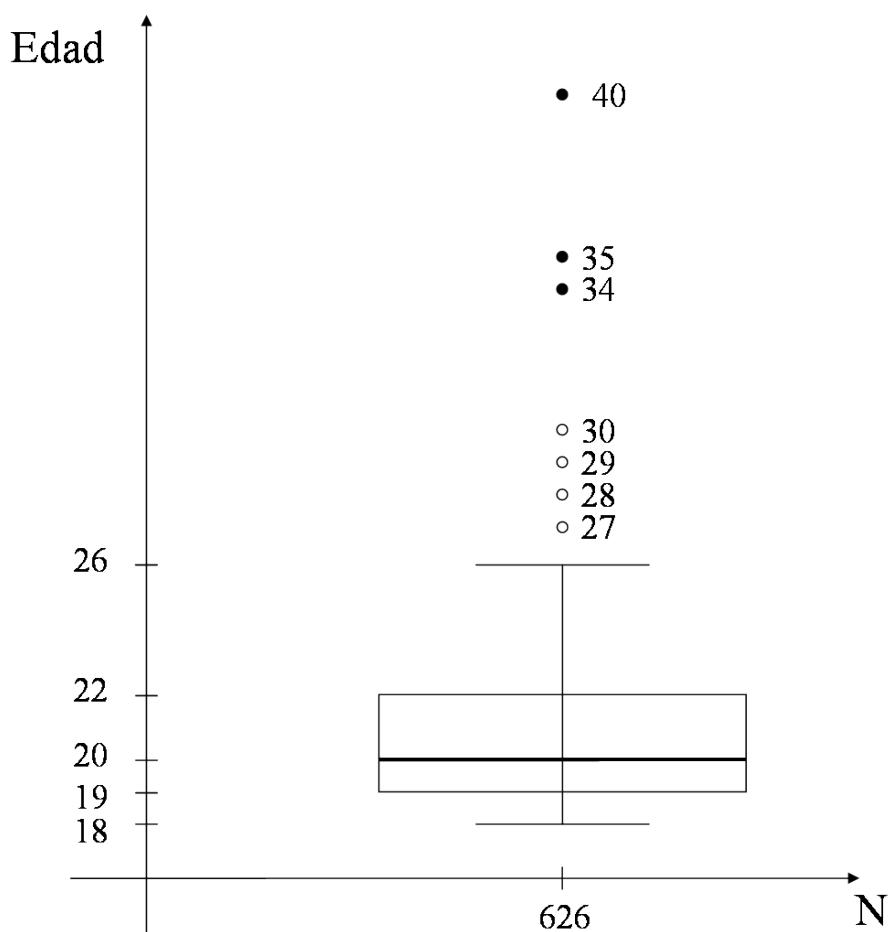


Figura 4.3: Diagrama de caja correspondiente a los datos de la tabla ??

Calcular el *ratio* de variación de 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	
Homicidio	0
Lesiones	31
Contra la libertad sexual	12
Hurto	69
Robo	306
Contra la salud pública	10

Fuente:Instituto Nacional de Estadística

El cálculo es sencillo y puede hacerse con la ayuda de una calculadora. El modo de la variable es *Robo*, cuya frecuencia es 306. La suma de todas las frecuencias es $N = 31+12+69+306+10 = 428$. De manera que:

$$RV = 1 - \frac{306}{428} = 0,28 = 28\%$$

Este valor indica que sólo el 28 % de los valores de la variable se encuentran fuera de la categoría *Robo*, o lo que es lo mismo, la mayoría de valores se concentran en la moda.

Ejemplo 2

Calcular el rango, rango intercuantílico, varianza, desviación estándar, valores atípicos y valores extremos para la variable “autobuses matriculados 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 ??

	A	
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 4.4: 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^{er} cuartil=264, 25 y 3^{er} cuartil =367. El rango intercuartílico es por tanto: $RIC = 367 - 264, 25 = 102, 75$.

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

1. Escribimos las palabras “Varianza”, “Desv. típica”, “Mínimo”, “Máximo” y “Rango” en las casillas C20 a C24, 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 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. **Mínimo.** Nos situamos en la casilla D22 y escribimos =Mín(A1:A12). Al pulsar *Enter* obtenemos el valor de mínimo de la variable.
5. **Máximo.** Nos situamos en la casilla D23 y escribimos =Máx(A1:A12). Al pulsar *Enter* obtenemos el valor de máximo de la variable.
6. **Rango.** Nos situamos en la casilla D24 y escribimos =D23-D22. Al pulsar *Enter* obtenemos el rango de la variable.

La siguiente tabla resume los resultados obtenidos hasta el momento:

Mediana	286
1 ^{er} cuartil	264, 25
3 ^{er} cuartil	367
RIC	102, 75
Varianza	14902, 24
Desviación típica	122, 07
Mínimo	158
Máximo	634
Rango	476

Finalmente calcularemos los valores atípicos y extremos. Para ello primero calcularemos los límites siguientes:

$$LAI = 264,25 - 1,5 \cdot 102,75 = 110,13$$

$$LAS = 367 + 1,5 \cdot 102,75 = 521,13$$

$$LEI = 264,25 - 3 \cdot 102,75 = -44$$

$$LES = 367 + 3 \cdot 102,75 = 675,25$$

El máximo valor no atípico es 427, el mínimo valor no atípico es 158 y el único valor atípico es 634. En este ejemplo no hay valores extremos.

Con todos estos datos estamos en disposición de dibujar el diagrama de caja de la variable. No obstante, Calc no ofrece la posibilidad de tal representación (tampoco es posible con Excel). Se requieren herramientas de software estadístico más complejas para ello (como R o SPSS).

Ejemplo 3

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*. Calcular el intervalo de peso $[\bar{x} - 2s, \bar{x} + 2s]$. ¿Cuál será el porcentaje mínimo de alumnos comprendidos en este intervalo?

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 $B15$ (n^o 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 $=SUMA.PRODUCTO(A2:A15;B2:B15)$. 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 $=SUMA.PRODUCTO(B2:B15;C2:C15)/(SUMA(B2:B15)-1)$. Al pulsar *Enter* obtenemos el valor de la varianza muestral en la casilla $A18$ (ver figura ??). 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		
...			

Figura 4.5: Hoja de cálculo del ejemplo 3.

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

4. La desviación típica se calcula como la raíz cuadrada de la varianza: nos colocamos por ejemplo en la casilla A19, escribimos la fórmula $=RAÍZ(A18)$ y pulsamos *Enter*. El resultado es 6,12.
5. El intervalo $[\bar{x} - 2s, \bar{x} + 2s]$ es igual a: $[73, 18 - 2 \cdot 6, 12, 73, 18 + 2 \cdot 6, 12] = [60, 94, 85, 42]$. Según la desigualdad de Chebichev en este intervalo debe haber un mínimo de $1 - \frac{1}{2^2} = 0,75 = 75\%$ valores de la variable. Podemos comprobar que, en efecto, esto es así, ya que la frecuencias de los valores en este intervalo suman: $10 + 20 + 25 + 15 + 35 + 44 + 50 + 37 + 22 + 15 + 10 = 283$. Esto representa un $\frac{283}{300} = 0,943 = 94,3\%$ de los valores de la variable.

Ejemplo 4

Calcular el rango 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 rango, la varianza y la desviación estándar.

Estadísticas judiciales 2005	
Estadística de lo Penal. Condenados. Resultados autonómicos	
Condenados según edad y sexo	
Unidades: n° 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 ^{er} cuartil	27,04
3 ^{er} cuartil	42,65

De aquí deducimos que el rango 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 rango 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 consideraremos 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 click 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 ??). 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 4.6: 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.

4.6. Ejercicios propuestos

Ejercicio 1

Calcular el rango, rango intercuantílico, varianza, desviación estándar, valores atípicos y valores extremos para el número de farmacias por municipios en Mallorca a partir de los datos de la tabla ??.

Tabla 4.2: Farmacias en Mallorca, por municipio (fuente: Col.legi Oficial d'Apotecaris de les Illes Balears, septiembre 2005)

Alaró	1	Capdepera	4	Llucmajor	10	Salines (ses)	2
Alcúdia	6	Consell	1	Manacor	14	Sant Joan	1
Algaida	1	Costitx	1	Mancor de la Vall	1	Sant Llorenç des Cardassar	6
Andratx	6	Deià	1	Maria de la Salut	1	Santa Maria del Camí	1
Ariany	1	Escrava	1	Marratxí	8	Santanyí	8
Artà	3	Esporles	1	Montuïri	1	Selva	2
Banyalbufar	1	Estellenchs	1	Muro	3	Sencelles	1
Binissalem	2	Felanitx	9	Palma	140	Sineu	2
Búger	1	Fornalutx	1	Petra	1	Sóller	5
Bunyola	2	Inca	8	Pollença	5	Son Servera	4
Calvià	27	Lloret de Vistalegre	1	Porreres	1	Santa Margalida	4
Campanet	1	Lloseta	2	Puigpunyent	1	Valldemossa	1
Campos	4	Llubí	1	Pobla (sa)	5	Vilafranca de Bonany	1

Ejercicio 2

A partir de los datos de la tabla ?? calcular la media, la varianza y la desviación típica de la variable “Edad de víctimas de accidentes en 2006”.

Tabla 4.3: Edad de víctimas de accidentes en 2006 (fuente DGT)

Edad (años)	Nº víctimas
0 a 4	343
5 a 14	1172
15 a 17	333
18 a 24	918
25 a 64	5026
65 a 80	2947

Ejercicio 3

Para incentivar a los trabajadores de una empresa de mensajeros la dirección de la empresa ha decidido conceder un suplemento salarial a la persona que hace las entregas con mayor rapidez. Los trabajadores de la empresa se organizan en dos turnos. En el turno de la mañana, debido al tráfico, el tiempo medio de entrega es de 30 minutos, con una varianza de 100, mientras que en el turno de la tarde la media es de 20 minutos con una varianza de 49. El mensajero más veloz del turno de mañana tarda una media de 25 minutos en hacer sus entregas y el de la tarde 15 minutos. Utiliza *z-scores* para decidir a qué mensajero aumentar el sueldo.

Módulo II

En este módulo avanzamos en el estudio de la Estadística Descriptiva con el estudio de las relaciones entre varias variables estadísticas. Al igual que en módulo anterior se explica cómo resolver los problemas con la ayuda de aplicaciones informáticas.

En la segunda parte del módulo se dan los fundamentos de Probabilidad necesarios para el estudio de la Estadística Inferencial.

Capítulo 5

Relaciones entre dos variables estadísticas

En los capítulos anteriores hemos aprendido a representar (mediante tablas de frecuencias y gráficas) y analizar (calculando estadísticos de tendencia central y de variabilidad) los datos correspondientes a una única variable estadística.

Sin embargo, es habitual que al hacer un estudio estadístico obtengamos datos de varias variables (por ejemplo, años de convivencia y número de denuncias por agresión en un estudio sobre violencia de género) y nos interesaría estudiar las relaciones de dependencia entre ellas. En este tema aprenderemos a crear tablas de frecuencias de dos variables, a representarlas gráficamente de manera conjunta y a calcular el grado de asociación que existe entre ellas.

El estudio conjunto de dos variables estadísticas recibe el nombre de **análisis bivariante**, en contraposición con el **análisis univariante**, referido a una única variable.

5.1. Tablas de contingencia

Las tablas mostradas en los temas anteriores mostraban valores de frecuencia referidos a una única variable. Para poner de manifiesto la relación entre dos variables, es conveniente mostrar valores de frecuencias referentes a ambas variables.

En este caso se consideran las frecuencias absolutas conjuntas n_{ij} , que representan el número de elementos de la muestra cuyo valor de la primera variable es x_i y de la segunda es y_j . Las tablas de frecuencias absolutas de dos variables se denominan **tablas de contingencia**.

La forma general de una tabla de contingencia es la siguiente:

$X \setminus Y$	y_1	y_2	\cdots	y_l	Suma
x_1	n_{11}	n_{12}	\cdots	n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	\cdots	n_{2l}	$n_{2\bullet}$
\vdots			\vdots		\vdots
x_k	n_{k1}	n_{k2}	\cdots	n_{kl}	$n_{k\bullet}$
Suma	$n_{\bullet 1}$	$n_{\bullet 2}$	\cdots	$n_{\bullet l}$	N

donde

- X e Y son los nombres de las variables

- k e l el número de valores diferentes que toman X e Y , respectivamente
- x_i e y_j representan los valores que toman las variables
- n_{ij} es el número de veces que aparecen de manera simultánea los valores x_i y y_j (frecuencia absoluta de x_i y y_j).
- $n_{i\bullet}$ y $n_{\bullet j}$ son el número de veces que aparecen los valores x_i y y_j , respectivamente (frecuencias absolutas parciales de x_i y y_j , respectivamente). Se verifica que $n_{i\bullet} = n_{i1} + n_{i2} + \dots + n_{il}$ y $n_{\bullet j} = n_{1j} + n_{2j} + \dots + n_{kj}$
- N es el número total de valores conjuntos: $N = n_{11} + n_{12} + \dots + n_{kl}$

Consideremos, por ejemplo, el caso de un estudio sobre seguridad vial que analiza la percepción de los conductores sobre la importancia del uso del cinturón en la prevención de accidentes. Para ello se entrevista a un conjunto de conductores y se consideran dos variables: “uso habitual del cinturón” y “victima de algún accidente”. Ambas variables pueden tomar dos únicos valores (sí o no).

Si los datos del estudio de nuestro ejemplo son: 300 personas entrevistadas; 80 han sufrido un accidente, de las cuales 60 usan habitualmente el cinturón de seguridad; de las 220 personas que no han sufrido accidente usan cinturón de seguridad 90. La tabla de contingencia sería la siguiente:

Tabla 5.1: Seguridad vial: uso cinturón vs. víctimas accidentes

Uso cinturón \ Víctima accidente	Si	No	Suma
Sí	60	90	150
No	20	130	150
Suma	80	220	300

En este caso X es la variable ‘Uso del cinturón’, que toma 2 valores posibles ($k = 2$), $x_1=\text{Si}$ y $x_2=\text{No}$. Y es la variable ‘Víctima de accidente’, que también toma 2 valores ($l = 2$), $y_1=\text{Si}$ y $y_2=\text{No}$. Las frecuencias absolutas conjuntas son $n_{11} = 60$, $n_{12} = 90$, $n_{21} = 20$ y $n_{22} = 130$. Las frecuencias absolutas parciales de la primera variable son $n_{1\bullet} = 150$ y $n_{2\bullet} = 150$, las de la segunda $n_{\bullet 1} = 80$ y $n_{\bullet 2} = 200$ y el número total de valores es $N = 300$.

Las expresiones para el cálculo de las frecuencias relativas y porcentajes conjuntos son similares a las explicadas en la sección anterior, pero aplicadas a estudiar la relación entre los valores conjuntos:

$$f_{ij} = \frac{n_{ij}}{n_{\bullet j}} \quad p_{ij} = \frac{n_{ij}}{n_{\bullet j}} \times 100 \%$$

Para el ejemplo sobre seguridad vial los porcentajes conjuntos de uso de cinturón respecto a ser víctima de accidente se muestran en la tabla ??.

Los porcentajes de la tabla deben mirarse por columnas: en la columna para el valor *Si* de la variable “victima de accidente” el valor 75,00% se obtiene al dividir 60 entre el total para la columna, 80, y sacar el porcentaje. Por otra parte 25,00% = $\frac{20}{80} \times 100$. Los porcentajes para *No* se obtienen de la forma siguiente: 40,91% = $\frac{90}{220} \times 100$ y 59,09% = $\frac{130}{220} \times 100$.

Observando los valores de la tabla anterior comprobamos que el uso del cinturón de seguridad es más frecuente entre aquellas personas que han sufrido algún accidente de tráfico que entre

Tabla 5.2: Seguridad vial: uso cinturón vs. víctimas accidentes

Uso cinturón	Victimas Accidente		Suma Filas
	Sí	No	
Si	60 75,00%	90 40,91%	150 50,00%
No	20 25,00%	130 59,09%	150 50,00%
Suma columnas	80 100,00%	220 100,00%	300 100,00%

aquellas que no lo han sufrido (75 % contra 40,91 %). De esta observaciones podemos deducir que existe una relación entre las variables. En las siguientes secciones aprenderemos a cuantificar este nivel de asociación.

5.2. Representación gráfica conjunta de dos variables

5.2.1. Diagramas de barras dobles

Como ya se ha comentado en el tema 2 una manera sencilla de representar de manera conjunta dos variables es mediante un diagrama de barras dobles.

Para el ejemplo sobre la seguridad vial se pueden representar los porcentajes relativos de las variables mediante el diagrama de la figura ??.

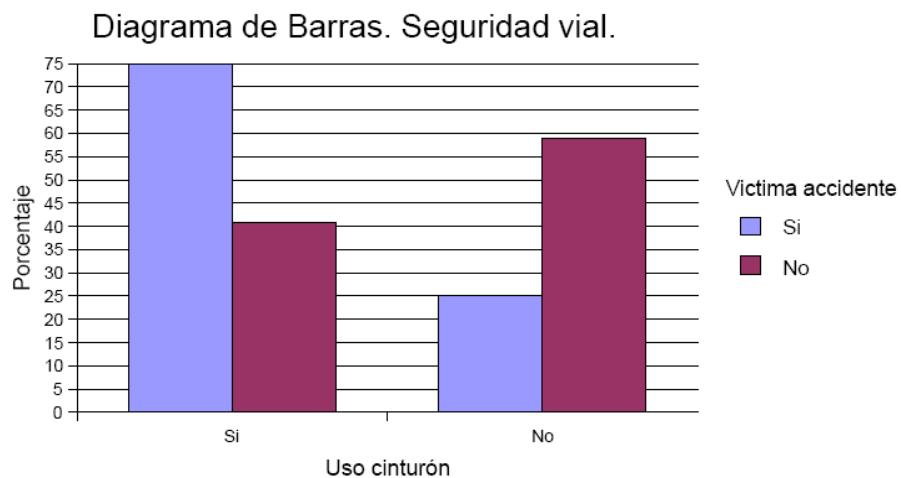


Figura 5.1: Diagrama de barras doble de porcentajes de la tabla ??

5.2.2. Diagramas de dispersión

Una manera mejor que los diagramas de barras dobles para representar gráficamente las relaciones entre dos variables ordinales o cuantitativas son los diagramas de dispersión. Estos

Tabla 5.3:

	Nota Media Bachillerato	Nota Sociología
Alumno 1	7,7	10
Alumno 2	7	2,25
Alumno 3	5,5	4,75
Alumno 4	6,2	4,25
Alumno 5	6,9	4,75
Alumno 6	5,8	0,75
Alumno 7	7,9	6
Alumno 8	6,7	3,75
Alumno 9	7,2	5
Alumno 10	5,4	5,75
Alumno 11	6,6	4,75
Alumno 12	8	6,25
Alumno 13	6,8	0,75
Alumno 14	7,1	1,25
Alumno 15	7	1,65
Alumno 16	5,8	8,5
Alumno 17	6,7	0,65

diagramas muestran en una misma gráfica los valores de ambas variables.

Consideremos el ejemplo de la siguiente tabla, en la que se muestran las notas de la asignatura de Sociología de unos alumnos de los estudios de Educación Social y se comparan con sus notas medias de Bachillerato.

Cabría esperar, en un principio, que aquellos alumnos que fueron buenos estudiantes de Bachillerato tengan mejores notas en su etapa universitaria. Para visualizar de manera gráfica la relación entre las variables ‘Nota de bachillerato’ y ‘Nota de sociología’ representamos los valores en el diagrama de dispersión de la figura ???. Cada punto de este diagrama representa la nota de un alumno: el valor en el eje horizontal representa su nota de bachillerato y el del eje vertical su nota de sociología.

Repetimos el diagrama para los datos de la tabla siguiente, en la que se comparan, para los mismos alumnos, las notas de bachillerato con las de la asignatura de Estadística. El diagrama se muestra en la figura ???.

Si observamos ambos diagramas vemos como en el primer caso los puntos parecen distribuidos al azar mientras que en la segunda gráfica parece que siguen una línea ascendente. Esta estructura de línea recta ascendente indica que los alumnos con peores notas de bachillerato son también los que obtienen peores notas de Estadística mientras que los que obtuvieron mejores notas de bachillerato son también los mejores alumnos de Estadística.

El hecho de que la distribución de los puntos en el diagrama de dispersión no sea aleatoria indica una relación de dependencia entre las variables. En las siguientes secciones aprenderemos a cuantificar esta relación.

La no existencia de relación entre las notas de Sociología y las de bachillerato podría indicar que se trata de una asignatura autocontenido, que no requiere conocimientos previos aprendidos

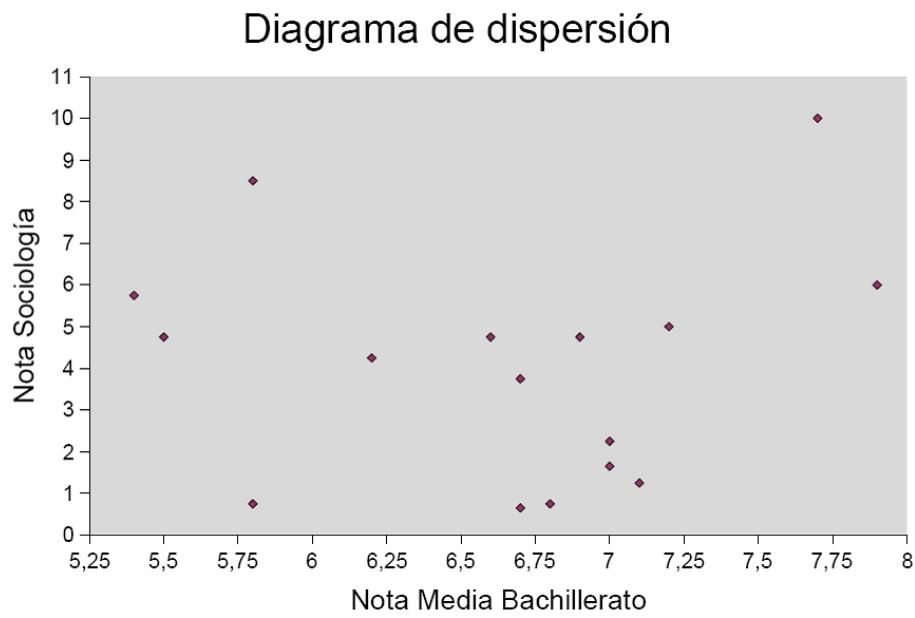


Figura 5.2: Diagrama de dispersión para los datos de la tabla ??

Tabla 5.4:

	Nota Media Bachillerato	Nota Estadística
Alumno 1	7,7	9
Alumno 2	7	7,5
Alumno 3	5,5	5
Alumno 4	6,2	5
Alumno 5	6,9	7
Alumno 6	5,8	5,5
Alumno 7	7,9	8,5
Alumno 8	6,7	6
Alumno 9	7,2	6,5
Alumno 10	5,4	4
Alumno 11	6,6	6
Alumno 12	8	7
Alumno 13	6,8	8
Alumno 14	7,1	7,5
Alumno 15	7	9
Alumno 16	5,8	4,5
Alumno 17	6,7	7

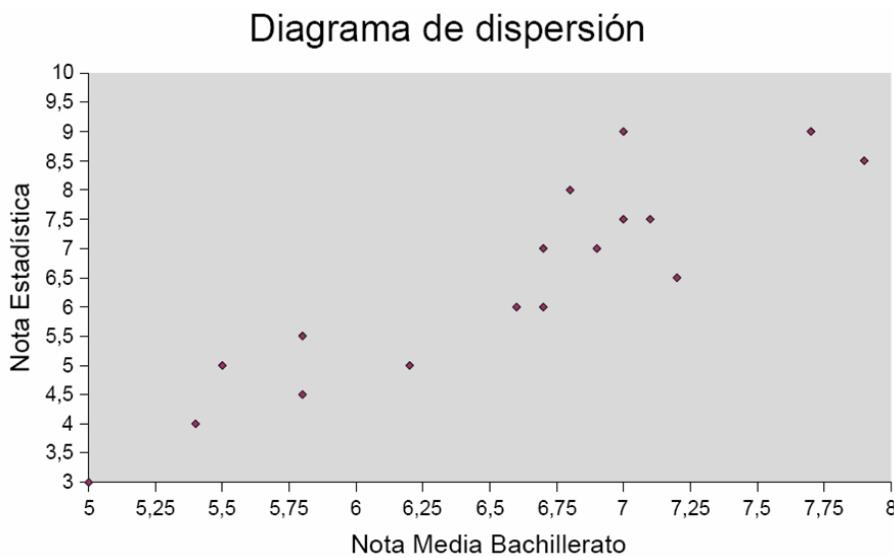


Figura 5.3: Diagrama de dispersión para los datos de la tabla ??

durante el bachillerato. Mientras que la dependencia entre las variables en el caso de las notas de Estadística sugiere que una buena base matemática en el bachillerato facilita la obtención de buenas notas durante la carrera.

5.3. Cuantificación de la relación entre variables estadísticas

Los principales estadísticos que permiten cuantificar la relación entre dos variables estadísticas son el coeficiente de contingencia y el de correlación. El primero se puede aplicar a cualquier tipo de variables mientras que el segundo sólo se define para variables de tipo cuantitativo.

5.3.1. Coeficiente de contingencia

Este estadístico permite medir el grado de *dependencia* entre dos variables estadísticas cualesquiera, cualitativas, ordinales o cuantitativas, cuyos datos estén organizados en una tabla de contingencia.

Antes de definir el coeficiente de contingencia es preciso clarificar primero la noción de dependencia/independencia entre variables estadísticas.

Se dice que dos variables x e y son **estadísticamente independientes** si para todos los valores de frecuencia conjunta de la tabla de contingencia se cumple que

$$\frac{n_{ij}}{N} = \frac{n_{i\bullet}}{N} \cdot \frac{n_{\bullet j}}{N}$$

Esta definición formaliza la intuición de que cuando dos variables son independientes la proporción de valores de cada una de ellas es la misma tengamos en cuenta o no a la otra variable.

Por ejemplo, las variables ‘Uso del cinturón’ y ‘Víctima de accidente’ de la tabla ?? no son estadísticamente independientes ya que

$$\frac{n_{11}}{N} = \frac{60}{300} = 0,2 \neq \frac{n_{1\bullet}}{N} \cdot \frac{n_{\bullet 1}}{N} = \frac{150}{300} \cdot \frac{80}{300} = 0,133$$

$$\frac{n_{12}}{N} = \frac{90}{300} = 0,3 \neq \frac{n_{1\bullet}}{N} \cdot \frac{n_{\bullet 2}}{N} = \frac{150}{300} \cdot \frac{220}{300} = 0,367$$

$$\frac{n_{21}}{N} = \frac{20}{300} = 0,067 \neq \frac{n_{2\bullet}}{N} \cdot \frac{n_{\bullet 1}}{N} = \frac{150}{300} \cdot \frac{80}{300} = 0,133$$

$$\frac{n_{22}}{N} = \frac{130}{300} = 0,433 \neq \frac{n_{2\bullet}}{N} \cdot \frac{n_{\bullet 2}}{N} = \frac{150}{300} \cdot \frac{220}{300} = 0,367$$

(en realidad bastaría comprobar que para uno de los valores no se cumple la condición para decidir que las variables no son independientes).

Aunque dos variables no sean estadísticamente independientes según la anterior definición ello no significa que sean totalmente independientes entre sí. El coeficiente de contingencia mide el grado de dependencia de las variables.

Este coeficiente (también conocido como **coeficiente C de contingencia de Pearson**) se define como:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

donde χ^2 es el estadístico **chi-cuadrado**, que se define a continuación:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

k y l son las cantidades de valores diferentes que toma cada variable y $e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N}$.

Interpretación del coeficiente de contingencia C toma un valor mínimo de 0 cuando las variables son completamente independientes y un valor máximo dado por la siguiente expresión:

$$\sqrt{1 - \frac{1}{\min(k, l)}} \quad \text{si } k \text{ y } l \text{ son mayores o iguales a 2}$$

Cuanto mayor es el valor del coeficiente de contingencia mayor es el grado de dependencia entre las variables.

Para el ejemplo de la tabla ??:

$$\begin{aligned} e_{11} &= \frac{n_{1\bullet} \cdot n_{\bullet 1}}{N} = \frac{150 \cdot 80}{300} = 40 \\ e_{12} &= \frac{n_{1\bullet} \cdot n_{\bullet 2}}{N} = \frac{150 \cdot 220}{300} = 110 \\ e_{21} &= \frac{n_{2\bullet} \cdot n_{\bullet 1}}{N} = \frac{150 \cdot 80}{300} = 40 \\ e_{22} &= \frac{n_{2\bullet} \cdot n_{\bullet 2}}{N} = \frac{150 \cdot 220}{300} = 110 \end{aligned}$$

$$\begin{aligned} \chi^2 &= \frac{(n_{11} - e_{11})^2}{e_{11}} + \frac{(n_{12} - e_{12})^2}{e_{12}} + \frac{(n_{21} - e_{21})^2}{e_{21}} + \frac{(n_{22} - e_{22})^2}{e_{22}} = \\ &= \frac{(60 - 40)^2}{40} + \frac{(90 - 110)^2}{110} + \frac{(20 - 40)^2}{40} + \frac{(130 - 110)^2}{110} = 27,27 \end{aligned}$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{27,27}{300 + 27,27}} = 0,288$$

Como el valor máximo de C es $\sqrt{1 - \frac{1}{\min(k,l)}} = \sqrt{1 - \frac{1}{2}} = 0,707$ el valor hallado representa un $\frac{0,288}{0,707} = 0,407 = 40,7\%$ del valor máximo, lo cual indica un cierto grado de dependencia entre las variables, pero no muy fuerte.

5.3.2. Coeficiente de correlación

El término **correlación** se utiliza en Estadística para denotar la relación entre dos o más variables. El coeficiente de correlación (también llamado **coeficiente de correlación lineal de Pearson**) permite medir el grado de relación lineal entre dos variables cuantitativas, es decir, permite decir en qué medida el diagrama de dispersión de las variables forma una línea recta.

Este estadístico se define como:

$$r = \frac{\text{Cov}}{s_X \cdot s_Y}$$

donde s_X y s_Y son las desviaciones típicas de las variables X e Y , respectivamente (la desviación típica se definió en el tema anterior) y Cov es un nuevo estadístico llamado **covarianza**. La covarianza se define de manera diferente cuando se aplica a datos de una población o de una muestra (de manera similar a lo que ocurría con la varianza):

$$\text{Cov} = \frac{\sum_{i=1}^k \sum_{j=1}^l n_{ij}(x_i - \bar{x})(y_j - \bar{y})}{N} = \frac{n_{11}(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + n_{kl}(x_k - \bar{x})(y_l - \bar{y})}{N} \quad (\text{covarianza poblacional})$$

$$\text{Cov} = \frac{\sum_{i=1}^k \sum_{j=1}^l n_{ij}(x_i - \bar{x})(y_j - \bar{y})}{N-1} = \frac{n_{11}(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + n_{kl}(x_k - \bar{x})(y_l - \bar{y})}{N-1} \quad (\text{covarianza muestral})$$

donde \bar{x} y \bar{y} son las medias de las variables X e Y , respectivamente, definidas en el tema 3.

Tomemos por ejemplo los datos de la siguiente tabla de contingencia en la que se relacionan la edad (en años) y el número de hijos para un grupo de 100 mujeres:

Edad	Hijos	0	1	2	3
20	19	1	1	0	
25	12	5	3	0	
30	9	6	4	1	
35	5	7	4	2	
40	3	5	10	3	

Si llamamos x a la variable ‘Edad’ y y a la variable ‘Número de hijos’, la correlación entre ellas se calcula de la siguiente forma:

1. Para poder calcular las medias y las varianzas debemos hallar primero las frecuencias absolutas de las variables. Para ello sumamos todos los valores de frecuencias conjuntas, tanto por filas como por columnas:

Edad	Hijos	0	1	2	3	Suma
20	19	1	1	0	21	
25	12	5	3	0	20	
30	9	6	4	1	20	
35	5	7	4	2	18	
40	3	5	10	3	21	
Suma	48	24	22	6		

Observando la nueva tabla vemos que las frecuencias absolutas para la variable ‘Edad’ son: $n_{20\bullet} = 21$ (21 personas de 20 años), $n_{25\bullet} = 20$ (20 personas de 25 años), etc. Y las frecuencias absolutas para la variable ‘Número de hijos’ son: $n_{\bullet 0} = 48$ (48 casos de personas sin hijos), $n_{\bullet 1} = 24$ (24 casos de personas con un sólo hijo), etc. Por otra parte, la suma de todas las frecuencias conjuntas es $N = 19 + 1 + 1 + \dots + 10 + 3 = 100$.

- Ahora podemos calcular las medias y varianzas de cada variable. Como deseamos extrapolar nuestro análisis estadístico a un grupo mayor, calcularemos en este caso varianzas muestrales:

$$\bar{x} = \frac{21 \cdot 20 + 20 \cdot 25 + 20 \cdot 30 + 18 \cdot 35 + 21 \cdot 40}{100} = 29,9$$

$$\bar{y} = \frac{48 \cdot 0 + 24 \cdot 1 + 22 \cdot 2 + 6 \cdot 3}{100} = 0,86$$

$$\text{Var}_X = \frac{21 \cdot (20 - 29,9)^2 + 20 \cdot (25 - 29,9)^2 + \dots + 21 \cdot (40 - 29,9)^2}{99} = 52,01$$

$$\text{Var}_Y = \frac{48 \cdot (0 - 0,86)^2 + 24 \cdot (1 - 0,86)^2 + \dots + 6 \cdot (3 - 0,86)^2}{99} = 0,93$$

$$s_X = \sqrt{52,01} = 7,21$$

$$s_Y = \sqrt{0,93} = 0,96$$

- Finalmente calculamos la covarianza (muestral) y la correlación:

$$\text{Cov} = \frac{19 \cdot (20 - 29,9) \cdot (0 - 0,86) + 12 \cdot (25 - 29,9) \cdot (0 - 0,86) + \dots + 2 \cdot (35 - 29,9) \cdot (3 - 0,86) + 3 \cdot (40 - 29,9) \cdot (3 - 0,86)}{99} = 3,72$$

$$r = \frac{3,72}{7,21 \cdot 0,96} = 0,54$$

Este valor indica una débil correlación lineal entre la edad de las mujeres y su número de hijos, para la muestra considerada, tal como muestra el correspondiente diagrama de dispersión (figura ??).

Cálculo de la covarianza y la correlación a partir de datos organizados en intervalos

Supongamos que queremos calcular la autocorrelación de las variables ‘Temperatura media anual’ (T , en $^{\circ}\text{C}$) y ‘Latitud’ (L , en $^{\circ}$) de varias ciudades (fuente *weatherbase.com*) a partir de la siguiente tabla de contingencia:

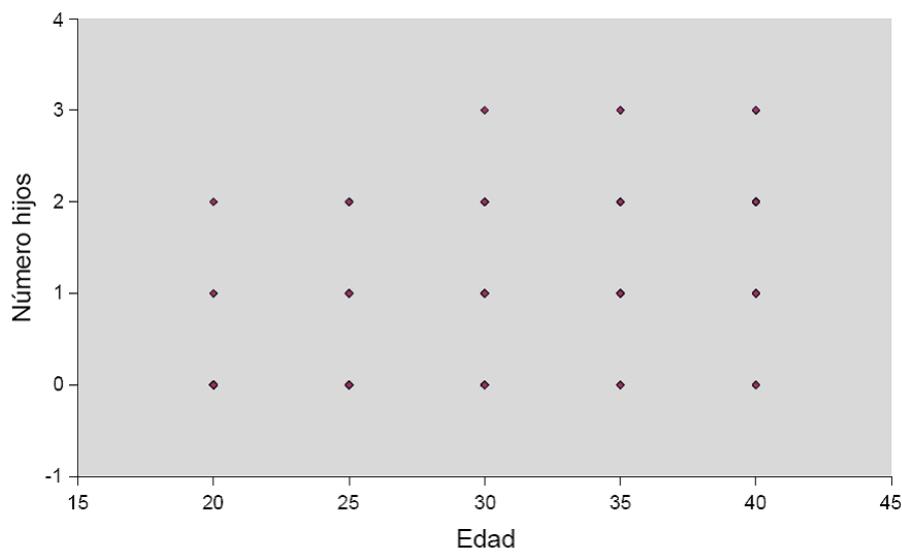


Figura 5.4: Diagrama de dispersión Edad-Número de hijos. El coeficiente de correlación es 0,54

T \ L	[0 – 10)	[10 – 20)	[20 – 30)	[30 – 40)	[40 – 50)	[50 – 60)
[0 – 5)	0	0	0	0	1	3
[5 – 10)	0	0	0	0	0	2
[10 – 15)	1	0	0	1	5	1
[15 – 20)	0	1	0	5	0	0
[20 – 25)	0	2	5	1	0	0
[25 – 30)	6	4	2	0	0	0

Los datos de ambas variables se hallan agrupados en intervalos, por lo que en primer lugar calcularemos los valores centrales de cada intervalo:

T \ L	5	15	25	35	45	55	Suma
2,5	0	0	0	0	1	3	4
7,5	0	0	0	0	0	2	2
12,5	1	0	0	1	5	1	8
17,5	0	1	0	5	0	0	6
22,5	0	2	5	1	0	0	8
27,5	6	4	2	0	0	0	12
Suma	7	7	7	7	6	6	

A partir de aquí el cálculo se hace como en el ejemplo anterior, utilizando valores centrales en lugar de los valores originales (calcularemos también varianzas y covarianzas muestrales pues deseamos generalizar los resultados de nuestro estudio).

$$\bar{T} = \frac{4 \cdot 2,5 + 2 \cdot 7,5 + \dots + 12 \cdot 27,5}{40} = 18,5$$

$$\bar{L} = \frac{7 \cdot 5 + 7 \cdot 15 + \dots + 6 \cdot 55}{40} = 29$$

$$\text{Var}_T = \frac{4 \cdot (2,5 - 18,5)^2 + 2 \cdot (7,5 - 18,5)^2 + \dots + 12 \cdot (27,5 - 18,5)^2}{39} = 68,21$$

$$\text{Var}_L = \frac{7 \cdot (5 - 29)^2 + 7 \cdot (15 - 29)^2 + \dots + 6 \cdot (55 - 29)^2}{39} = 291,28$$

$$s_T = \sqrt{68,21} = 8,26$$

$$s_L = \sqrt{291,28} = 17,07$$

$$\text{Cov} = \frac{0 \cdot (2,5 - 18,5) \cdot (5 - 29) + 0 \cdot (7,5 - 18,5) \cdot (5 - 29) + \dots + 0 \cdot (27,5 - 18,5) \cdot (55 - 29)}{39} = -119,49$$

$$r = \frac{-119,49}{8,26 \cdot 17,07} = -0,85$$

La conclusión es que existe una correlación lineal negativa bastante fuerte entre latitud y temperatura (a mayor latitud menor temperatura) tal y como muestra el diagrama de dispersión de la figura ??.

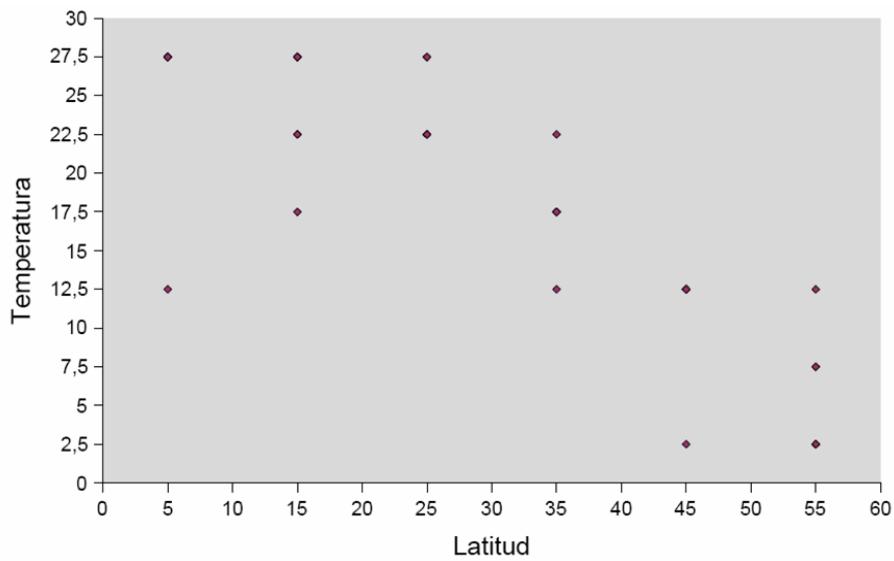


Figura 5.5: Diagrama de dispersión Latitud-Temperatura. El coeficiente de correlación es $-0,85$

Cálculo de la covarianza y correlación a partir de datos brutos

Al igual que ocurría con el cálculo de la varianza, el cálculo de la covarianza a partir de datos brutos puede hacerse de una manera muy sencilla utilizando las siguientes fórmulas:

$$\text{Cov} = \frac{\sum_{i=1}^N x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y} \quad (\text{covarianza poblacional})$$

$$\text{Cov} = \frac{\sum_{i=1}^N x_i \cdot y_i}{N-1} - \frac{N}{N-1} \cdot \bar{x} \cdot \bar{y} \quad (\text{covarianza muestral})$$

donde x_i e y_i representa los distintos valores de las variables, N es el número total de valores y \bar{x} y \bar{y} sus medias.

Calculemos por ejemplo la covarianza y el coeficiente de correlación para los datos de la tabla ?? (llamamos X a la nota de bachillerato y Y a la nota de la asignatura). Como no deseamos extrapolar los resultados de nuestro análisis al análisis de un grupo mayor, calcularemos varianzas y covarianza poblacionales.

$$\bar{x} = \frac{7,7+7+5,5+\dots+7+5,8+6,7}{17} = 6,72$$

$$\bar{y} = \frac{10+2,25+4,75+\dots+1,65+8,5+0,65}{17} = 4,18$$

$$\text{Var}_X = \frac{7,7^2+7^2+5,5^2+\dots+7^2+5,8^2+6,7^2}{17} - 6,72^2 = 0,57$$

$$\text{Var}_Y = \frac{10^2+2,25^2+4,75^2+\dots+1,65^2+8,5^2+0,65^2}{17} - 4,18^2 = 7,01$$

$$s_X = \sqrt{0,57} = 0,755$$

$$s_Y = \sqrt{7,01} = 2,648$$

$$\text{Cov} = \frac{7,7 \cdot 10 + 7 \cdot 2,25 + 5,5 \cdot 4,75 + \dots + 7 \cdot 1,65 + 5,8 \cdot 8,5 + 6,7 \cdot 0,65}{17} - 6,72 \cdot 4,18 = 0,28$$

$$r = \frac{0,28}{0,755 \cdot 2,648} = 0,14$$

Procediendo de manera similar para los datos de la tabla ?? obtendríamos los siguientes valores:

$$\bar{x} = 6,72$$

$$\bar{y} = 6,65$$

$$\text{Var}_X = 0,57$$

$$\text{Var}_Y = 2,2$$

$$s_X = \sqrt{0,57} = 0,755$$

$$s_Y = \sqrt{7,01} = 1,483$$

$$\text{Cov} = 0,93$$

$$r = \frac{0,93}{0,755 \cdot 1,483} = 0,82$$

Interpretación del coeficiente de correlación

El coeficiente de correlación toma valores entre -1 y 1. Un valor 0 indica que no existe relación lineal entre las variables. Un valor 1 o -1 indica que la relación lineal es máxima.

El signo positivo indica una correlación positiva, es decir, a mayores valores de una variable le corresponden mayores valores de la otra variable. Mientras que un signo negativo indica una

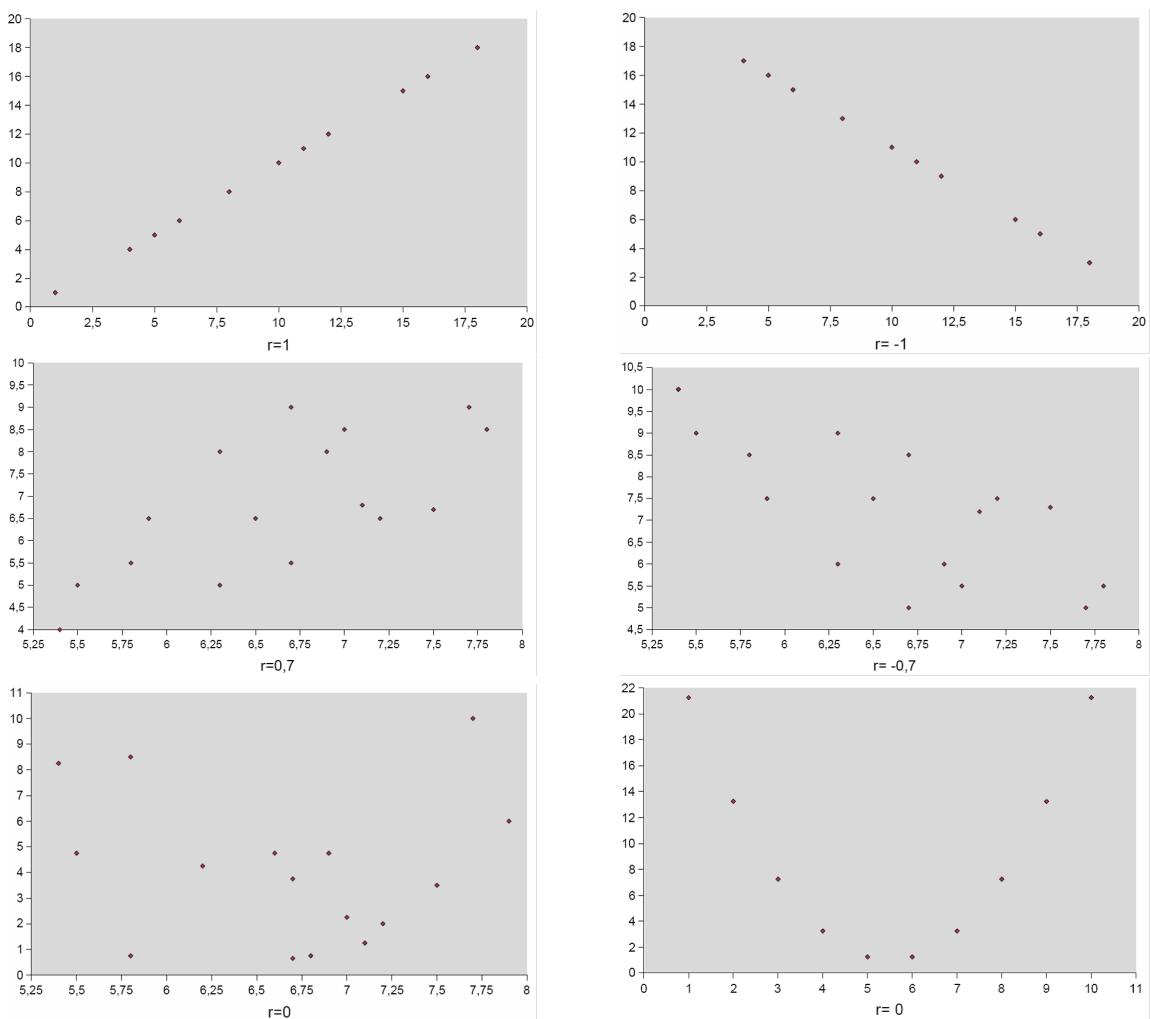


Figura 5.6: Diagramas de dispersión y correspondientes coeficientes de correlación

correlación negativa: a mayores valores de una variable le corresponden valores menores de la otra variable

En la figura ?? se muestran distintos diagramas de dispersión de 2 variables y sus correspondientes coeficientes de correlación.

A la hora de interpretar el coeficiente de correlación entre dos variables hay que tener presentes las siguientes observaciones:

- Un valor alto del coeficiente de correlación no implica una relación de causa-efecto entre las variables: los valores de una variable pueden depender de los de la otra, pero también puede pasar que ambos valores dependan de una tercera variable.
- La ausencia de relación entre dos variables (distribución prácticamente aleatoria de valores en el diagrama de dispersión) implica un coeficiente de correlación próximo a cero. Pero no siempre un coeficiente próximo a cero implica la ausencia de relación entre las variables: la relación entre ellas puede ser no lineal (ver Figura ??, inferior-derecha).
- Si las variables son estadísticamente independientes su coeficiente de correlación es cero. Sin embargo, que el coeficiente de correlación sea cero no implica necesariamente que las

variables sean independientes. Sólo el coeficiente de contingencia permite determinar la dependencia o independencia de dos variables.

5.4. Regresión lineal y predicción

Consideremos el diagrama de dispersión de la figura ??, correspondiente a la tabla de valores ??, y supongamos que queremos *predecir* la nota de la asignatura de Estadística de un alumno que obtuvo un 6,5 como nota media de bachillerato.

Evidentemente el valor que demos como resultado siempre puede ser erróneo pero deseamos hacer la mejor predicción posible suponiendo que la nota del alumno sigue la misma tendencia que las notas de sus compañeros.

Para ello debemos calcular la **recta de regresión lineal** de los datos, que indica la tendencia de los mismos. Esta recta, que a continuación calcularemos, se muestra en la figura ?? y es la que mejor se ajusta al conjunto de datos (la suma de las distancias de cada punto a la recta es mínima). Una vez calculada la recta, la mejor estimación que podemos hacer del valor solicitado es la nota de Estadística correspondiente a la nota 6,5 de bachillerato según la recta de regresión (ver figura ??). En este caso el valor es 6,29.

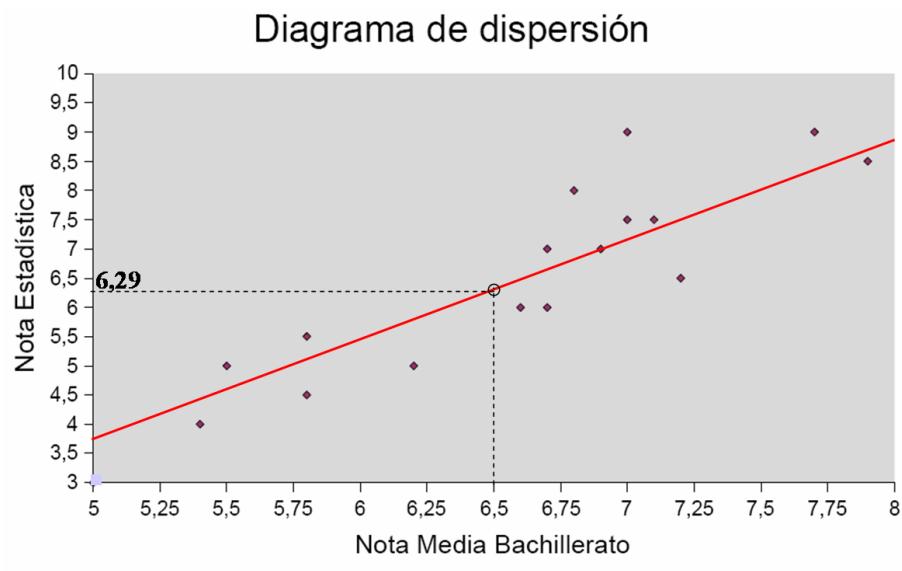


Figura 5.7: Diagrama de dispersión para la tabla ?? y recta de regresión

El cálculo de la recta de regresión se hace mediante las siguientes fórmulas:

$$\hat{Y} = ax + b$$

$$a = \frac{\text{Cov}}{\text{Var}_X}$$

$$b = \bar{y} - a \cdot \bar{x}$$

donde \hat{Y} denota el valor estimado de la variable Y a partir del valor conocido x . En nuestro caso los valores de covarianza, varianzas y esperanzas se han calculado en la sección anterior, de manera que $a = \frac{0,93}{0,57} = 1,63$, $b = 6,65 - 1,63 \cdot 6,72 = -4,3$ y $\hat{Y} = 1,63 \cdot 6,5 - 4,3 = 6,29$.

5.5. Análisis bivariante con ayuda del ordenador

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 C de contingencia 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 como la mostrada en la sección ??.

	A	B	C
1		Varón	Mujer
2	Reincidente		
3	No reincidente		
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:

- a) nos situamos en la casilla B4 y escribimos =SUMA(B2:B3). Al pulsar *Enter* obtenemos el valor $n_{\bullet 1} = 112001$.

- b) 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				

3. Para calcular los valores e_{ij} 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				

4. A continuación 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 $B9$
 - a partir de la esquina inferior derecha de $B9$ extendemos el cálculo a $C9$
 - 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 C de contingencia:
- 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 de la sección ??: escribimos $=RAÍZ(C12/(D4+C12))$ en $C13$, pulsamos *Enter* y obtenemos $C = 0,07$.

- c) Para decidir si este valor es alto o bajo debemos calcular el valor máximo de C , según la fórmula de la sección ??: como ambas variables constan de dos únicos valores, $\min\{k, l\} = 2$, por tanto escribimos =RAÍZ(1-1/2) en C14. Al pulsar *Enter* obtenemos $C_{max} = 0,71$.
- d) La proporción de C respecto de C_{max} se calcula en C15 con la fórmula =100*C13/C14. El valor obtenido es 9,5 %.

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 obtenido (9,5 % respecto al máximo posible) indica que las variables ‘Reincidencia’ y ‘Sexo’ del delincuente son prácticamente independientes: la probabilidad de ser reincidente 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: Consellería de Medi Ambient y Instituto Nacional de Meteorología).

Año	Precipitationes (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.

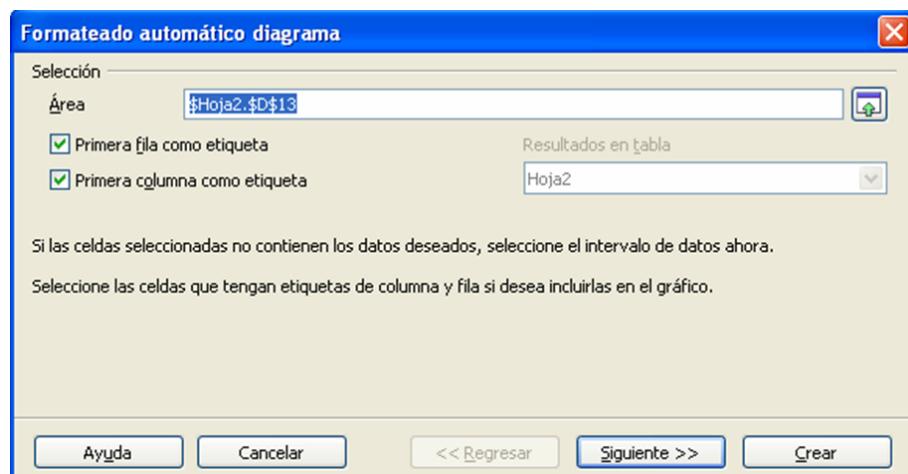
Calculamos la recta de regresión con la fórmula de la sección ???. Para utilizar la fórmula debemos calcular:

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 $88, 31$)
4. 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 $b = 139, 96$.

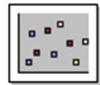
La ecuación de la recta de regresión es por tanto: $\hat{Y} = -0, 12X + 139, 96$. De manera que el valor estimado para $x = 550$ será: $\hat{Y} = -0, 12 \cdot 550 + 139, 96 = 73, 96$.

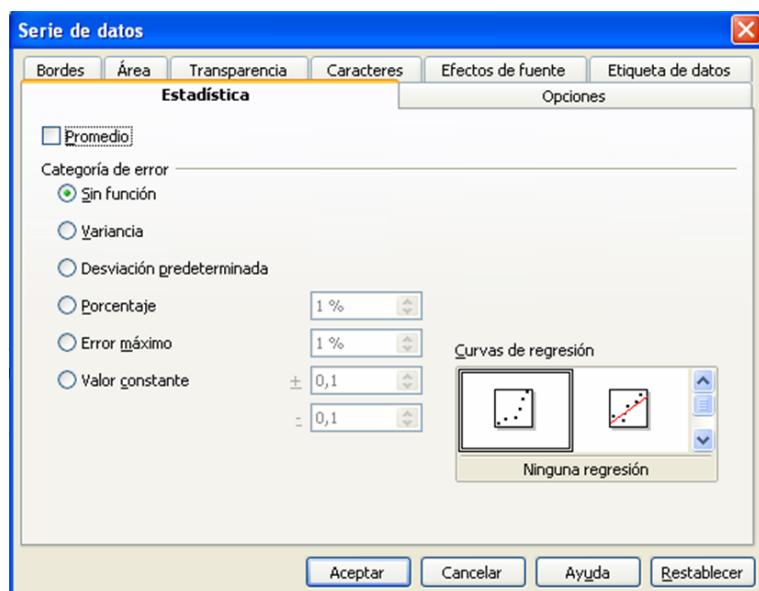
El diagrama de dispersión se dibuja fácilmente con Calc:

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



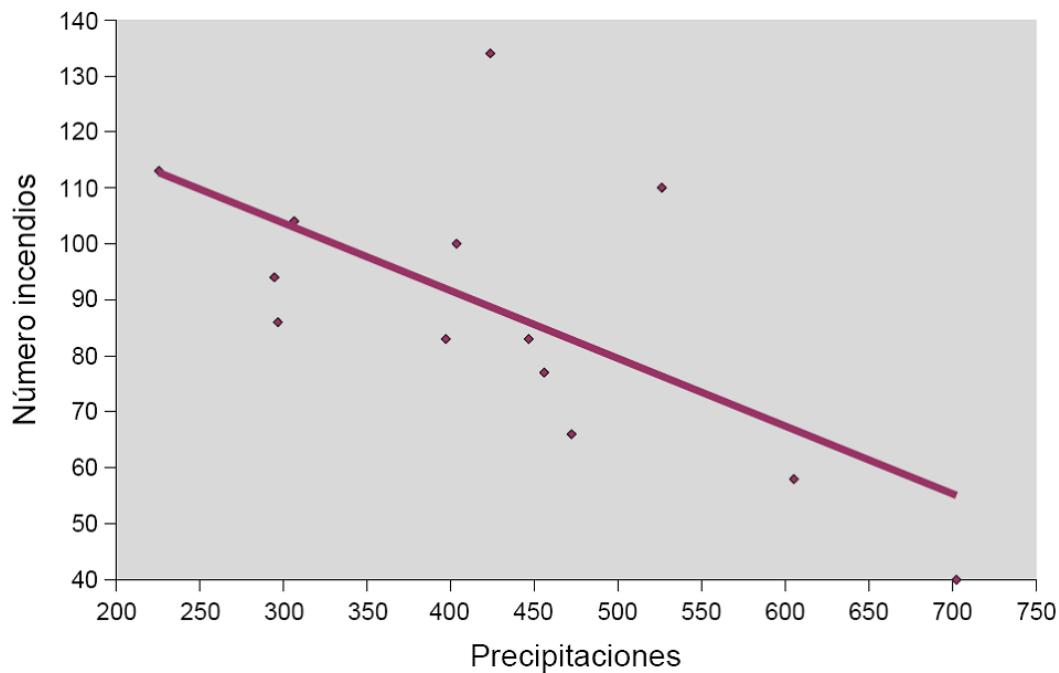
3. Seleccionamos con el cursor las casillas que contienen los datos (de $A2$ a $B14$) y desactivamos las opciones *Primera fila como etiqueta* y *Primera columna como etiqueta* del cuadro de diálogo. Pulsamos el botón *Siguiente*.

4. En el nuevo cuadro de diálogo que aparece seleccionamos la opción  y avanzamos.
5. En el siguiente diálogo desactivamos la opción *Eje Y* y avanzamos.
6. En el último diálogo desactivamos las opciones *Título de diagrama* y *Leyenda* y escribimos **Precipitaciones** e **Número incendios**, respectivamente, en las opciones *Título del Eje X* y *Título del Eje Y*.
7. Pulsamos la tecla *Crear* y el diagrama aparece en la posición seleccionada. Ahora podemos reescalarlo con el cursor a un tamaño mayor.
8. Si deseamos dibujar la recta de regresión procedemos del siguiente modo:
 - a) Nos situamos sobre el diagrama y hacemos doble-click sobre cualquiera de los puntos dibujados.
 - b) Aparece una nueva ventana de diálogo en la que seleccionamos la pestaña *Estadística*:



- c) Dentro de las opciones de *Curvas de regresión* seleccionamos el ícono  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:



5.6. Ejercicios propuestos

Ejercicio 1

Calcular el coeficiente C de contingencia para las variables ‘Tipo de delito’ y ‘Edad’ de los condenados en el año 2006 a partir de los siguientes datos y comentar los resultados.

Estadísticas judiciales 2006							
Estadística de lo Penal. Condenados. Resultados nacionales							
Condenados según tipo de delito, edad y sexo							
Unidades: nº de condenados							
	De 18 a 20 años	De 21 a 25 años	De 26 a 30 años	De 31 a 35 años	De 36 a 40 años	De 41 a 50 años	De 51 a 60 años
	Ambos sexos						
Homicidio y formas	12	78	78	72	73	93	61
De las lesiones	629	3.029	3.712	3.295	3.118	3.985	1.523
Contra la libertad	68	270	438	503	527	808	361
Contra el orden público	314	943	1.074	912	841	965	320

Fuente: Instituto Nacional de Estadística

Ejercicio 2

Hallar la covarianza y el coeficiente de correlación para las variables ‘Población residente de Alemania y Reino Unido’ y ‘Tasa de ocupación hotelera’ en Mallorca a partir de los datos de la siguiente tabla (fuentes: IBAB y Conselleria de Turisme).

Año	Residentes Alemania y Reino Unido	Tasa ocupación hotelera
1998	13191	83,9
1999	15955	83,7
2000	18943	79,5
2001	22028	78,6
2002	24934	72,2
2003	28147	72,4
2004	25293	73
2005	29307	72,8

Ejercicio 3

Calcular la recta de regresión lineal para los datos del ejercicio anterior y predecir a partir de ella el valor de la tasa de ocupación hotelera si el número de residentes alemanes y británicos llega a 35000. Dibujar el diagrama de dispersión y representar sobre él la recta de regresión.

Módulo III

En este módulo se estudian las técnicas básicas de la Estadística Inferencial, que permiten conocer el grado de fiabilidad de la generalización de los datos obtenidos mediante la Estadística Descriptiva a conjuntos de datos mayores.