

Conceptos básicos de estadística bidimensional

DISTRIBUCIÓN DE FRECUENCIAS BIDIMENSIONAL

Supongamos que estamos interesados en observar dos caracteres sobre cada individuo de la población objeto de un cierto estudio. Tendremos entonces un par de números (x, y) asociados a cada individuo de la población, lo que dará lugar a una **variable estadística bidimensional** (X, Y) . Así, una distribución de frecuencias bidimensional es el conjunto de valores (x, y) de la variable estadística (X, Y) junto con las correspondientes frecuencias (absolutas o relativas).

Tablas de frecuencias

Según la naturaleza de la población en la que se observan los caracteres X e Y , se emplean dos tipos de tablas de frecuencias: las **tablas de datos apareados** (de entrada simple) y las **tablas de doble entrada**.

Ejemplo 1

Supongamos que se han observado las edades de cinco niños y sus pesos respectivos, habiéndose obtenido los siguientes datos:

(2, 10) (4, 18) (6, 25) (7, 33) (8, 34)

Llamando X a la variable “edad”, medida en años, e Y a la variable “peso”, medida en kg, podemos organizar los datos en una tabla de datos apareados o tabla de entrada simple como la siguiente:

x_i	y_i	f_i
2	10	1
4	18	1
6	25	1
7	33	1
8	34	1
		$n = 5$

Está claro que al ser 1 la frecuencia absoluta de cada par ordenado (x_i, y_i) , podríamos prescindir de esta columna y escribir la tabla en un par de filas (ahora adquiere más sentido si cabe lo de datos apareados):

x_i	2	4	6	7	8
y_i	10	18	25	33	34

Ejemplo 2

Supongamos que se han sometido 50 alumnos a una prueba formada por dos test A y B, que puntúan de 1 a 3, y que se han obtenido las siguientes puntuaciones (el primer número es la puntuación del test A y el segundo el del test B):

(1, 1) (2, 1) (2, 2) (1, 3) (2, 1) (2, 2) (1, 3) (2, 2) (3, 1) (3, 3)
 (2, 3) (1, 1) (3, 3) (2, 1) (3, 1) (1, 2) (1, 1) (2, 3) (3, 2) (3, 3)
 (2, 2) (2, 2) (1, 2) (2, 2) (1, 1) (2, 3) (2, 2) (1, 1) (2, 2) (3, 1)
 (1, 2) (2, 3) (1, 3) (2, 1) (2, 2) (2, 1) (1, 1) (2, 2) (2, 3) (1, 1)
 (3, 3) (2, 2) (2, 1) (3, 3) (1, 1) (1, 2) (2, 1) (1, 1) (3, 3) (2, 2)

Llamemos X a la puntuación obtenida en el test A e Y a la obtenida en el test B. Después de hacer un recuento, podemos agrupar los datos en una tabla de entrada simple (datos apareados) o bien en una tabla de doble entrada, como las siguientes:

x_i	y_i	f_i
1	1	9
1	2	4
1	3	3
2	1	7
2	2	12
2	3	5
3	1	3
3	2	1
3	3	6
		$n = 50$

$X \backslash Y$	$y_1 = 1$	$y_2 = 2$	$y_3 = 3$	Total
$x_1 = 1$	9	4	3	16
$x_2 = 2$	7	12	5	24
$x_3 = 3$	3	1	6	10
Total	19	17	14	$n = 50$

La estructura general de estas tablas es:

Tabla simple		
x_i	y_i	f_i
x_1	y_1	f_1
x_2	y_2	f_2
...	...	...
x_i	y_i	f_i
...	...	...
x_k	y_k	f_k
		n

Tablas de doble entrada						
$X \backslash Y$	y_1	...	y_j	...	y_q	$f_{i\bullet}$
x_1	f_{11}	...	f_{1j}	...	f_{1q}	$f_{1\bullet}$
...	...	...	...	...	...	...
x_i	f_{i1}	...	f_{ij}	...	f_{iq}	$f_{i\bullet}$
...	...	...	...	...	...	...
x_p	f_{p1}	...	f_{pj}	...	f_{pq}	$f_{p\bullet}$
$f_{\bullet j}$	$f_{\bullet 1}$	...	$f_{\bullet j}$	...	$f_{\bullet q}$	$f_{\bullet\bullet} = n$

donde f_i es la frecuencia absoluta del par (x_i, y_i) , y f_{ij} es la frecuencia absoluta del par (x_i, y_j) . Además:

- ✓ $f_{i\bullet} = \sum_{j=1}^q f_{ij} = f_{i1} + f_{i2} + \dots + f_{iq}$, o sea, el número de individuos con la modalidad x_i de la variable X , y todas las de la variable Y . A esta frecuencia se le llama frecuencia absoluta de x_i independientemente de Y .
- ✓ $f_{\bullet j} = \sum_{i=1}^p f_{ij} = f_{1j} + f_{2j} + \dots + f_{pj}$, o sea, el número de individuos con la modalidad y_j de la variable Y , y todas las de la variable X . A esta frecuencia se le llama frecuencia absoluta de y_j independientemente de X .
- ✓ $f_{\bullet\bullet} = \sum_{i=1}^p f_{i\bullet} = \sum_{j=1}^q f_{\bullet j} = \sum_i \sum_j f_{ij} = n$ (total de individuos de la población).

Distribuciones marginales

Se llama **distribución marginal** de X (respectivamente, de Y) a la distribución de la variable X (respectivamente, Y) independientemente de los valores que toma la variable X (respectivamente, Y).

Dependiendo de cómo venga dada la distribución conjunta, por la tabla de datos apareados (tabla simple) o por la tabla de doble entrada, las distribuciones de frecuencias absolutas de X y de Y son:

Datos apareados				Doble entrada			
Distribución marginal de X		Distribución marginal de Y		Distribución marginal de X		Distribución marginal de Y	
x_i	f_i	y_j	f_j	x_i	$f_{i\bullet}$	y_j	$f_{\bullet j}$
x_1	f_1	y_1	f_1	x_1	$f_{1\bullet}$	y_1	$f_{\bullet 1}$
x_2	f_2	y_2	f_2	x_2	$f_{2\bullet}$	y_2	$f_{\bullet 2}$
...	...	...	...	...	...	...	...
x_i	f_i	y_j	f_j	x_i	$f_{i\bullet}$	y_j	$f_{\bullet j}$
...	...	...	...	...	...	...	...
x_k	f_k	y_k	f_k	x_p	$f_{p\bullet}$	y_q	$f_{\bullet q}$
	n		n		$f_{\bullet\bullet} = n$		$f_{\bullet\bullet} = n$

Se trata de dos distribuciones unidimensionales como las estudiadas anteriormente. Para cada una de ellas se calculan todos los estadísticos usuales en tales distribuciones. En particular, las **medias marginales**, las **varianzas marginales** y las **desviaciones típicas marginales** son:

		Datos apareados	Doble entrada
Medias marginales	Variable X	$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^p x_i f_{i\bullet}}{n}$
	Variable Y	$\bar{y} = \frac{\sum_{j=1}^k y_j f_j}{n}$	$\bar{y} = \frac{\sum_{j=1}^q y_j f_{\bullet j}}{n}$
Varianzas marginales	Variable X	$\sigma_X^2 = \frac{\sum_{i=1}^k x_i^2 f_i}{n} - \bar{x}^2$	$\sigma_X^2 = \frac{\sum_{i=1}^p x_i^2 f_{i\bullet}}{n} - \bar{x}^2$
	Variable Y	$\sigma_Y^2 = \frac{\sum_{j=1}^k y_j^2 f_j}{n} - \bar{y}^2$	$\sigma_Y^2 = \frac{\sum_{j=1}^q y_j^2 f_{\bullet j}}{n} - \bar{y}^2$
Desviaciones típicas marginales	Variable X	$\sigma_X = +\sqrt{\sigma_X^2}$	
	Variable Y	$\sigma_Y = +\sqrt{\sigma_Y^2}$	

Covarianza

En las distribuciones bidimensionales se emplea otro estadístico que refleja el promedio de los productos de las desviaciones de cada una de las variables respecto a sus respectivas medias. Se llama **covarianza** de X e Y :

$$\sigma_{XY} = \overline{(X - \bar{X})(Y - \bar{Y})}$$

Cuando los datos vienen dados en una tabla simple (datos apareados) tenemos:

$$\sigma_{XY} = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y}) f_i}{n} = \frac{\sum_{i=1}^k x_i y_i f_i}{n} - \bar{x} \bar{y}$$

Cuando los datos vienen dados en una tabla de doble entrada se tiene:

$$\sigma_{XY} = \frac{\sum_{i=1}^p \sum_{j=1}^q (x_i - \bar{x})(y_j - \bar{y}) f_{ij}}{n} = \frac{\sum_{i=1}^p \sum_{j=1}^q x_i y_j f_{ij}}{n} - \bar{x} \bar{y}$$

En ambos casos:

$$\sigma_{XY} = \overline{(x \cdot y)} - \bar{x} \cdot \bar{y}$$

Para el ejemplo 2 de la primera página, tendríamos

x_i	y_i	f_i	$x_i f_i$	x_i^2	$x_i^2 f_i$	$y_i f_i$	y_i^2	$y_i^2 f_i$	$x_i y_i$	$x_i y_i f_i$
1	1	9	9	1	9	9	1	9	1	9
1	2	4	4	1	4	8	4	16	2	8
1	3	3	3	1	3	9	9	27	3	9
2	1	7	14	4	28	7	1	7	2	14
2	2	12	24	4	48	24	4	48	4	48
2	3	5	10	4	20	15	9	45	6	30
3	1	3	9	9	27	3	1	3	3	9
3	2	1	3	9	9	2	4	4	6	6
3	3	6	18	9	54	18	9	54	9	54
		$n = 50$	94		202	95		213		187

Para el mismo ejemplo, si organizamos los datos en una tabla de doble entrada, tendremos

$X \backslash Y$	1	2	3	$f_{i\cdot}$	$x_i f_{i\cdot}$	x_i^2	$x_i^2 f_{i\cdot}$	$\sum x_i y_j f_{ij}$
1	9	4	3	16	16	1	16	26
2	7	12	5	24	48	4	96	92
3	3	1	6	10	30	9	90	69
$f_{\cdot j}$	19	17	14	$n = 50$	94		202	187
$y_j f_{\cdot j}$	19	34	42	95	$\sum x_i y_j f_{ij}$			
y_j^2	1	4	9					
$y_j^2 f_{\cdot j}$	19	68	126	213				
$\sum x_i y_j f_{ij}$	32	62	93	187				

Las medias, varianzas, desviaciones típicas y covarianzas resultan:

$$\bar{x} = \frac{94}{50} = 1,88$$

$$\bar{y} = \frac{95}{90} = 1,9$$

$$\sigma_x^2 = \frac{202}{50} - 1,88^2 = 0,51$$

$$\sigma_y^2 = \frac{213}{50} - 1,9^2 = 0,65$$

$$\sigma_x = +\sqrt{\sigma_x^2} = +\sqrt{0,51} = 0,71$$

$$\sigma_y = +\sqrt{\sigma_y^2} = +\sqrt{0,65} = 0,81$$

$$\sigma_{xy} = \frac{187}{50} - 1,88 \cdot 1,9 = 0,168$$

PREDICCIÓN BIDIMENSIONAL: REGRESIÓN

La velocidad que alcanza una piedra que dejamos caer y el tiempo transcurrido desde que la soltamos están relacionadas por la expresión $v = g \cdot t$, donde $g \cong 9,81 \text{ m/s}^2$ es la aceleración de la gravedad. Entre las variables t y v existe una **relación funcional** que permite calcular exactamente el valor de una de ellas conocida la otra.

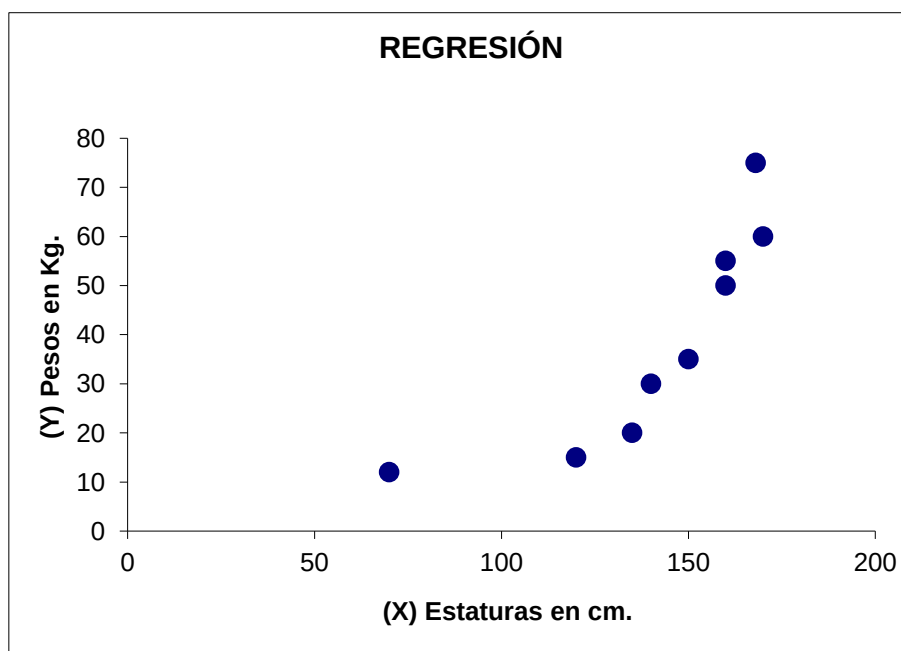
En un conjunto de personas a las que pesamos y tallamos se comprueba que, aproximadamente, a más estatura corresponde mayor peso. Pero no hay ninguna función matemática que ligue a esas dos variables. Se trata de una **relación estadística**.

La **regresión** consiste en la búsqueda de una función que exprese lo mejor posible la relación existente entre dos variables estadísticas. Esto permitirá predecir el valor de una de las variables para un valor dado de la otra.

Ejemplo 3

En la tabla siguiente se dan los pesos y estaturas de los nueve miembros de una familia:

Estatura (cm)	X	168	170	160	160	150	140	135	120	70
Peso (kg)	Y	75	60	55	50	35	30	20	15	12



La nube de puntos correspondiente a la variable bidimensional (X, Y) se ajusta bastante bien, aunque no exactamente, a una recta en las edades adultas. En cambio el “ajuste” en los niños no es tan bueno.

La búsqueda de “buenas” rectas de ajuste es el objetivo de la **regresión lineal**.

Recta de regresión de Y sobre X

Se trata de hallar la ecuación de la recta r que mejor se ajuste a la nube de puntos en el siguiente sentido:

- r pasa por el punto $G(\bar{x}, \bar{y})$ (**centro de gravedad** de la distribución).
- La suma de los cuadrados de las diferencias de ordenadas entre los puntos-nube y los puntos-recta r , para cada valor x_i de X , $S_{Y/X} = \sum d_i^2 = \sum (y_{in} - y_{ir})^2$, ha de ser mínima.

La **recta de regresión de Y sobre X** tiene por ecuación:

$$r_{Y/X} : y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2} (x - \bar{x})$$

Se emplea para predecir el valor de Y para un valor de X . Para un individuo que mida x_o cm, predice un peso de

$$y_o = \bar{y} + \frac{\sigma_{XY}}{\sigma_X^2} (x_o - \bar{x}) \text{ kilogramos.}$$

La *pendiente* de la recta $r_{Y/X}$, $m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2}$, se llama **coeficiente de regresión de Y sobre X** .

Recta de regresión de X sobre Y

Se trata de hallar la ecuación de la recta r que mejor se ajuste a la nube de puntos en el siguiente sentido:

- r pasa por el punto $G(\bar{x}, \bar{y})$ (**centro de gravedad** de la distribución).
- La suma de los cuadrados de las diferencias de ordenadas entre los puntos-nube y los puntos-recta r , para cada valor y_j de Y , $S_{X/Y} = \sum D_i^2 = \sum (x_{in} - x_{ir})^2$, ha de ser mínima.

La **recta de regresión de X sobre Y** tiene por ecuación:

$$r_{X/Y} : x - \bar{x} = \frac{\sigma_{XY}}{\sigma_Y^2} (y - \bar{y})$$

Se emplea para predecir el valor de X para un valor de Y . Para un individuo que pese y_o kilogramos, predice una

$$x_o = \bar{x} + \frac{\sigma_{XY}}{\sigma_Y^2} (y_o - \bar{y}) \text{ cm.}$$

La *pendiente* de la recta $r_{X/Y}$, $m_{X/Y} = \frac{\sigma_{XY}}{\sigma_Y^2}$, se llama **coeficiente de regresión de X sobre Y** .

Observación

La nube de puntos presentada es de tal forma que sugiere la idea de proceder a una aproximación mediante una recta (*ajuste lineal*). Otras veces sería más adecuado un ajuste parabólico, exponencial u otros.

las rectas de regresión de Y sobre X y de X sobre Y se pueden obtener siempre, cualquiera que sea la forma de la nube de puntos (otra cosa es que el ajuste sea más o menos adecuado).

Para el ejemplo 3 se tendría (observa que como las frecuencias absolutas son iguales a 1, prescindimos de las columnas donde aparecen tales frecuencias absolutas y no es necesaria una tabla de doble entrada, más cómodas cuando algunas de las frecuencias absolutas son distintas de 1):

Talla (cm) x_i	Peso (kg) y_i	x_i^2	y_i^2	$x_i y_i$
168	75	28224	5625	12600
170	60	28900	3600	10200
160	55	25600	3025	8800
160	50	25600	2500	8000
150	35	22500	1225	5250
140	30	19600	900	4200
135	20	18225	400	2700
120	15	14400	225	1800
70	12	4900	144	840
1273	352	187949	17644	54390

Las medias, varianzas, desviaciones típicas y covarianza son las siguientes:

$$\bar{x} = \frac{1273}{9} = 141,44$$

$$\bar{y} = \frac{352}{9} = 39,11$$

$$\sigma_x^2 = \frac{187949}{9} - 141,44^2 = 877,95$$

$$\sigma_y^2 = \frac{17644}{9} - 39,11^2 = 430,85$$

$$\sigma_x = +\sqrt{\sigma_x^2} = +\sqrt{877,95} = 29,63$$

$$\sigma_y = +\sqrt{\sigma_y^2} = +\sqrt{430,85} = 20,76$$

$$\sigma_{xy} = \frac{54390}{9} - 141,44 \cdot 39,11 = 511,61$$

Los coeficientes de regresión son:

$$m_{Y/X} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{511,61}{877,95} = 0,58$$

$$m_{X/Y} = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{511,61}{430,85} = 1,19$$

La recta de regresión de Y sobre X ($r_{Y/X}$) tiene por ecuación:

$$y - 39,11 = 0,58(x - 141,44) \Rightarrow y = 0,58x - 42,93$$

Se emplea para predecir el valor de Y para un valor de X . Por ejemplo, para un individuo que mida $x_o = 120$ cm predice un peso de $y_o = 0,58 \cdot 120 - 42,93 = 26,67$ kg.

La recta de regresión de X sobre Y ($r_{X/Y}$) tiene por ecuación:

$$x - 141,44 = 1,19(y - 39,11) \Rightarrow x = 1,19y + 94,9$$

Se emplea para predecir el valor de X para un valor de Y . Por ejemplo, para un individuo que pese $y_o = 35$ kg predice una talla de $x_o = 1,19 \cdot 35 + 94,9 = 136,55$ cm.

Ahora podrías calcular las rectas de regresión y alguna que otra predicción utilizando los datos del ejemplo 2, donde se sometía a 50 alumnos a una prueba formada por dos test, A y B. En este caso, dada la puntuación de un alumno en el test A, se podría predecir su puntuación en el test B.

CORRELACIÓN LINEAL

Correlación lineal es un término estadístico que pretende medir el grado de ajuste existente entre la recta de regresión y la nube de puntos (las predicciones serán buenas cuando cuanto más correlación haya entre las variables).

Coeficiente de correlación lineal de Pearson

Una medida de la bondad de este ajuste es el **coeficiente de correlación de Pearson**, definido como:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

En general, a los cambios producidos en una variable X , pueden acompañar cambios en la otra variable Y . Incrementos en los valores de X pueden reflejar modificaciones en los correspondientes valores de Y en forma de aumentos o disminuciones.

Recordemos que los coeficientes de regresión (pendientes de las rectas de regresión) son:

$$m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} \qquad m_{X/Y} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

Dos variables están *correlacionadas positivamente (correlación directa)* si las dos crecen o disminuyen en el mismo sentido: al crecimiento (decrecimiento) de una de ellas le acompaña el crecimiento (decrecimiento) de la otra. En este caso, las pendientes de las rectas de regresión son ambas positivas y, por tanto: $\sigma_{XY} > 0 \Rightarrow \rho > 0$

Dos variables están *correlacionadas negativamente (correlación inversa)* si las dos crecen o disminuyen en sentido contrario: al crecimiento (decrecimiento) de una de ellas le acompaña el decrecimiento (crecimiento) de la otra. En este caso, las pendientes de las rectas de regresión son ambas negativas y, por tanto: $\sigma_{XY} < 0 \Rightarrow \rho < 0$

Se dice que dos variables son *incorreladas (correlación nula)* cuando no existe ninguna correlación entre ellas. Esto ocurre cuando $\sigma_{XY} = 0 \Rightarrow \rho = 0$; o bien cuando la covarianza es muy próxima a cero: $\sigma_{XY} \approx 0$.

Relación entre el coeficiente de correlación ρ y los coeficientes de regresión

Observemos que $\rho^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{\sigma_{XY}}{\sigma_X^2} \cdot \frac{\sigma_{XY}}{\sigma_Y^2} = m_{Y/X} \cdot m_{X/Y}$, de donde

$$\rho = \pm \sqrt{m_{Y/X} \cdot m_{X/Y}}$$

(se tomará el mismo signo que el de la covarianza, tal y como se ha visto anteriormente).

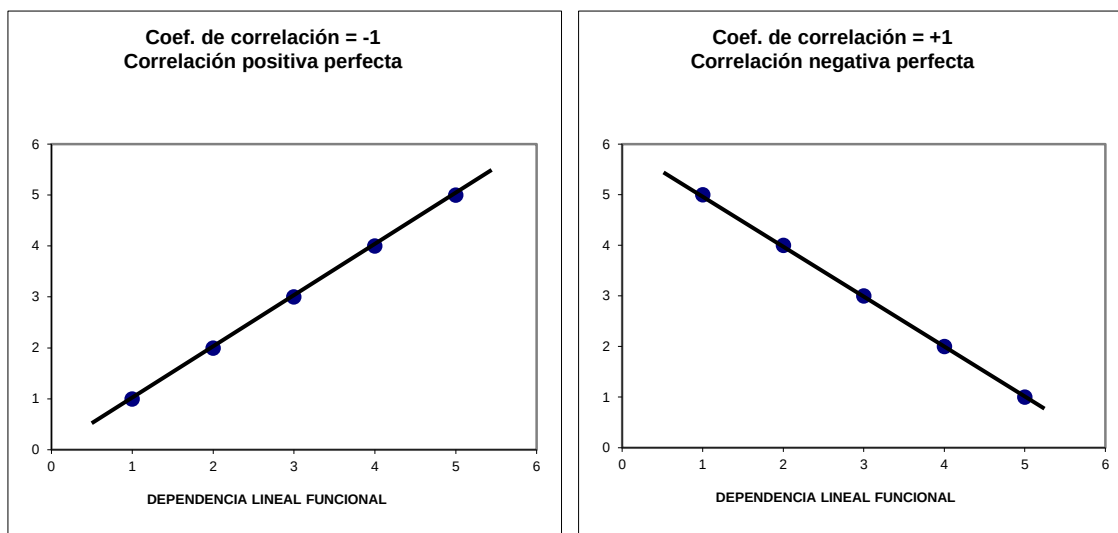
Interpretación de los posibles valores del coeficiente de correlación ρ y significado de la posición relativa de las rectas de regresión

Es posible demostrar que el cuadrado del coeficiente de correlación es menor o igual que 1, es decir $\rho^2 \leq 1$, y que por tanto $-1 \leq \rho \leq 1$.

✓ Si $\rho^2 = 1$ (es decir, si $\rho = 1$ ó $\rho = -1$), entonces todos los puntos del diagrama de dispersión o nube de puntos se encuentran sobre la recta de regresión de Y sobre X , o sea, que están alineados y por tanto el ajuste es perfecto: hay **dependencia funcional lineal**. Además $\rho^2 = m_{Y/X} \cdot m_{X/Y} = 1 \Rightarrow m_{Y/X} = \frac{1}{m_{X/Y}}$, luego las dos rectas

de regresión tienen la misma pendiente. Al pasar ambas por el centro de gravedad $G(\bar{x}, \bar{y})$, entonces coinciden.

- Si $\rho = +1$ hay correlación positiva perfecta (dependencia funcional).
- Si $\rho = -1$ hay correlación negativa perfecta (dependencia funcional).



✓ Si $\rho^2 \approx 1$ (es decir si $\rho \approx -1$ ó $\rho \approx 1$), entonces el ajuste es muy bueno y las predicciones que se hagan mediante las rectas de regresión serán muy fiables. Además $\rho^2 = m_{Y/X} \cdot m_{X/Y} \approx 1 \Rightarrow m_{Y/X} \approx \frac{1}{m_{X/Y}}$, luego las dos rectas de regresión tienen pendientes muy próximas.

- Si $\rho \approx +1$ hay una fuerte correlación positiva.
- Si $\rho \approx -1$ hay una fuerte correlación negativa.

✓ Si $\rho^2 = 0$ ($\rho = 0$), no existe correlación entre X e Y . La pendiente de la recta de regresión de Y sobre X es

$$m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\rho \sigma_X \sigma_Y}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X} = 0, \text{ luego } r_{Y/X} : y = \bar{y} \text{ (recta horizontal). De manera similar}$$

$$m_{X/Y} = \frac{\sigma_{XY}}{\sigma_Y^2} = \frac{\rho \sigma_X \sigma_Y}{\sigma_Y^2} = \rho \frac{\sigma_X}{\sigma_Y} = 0, \text{ luego } r_{X/Y} : x = \bar{x} \text{ (recta vertical). Hemos demostrado que, en este caso,}$$

las dos rectas son perpendiculares.

✓ Si $\rho^2 \approx 0$ ($\rho \approx 0$), el ajuste es malo. La pendiente de la recta de regresión de Y sobre X es

$$m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\rho \sigma_X \sigma_Y}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X} \approx 0 \text{ (es casi horizontal), mientras que la pendiente de la recta de regresión de}$$

$$X \text{ sobre } Y, \frac{1}{m_{X/Y}}, \text{ toma valores muy grandes en valor absoluto (es casi vertical), es decir, las rectas de regresión}$$

están muy separadas.

En el ejemplo 2 se tenía que $\sigma_X = 0,71$, $\sigma_Y = 0,81$, $\sigma_{XY} = 0,168$. Entonces $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0,168}{0,71 \cdot 0,81} = 0,29$. Como

el coeficiente de correlación es próximo a cero, la correlación entre las dos variables es escasa.

Analiza tú la correlación para el ejemplo 3.

Ejercicios y problemas

1. De un estudio sobre la influencia del consumo de tabaco en el índice de mortalidad de la población, se obtuvieron los siguientes datos:

Número de cigarrillos al día	3	5	6	15	20	40
Índice de mortalidad	0,2	0,3	0,3	0,5	0,7	1,7

Estudiar la correlación entre ambas variables e interpretarla. Hallar la recta de regresión que permite predecir la mortalidad conocido el consumo de cigarrillos. ¿Qué mortalidad se puede esperar para un consumidor de 30 cigarrillos diarios?

2. Las estaturas y los pesos, en cm y kg, de un grupo de seis personas vienen recogidas en la tabla:

Estaturas	168	174	180	175	158	162
Pesos	65	70	73	68	55	62

Si una persona pesa 71 kg, ¿qué estatura se le puede predecir? Si una persona mide 183 cm, ¿qué peso se le puede predecir? ¿Cómo de fiables son las predicciones? Justificar la respuesta.

3. Los cuarenta alumnos de un curso realizan dos pruebas: una de Matemáticas, X , en la que se puntúa de 1 a 5, y otra de Redacción, Y , en la que se puntúa de 1 a 3. Los resultados se dan en la siguiente tabla

$X \backslash Y$	3	2	1
1	1	4	3
2	1	3	6
3	3	5	2
4	4	2	2
5	3	1	0

Hallar las rectas de regresión y el coeficiente de correlación lineal de Pearson. Interpretar los resultados.

4. Las tallas y pesos de 24 personas vienen dadas en la siguiente tabla de doble entrada con datos agrupados:

		Pesos (Y)			
Tallas (X)	$x_i \backslash y_j$	50–55	55–60	60–65	65–70
		52,5	57,5	62,5	67,5
155–160	157,5	1	2	3	0
160–165	162,5	0	1	2	1
165–170	167,5	0	3	0	2
170–175	172,5	0	2	3	4

Calcular: a) Las medias y las desviaciones típicas marginales de talla y peso. b) Peso medio y desviación típica media condicionada a los que miden entre 16 y 170 cm. c) Talla media y desviación típica condicionada a los que pesan entre 55 y 60 kg. d) Rectas de regresión y coeficiente de correlación lineal de Pearson. Interpretación.

5. La siguiente tabla muestra el número de gérmenes patógenos (en miles por cm^3) de un determinado cultivo según el tiempo transcurrido:

Número de horas	0	1	2	3	4	5
Número de gérmenes	20	26	33	41	47	53

Hallar la recta de regresión para predecir el número de gérmenes por cm^3 en función del tiempo transcurrido. ¿Qué cantidad de gérmenes se puede predecir que habrá cuando pasen 6 horas? ¿Cómo de buena es la predicción anterior?