

# Índice general

## I Módulo II: Interpretación de datos estadísticos bi-variantes.

Introducción a la probabilidad y las variables aleatorias	3
1. Relaciones entre dos variables estadísticas	5
1.1. Tablas de contingencia . . . . .	5
1.2. Representación gráfica conjunta de dos variables . . . . .	7
1.2.1. Diagramas de barras dobles . . . . .	7
1.2.2. Diagramas de dispersión . . . . .	7
1.3. Cuantificación de la relación entre variables estadísticas . . . . .	12
1.3.1. Coeficiente de contingencia . . . . .	12
1.3.2. Coeficiente de correlación . . . . .	14
1.4. Regresión lineal y predicción . . . . .	22
1.5. Análisis bivalente con ayuda del ordenador . . . . .	23
1.6. Ejercicios propuestos . . . . .	31



# Índice de cuadros

1.1. Seguridad vial: uso cinturón vs. víctimas accidentes . . . . .	7
1.2. . . . .	8
1.3. . . . .	11





## Parte I

# Módulo II: Interpretación de datos estadísticos bivariantes. Introducción a al probabilidad y las variables aleatorias



En este módulo se culmina 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 tema se dan los fundamentos de Probabilidad necesarios para el estudio de la Estadística Inferencial.



# Capítulo 1

## 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 interesará 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.

### 1.1. 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
- $k$  i  $l$  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  $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{il}$  y  $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj}$
- $N$  es el número total de valores conjuntos:  $N = n_{11} + n_{12} + \cdots + n_{kl}$

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 “víctima 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:



Cuadro 1.1: 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  $N = 300$ .

## 1.2. Representación gráfica conjunta de dos variables

### 1.2.1. 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 las frecuencias absolutas de las variables mediante el diagrama de la figura 1.1.

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

Consideremos el ejemplo de la tabla 1.2 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.

Cabe 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



Cuadro 1.2:

	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





Diagrama de Barras. Seguridad vial.

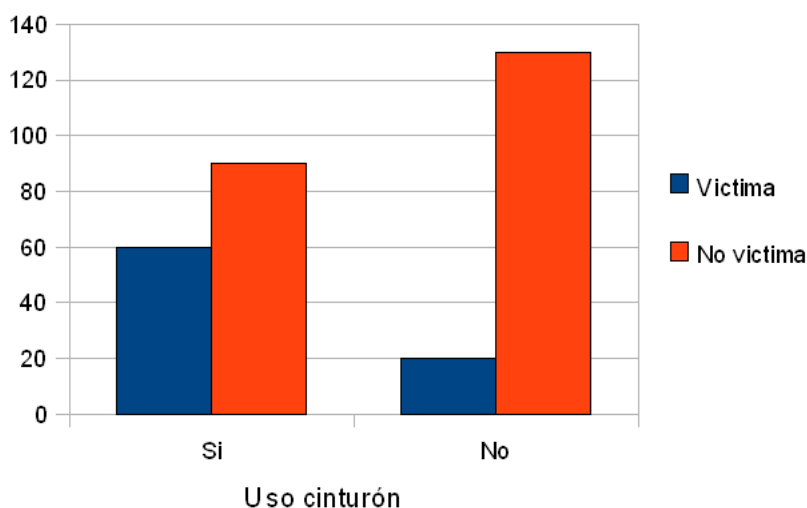


Figura 1.1: Diagrama de barras doble correspondiente a la tabla 1.1

bachillerato' y 'Nota de sociología' representamos los valores en el diagrama de dispersión de la figura 1.2. 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 1.3 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 1.3.

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 autocontenida, que no

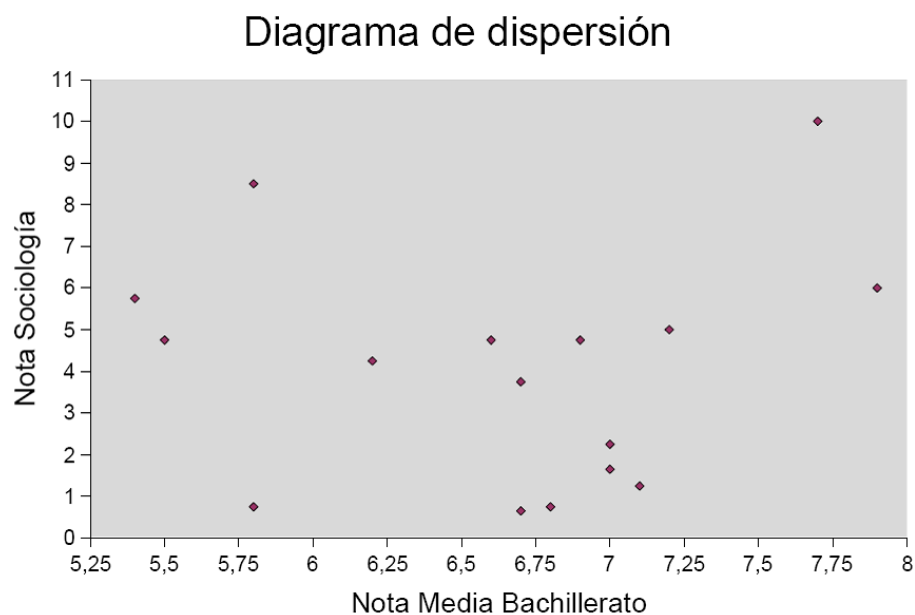


Figura 1.2: Diagrama de dispersión para los datos de la tabla 1.2

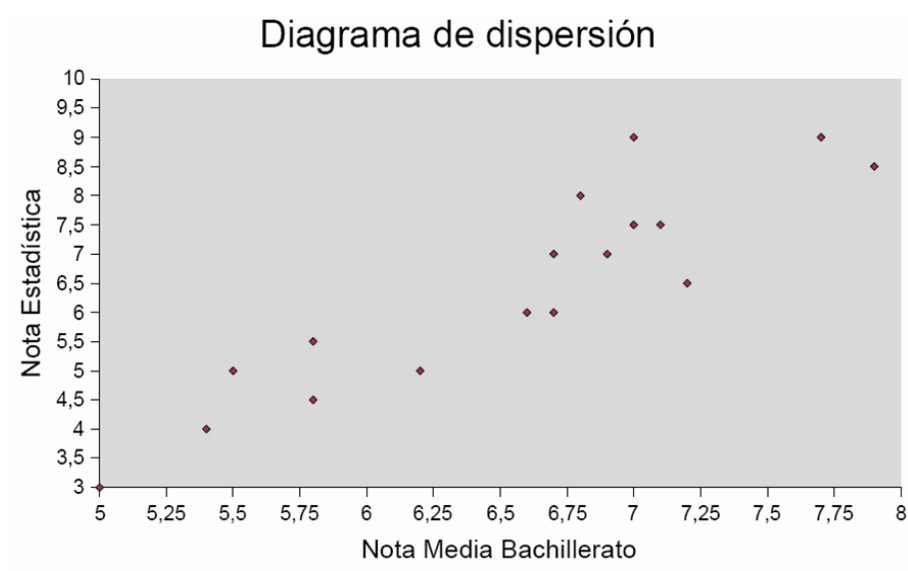


Figura 1.3: Diagrama de dispersión para los datos de la tabla 1.3



Cuadro 1.3:

	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



requiere conocimientos previos aprendidos 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.

### 1.3. 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.

#### 1.3.1. 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 1.1 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 dependientes 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  $\chi^2$  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 1.1:



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

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

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

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.

### 1.3.2. 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 de 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):



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

(covarianza poblacional)

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

(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  $N = 19 + 1 + 1 + \dots + 10 + 3 = 100$ .

2. 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:

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

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

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

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

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

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

3. Finalmente calculamos la covarianza (muestral) y la correlación:

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

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

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 1.4).

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



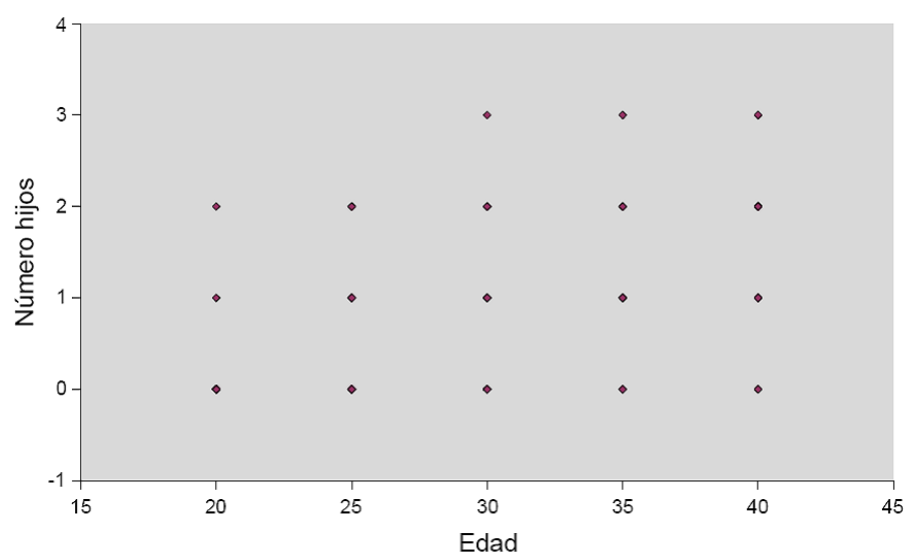


Figura 1.4: Diagrama de dispersión Edad-Número de hijos. El coeficiente de correlación es 0,54

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

$T \setminus L$	$[0 - 10)$	$[10 - 20)$	$[20 - 30)$	$[30 - 40)$	$[40 - 50)$	$[50 - 60)$
$[0 - 5)$	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).

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

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

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

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

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

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

$$\text{Cov} = \frac{0 \cdot (2,5 - 18,5) \cdot (5 - 29) + 0 \cdot (7,5 - 18,5) \cdot (15 - 29) + \dots + 0 \cdot (27,5 - 18,5) \cdot (55 - 29)}{39} = -119,49$$

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

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 1.5.

### 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:

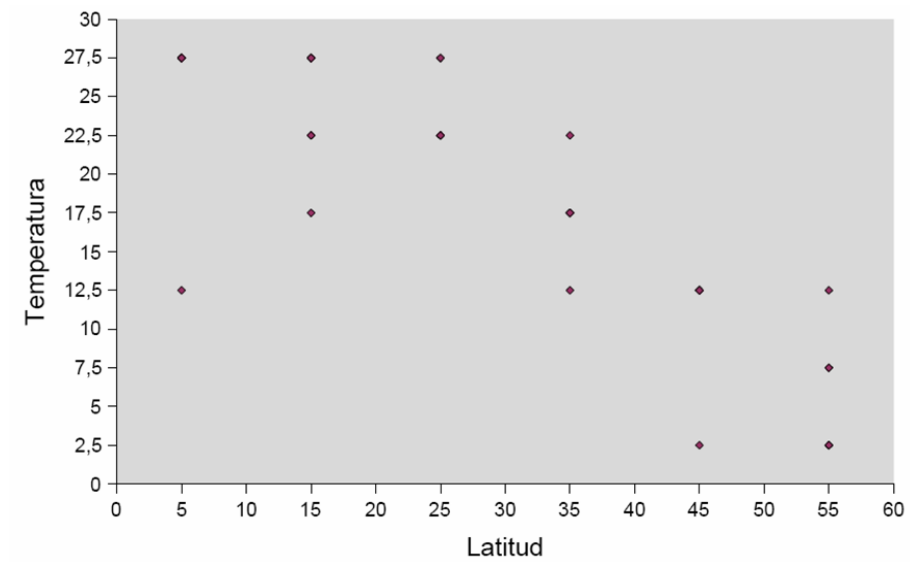


Figura 1.5: Diagrama de dispersión Latitud-Temperatura. El coeficiente de correlación es  $-0,85$

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

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

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 1.2 (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 1.3 obtendríamos los siguientes valores:

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

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

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

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

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

$$s_Y = \sqrt{2,2} = 1,483$$

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

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

### Interpretación del coeficiente de correlación

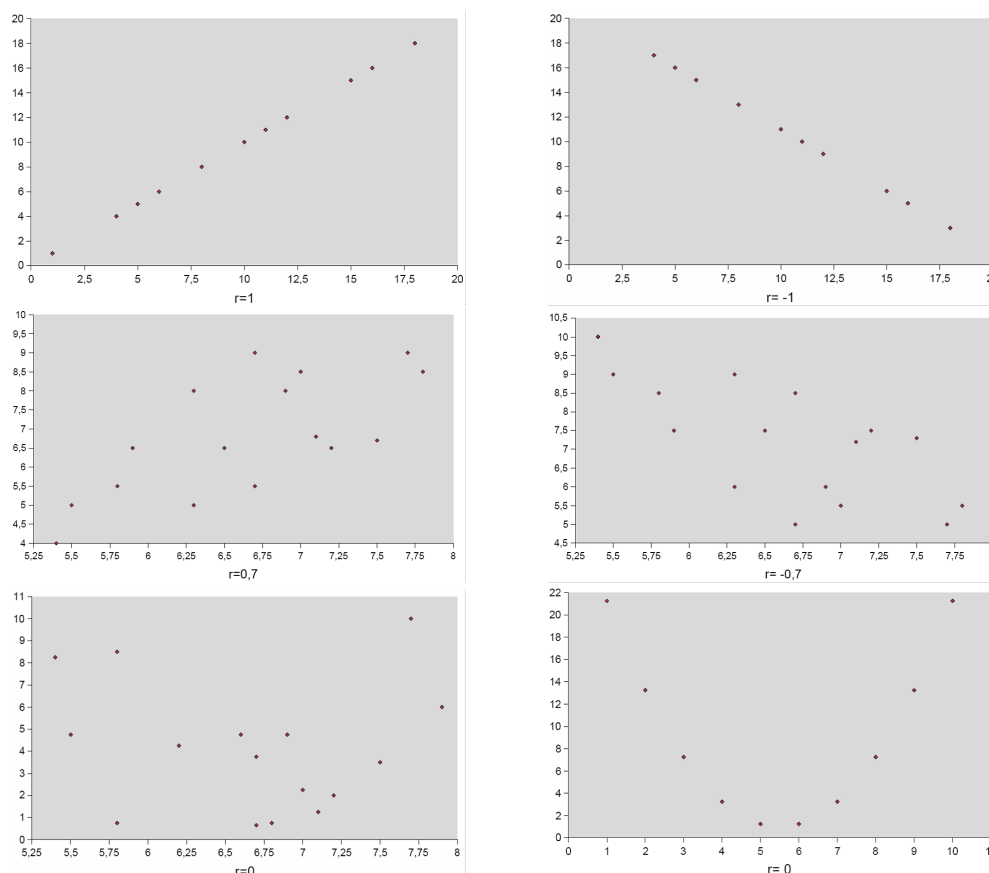


Figura 1.6: Diagramas de dispersión y correspondientes coeficientes de correlación

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

En la figura 1.6 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 1.6, 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.

## 1.4. Regresión lineal y predicción

Consideremos el diagrama de dispersión de la figura 1.3, correspondiente a la tabla de valores 1.3, 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 1.7 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 1.7). En este caso el valor es 6,29.

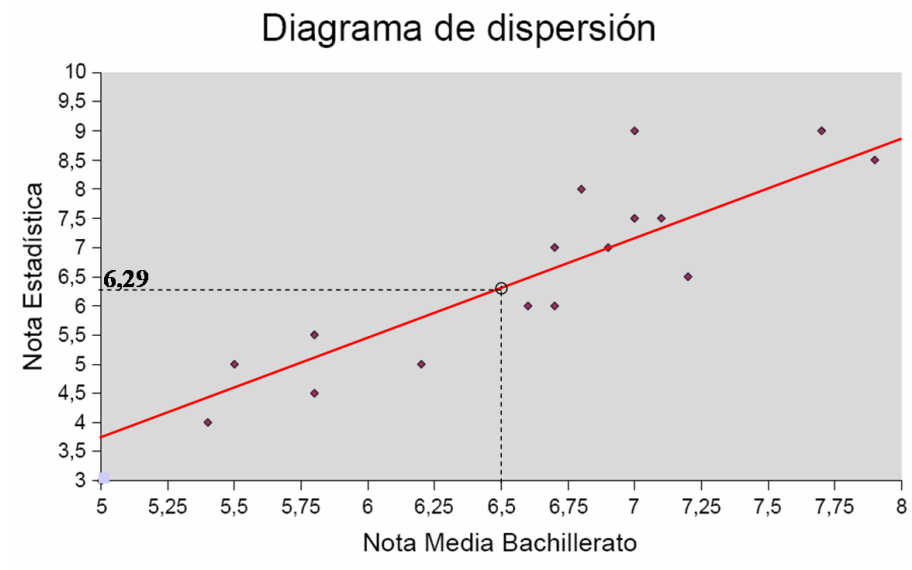


Figura 1.7: Diagrama de dispersión para la tabla 1.3 y recta de regresión

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  $\hat{Y} = 1,63 \cdot 6,5 - 4,3 = 6,29$ .

## 1.5. Análisis bivalente 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: n° de condenados				
	Reincidente		No reincidente	
	Varón	Mujer	Varón	Mujer
Total	26.771	1.352	85.230	8.625
<b>Notas:</b>				
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 1.1.

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

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

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

- a) nos situamos en la casilla D2 y escribimos `=SUMA(B2:C2)`. Al pulsar *Enter* obtenemos el valor  $n_{1\bullet} = 28123$ .





- b) el cálculo para las demás casillas de la columna se hace automáticamente situándonos con el cursor en la esquina inferior derecha de la casilla  $D2$ , pulsando el botón izquierdo del ratón y arrastrando el cursor hasta la casilla  $D3$ . Obtenemos:  $n_{2\bullet} = 93855$ .

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

- 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:

- a) escribimos la fórmula  $=B\$4*\$D2/\$D\$4$  en la casilla  $B6$
- b) a partir de la esquina inferior derecha de  $B6$  extendemos el cálculo a  $C6$
- c) 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  $= \text{SUMA}(B9:C10)$  y al pulsamos *Enter*. Obtenemos  $\chi^2 = 553,32$ .
- El coeficiente C de contingencia se calcula aplicando la fórmula de la sección 1.3.1: escribimos  $= \text{RAÍZ}(C12/(D4+C12))$  en C13, pulsamos *Enter* y obtenemos  $C = 0,07$ .
- Para decidir si este valor es alto o bajo debemos calcular el valor máximo de C, según la fórmula de la sección 1.3.1: como ambas variables constan de dos únicos valores,  $\min\{k, l\} = 2$ , por tanto escribimos  $= \text{RAÍZ}(1-1/2)$  en C14. Al pulsar *Enter* obtenemos  $C_{max} = 0,71$ .
- 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 contingen.	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 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: Consellería 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 de la sección 1.4. Para utilizar la fórmula debemos calcular:

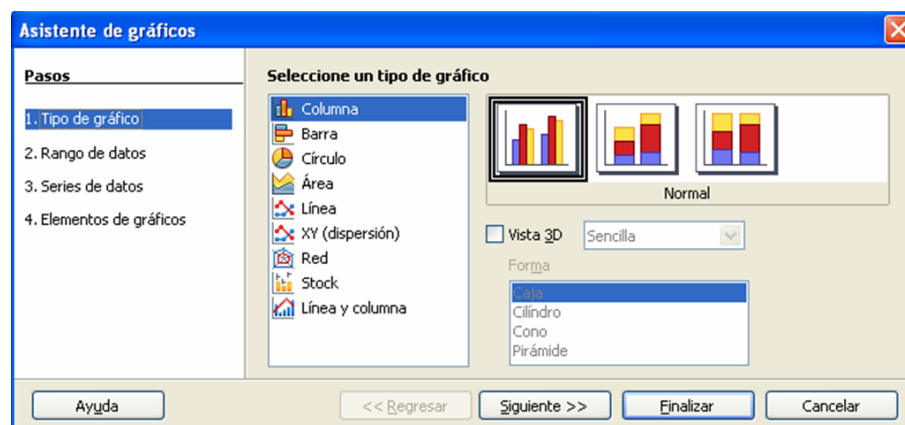
1. la covarianza ( $-1966,63$ , calculada en el ejemplo anterior),
2. la varianza de la primera variable (fórmula  $=\text{VARP}(A2:A14)$ , resultado  $16272,15$ ),
3. las medias de cada variable (fórmulas  $=\text{PROMEDIO}(A2:A14)$  y  $=\text{PROMEDIO}(B2:B14)$ , respectivamente, resultados  $427,36$  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  **Gráfico...** del menú *Insertar* y a continuación sobre una casilla cualquiera para insertar el gráfico en esa posición. Aparece el siguiente cuadro de diálogo:



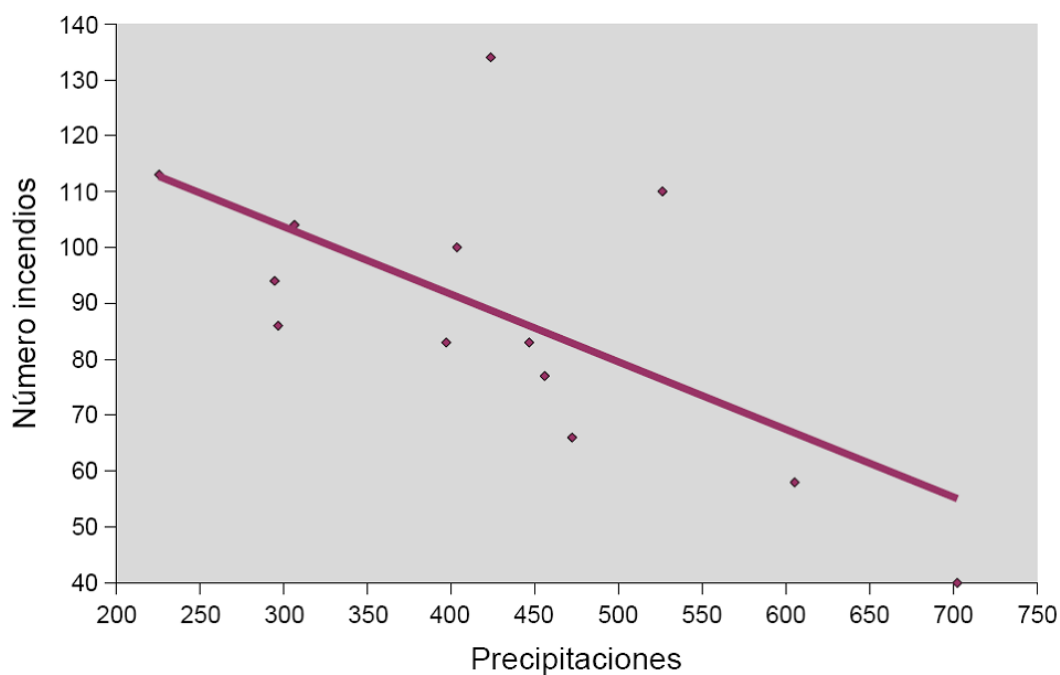
3. Seleccionamos la opción  **XY (dispersión)** y la opción *Sólo puntos*.
4. En el rango de datos escribimos A1:B14 y pulsamos el botón *Siguiente*.
5. En el diálogo *Series de datos* hacemos clic sobre *Valores X* y escribimos A2:A14 en *Rango para valores X*. Repetimos el proceso para los valores 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:

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



## 1.6. 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	Ambos sexos	Ambos sexos	Ambos sexos	Ambos sexos	Ambos sexos	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.