

3. 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

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

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

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

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 recorrido intercuartílico es por tanto: $RIC = 367 - 264, 25 = 102, 75$.

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

1. Escribimos las palabras “Varianza”, “Desv. típica”, “Coef. Variación”, “Mínimo”, “Máximo” y “Recorrido” en las casillas C20 a C25, por ejemplo, de la

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

Figura 25: Datos *brutos* del ejemplo 2

hoja de cálculo.

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

La siguiente tabla resume los resultados obtenidos hasta el momento:

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

Ejemplo 2

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

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

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

muestra formada por 300 alumnos del total de estudiantes de secundaria de la Baleares. Calcularemos por tanto la varianza muestral.

El cálculo se hace en dos pasos:

- 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.
- 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 26). El resultado final es 37,41.

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

Figura 26: Hoja de cálculo del ejemplo 3.

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

- 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.

Ejemplo 3

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

Estadísticas judiciales 2005	
Estadística de lo Penal. Condenados. Resultados autonómicos	
Condenados según edad y sexo	
Unidades: nº de condenados	
Ambos sexos	
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 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 recorrido intercuartílico es $RIC = 15,61$. Por otra parte, el valor mínimo de la variable *Edad* es 18 y, según el enunciado, el máximo es 70. De manera que el recorrido es $70 - 18 = 52$.

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

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

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

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

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

Figura 27: Hoja de cálculo del ejemplo 4.

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

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

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

3.1. Ejercicios propuestos

Ejercicio 1

Calcular el recorrido, recorrido intercuartílico, varianza, desviación estándar y coeficiente de variación para el número de farmacias por municipios en Mallorca a partir de los datos de la tabla 1.

Cuadro 1: 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	Esorca	1	Marratxí	8	Santanyí	8
Artà	3	Esporles	1	Montuïri	1	Selva	2
Banyalbufar	1	Estellencs	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 2 calcular la media, la varianza y la desviación típica de la variable “Edad de víctimas de accidentes en 2006”.

Cuadro 2: 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. Decide a qué mensajero aumentar el sueldo.

4. Cálculo de las medidas de simetría y apuntamiento con ayuda del ordenador

Ejemplo 1

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

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

Notas:

1.- EJC: equivalencia a jornada completa

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

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

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

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

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

Ejemplo 2

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

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

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

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

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

Figura 28: Hoja de cálculo del ejemplo 1.

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

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

Ejemplo 3

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

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

Debemos calcular:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

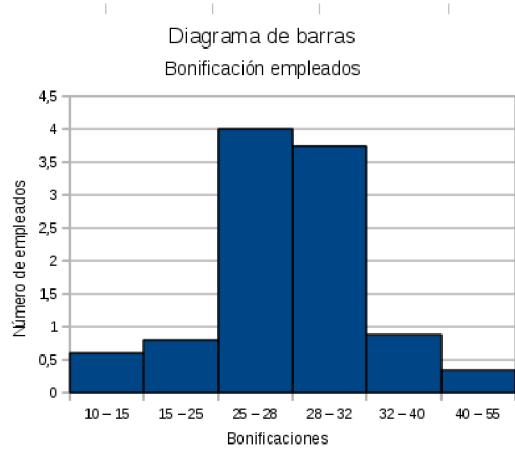


Figura 31: Histograma del ejemplo 3.

4.1. Ejercicios propuestos

Ejercicio 1

La puntuación que han obtenido 50 personas que se presentaron para ocupar un puesto en la plantilla de una empresa, ha sido la siguiente:

Puntuación	N. personas
14 - 18	3
18 - 20	6
20 - 25	11
25 - 28	15
28 - 32	8
32 - 36	7

Calcular la media y la varianza de las notas, así como el coeficiente de simetría y la curtosis de los datos. Interpretar los resultados obtenidos.

Ejercicio 2 La actividad física habitual en personas de la llamada tercera edad es recomendable por sus efectos beneficiosos tanto a nivel biomédico como psicológico. En un centro de salud se ha efectuado un estudio piloto en el que se han valorado el estado de salud general en 33 personas con edades comprendidas entre 65 y 75 años. La evaluación del estado de salud consistió en una relación de pruebas que puntuaban en una escala graduada entre 25 y 100. La tabla 3 muestra los datos sobre el estado de salud, así como una variable de control que señalaba si el sujeto realizaba ejercicio físico o no con cierta regularidad. Analizad, calculando la media, la desviación estándar, la simetría, la curtosis y el coeficiente de variación, los datos del estado de salud en función de si los sujetos siguen una pauta de actividad física determinada o no y en general sin distinguir los casos.

Caso	Activ.	Salud	Caso	Activ.	Salud	Caso	Activ.	Salud
1	No	38	12	No	49	23	Sí	63
2	No	41	13	No	50	24	Sí	75
3	No	41	14	No	51	25	Sí	70
4	No	42	15	No	51	26	Sí	78
5	No	43	16	No	55	27	Sí	78
6	No	47	17	No	55	28	Sí	81
7	No	47	18	No	42	29	Sí	82
8	No	49	19	Sí	52	30	Sí	82
9	No	49	20	Sí	53	31	Sí	66
10	No	49	21	Sí	55	32	Sí	66
11	No	49	22	Sí	57	33	Sí	69

Cuadro 3: Datos del estudio sobre el estado de salud en personas de la tercera edad (Activ: actividad física regular; Salud: puntuación en el estado de salud).

Ejercicio 3

A continuación se detallan los kilogramos de compost producidos por las diferentes plantas de triaje y compostaje de España en el 2004 (fuente INE):

11636 12654 6840 12358 15542 10600 10083 4000 4904 22909 19227 1700 15877
6184 30000 6000 1325 2200 1800 6984 900 7689 11404 5799 1585 6950 3100 1199
2694 7320 2500 2545 6515 2578 1812 2443 1314 6708 8668 9189 14027 3500 11128
48113 44563 8708 1653 3432 70576 1255 17444 37700 22631 8733 11200 5562 14884
3417

Se pide:

1. Utilizando los datos en “bruto”, calcular la media, mediana, coeficiente de simetría y curtosis de los datos.
2. Agrupando los datos en 8 intervalos, realizar los mismos cálculos que el apartado anterior.

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 de correlación de Pearson para las variables “Sexo” y “Reincidencia” de los condenados en el año 2006 a partir de los siguientes datos.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

La hoja de cálculo final muestra el siguiente aspecto:

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

Comentario.

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

Ejemplo 2

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

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

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

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

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

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

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

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

La hoja de cálculo final muestra el siguiente aspecto:

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

Comentario.

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

Ejemplo 3

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

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

- la covarianza ($-1966,63$, calculada en el ejemplo anterior),
- la varianza de la primera variable (fórmula $=VARP(A2:A14)$, resultado $16272,15$),
- 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 clic sobre el icono  del menú *Insertar* y a continuación sobre una casilla cualquiera para insertar el gráfico en esa posición. Aparece el siguiente cuadro de diálogo:



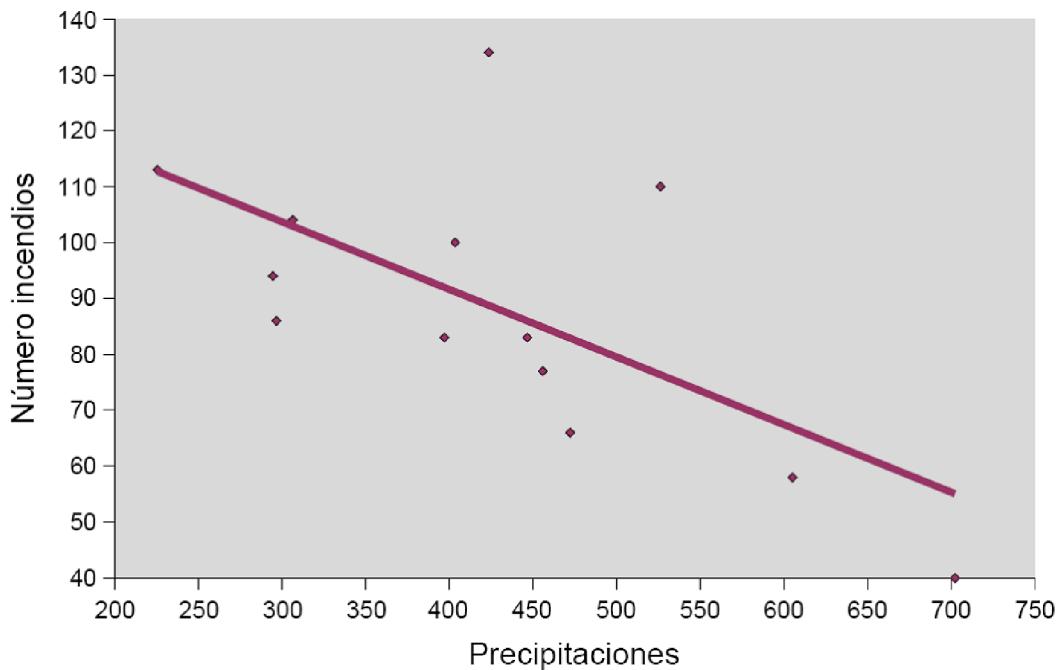
3. Seleccionamos la opción  y la opción *Sólo puntos*.
4. En el rango de datos escribimos $A1:B14$ y pulsamos el botón *Siguiente*.
5. En el diálogo *Series de datos* hacemos clic sobre *Valores X* y escribimos $A2:A14$ en *Rango para valores X*. Repetimos el proceso para los valores Y, cuyo rango es $B2 : B14$, y pulsamos *Siguiente*.
6. En el último diálogo desactivamos la opción *Mostrar leyenda* y escribimos **Precipitaciones** e **Número incendios**, respectivamente, en las opciones *Título del Eje X* y *Título del Eje Y*. También desactivamos la opción *Eje Y*.
7. Pulsamos la tecla *Finalizar* 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 clic sobre cualquiera de los puntos dibujados. Todos los puntos quedarán marcados.
- b) Hacemos clic con el botón derecho del ratón sobre cualquiera de los puntos y aparecerá un menu desplegable en el que seleccionamos la opción *Insertar Línea de Tendencia* ...
- c) Dentro de las opciones de *Línea de tendencia* seleccionamos el ícono



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

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



5.1. Ejercicios propuestos

Ejercicio 1

Calcular el coeficiente de correlación de Pearson 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.