Tema 9: Estadística en dos variables (bidimensional)

1. Distribución de frecuencias bidimensional

En el tema anterior se han estudiado las distribuciones unidimensionales obtenidas al observar sólo un carácter de la población.

Ahora estamos interesados en observar dos caracteres sobre cada individuo de la población objeto de estudio. Si los caracteres son cuantitativos, lo que hemos llamado variables estadísticas, tendremos un par de números (x, y) por cada unidad estadística o individuo de la población. Tenemos así 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).

Las distribuciones bidimensionales, como las unidimensionales, pueden venir dadas por los valores de sus componentes agrupados en intervalos (caso continuo) o por valores aislados (caso discreto). Es más, es posible que cada uno de los caracteres en cuestión tenga distinta naturaleza.

1.1. Tablas de frecuencias

Según la naturaleza de la población en la que se observan los caracteres X e Y, se emplean dos tipos de frecuencias:

- ✓ Tablas de datos apareados (de entrada simple).
- ✓ 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	n_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), podemos prescindir de la columna de las n_i.

Ejemplo 2:

Supongamos que se han sometido 50 alumnos a una prueba formada por dos tests. 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	n_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 \ Y	**1**	**2**	**3**	**Total**
1	9	4	3	**16**
2	7	12	5	**24**
3	3	1	6	**10**
Total	**19**	**17**	**14**	**N = 50**

La estructura general de estas tablas es:

Tabla simple		
x_i	y_i	n_i
x_1	y_1	n_1
x_2	y_2	n_2
...	...	...
x_i	y_i	n_i
...	...	...
x_k	y_k	n_k
		N

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

donde:

n_i = frecuencia absoluta del par (x_i, y_i).

n_{ij} = frecuencia absoluta del par (x_i, y_j).

$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} = n_{i1} + n_{i2} + ... + n_{iq}$ = "número de individuos con $X = x_i$ (y cualquier Y)" =
= frecuencia absoluta de x_i independientemente de Y.

$n_{\bullet j} = \sum_{i=1}^{p} n_{ij} = n_{1j} + n_{2j} + ... + n_{pj}$ = "número de individuos con $Y = y_j$ (y cualquier X)" =
= frecuencia absoluta de y_j independientemente de X.

$n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i}\sum_{j} n_{ij}$ = N = Total de individuos

1.2. 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:

Distribución marginal de X	
$\mathbf{x_i}$	$\mathbf{n_i}$
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k
	N

Distribución marginal de Y	
$\mathbf{y_j}$	$\mathbf{n_j}$
y_1	n_1
y_2	n_2
...	...
y_j	n_j
...	...
y_k	n_k
	N

Distribución marginal de X	
$\mathbf{x_i}$	$\mathbf{n_{i\bullet}}$
x_1	$n_{1\bullet}$
x_2	$n_{2\bullet}$
...	...
x_i	$n_{i\bullet}$
...	...
x_p	$n_{p\bullet}$
	$\mathbf{n_{\bullet\bullet} = N}$

Distribución marginal de Y	
$\mathbf{y_j}$	$\mathbf{n_{\bullet j}}$
y_1	$n_{\bullet 1}$
y_2	$n_{\bullet 2}$
...	...
y_j	$n_{\bullet j}$
...	...
y_q	$n_{\bullet q}$
	$\mathbf{n_{\bullet\bullet} = N}$

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:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} \qquad \text{o bien} \qquad \bar{x} = \frac{\sum_{i=1}^{p} n_{i\bullet} x_i}{N}$$

$$\bar{y} = \frac{\sum_{j=1}^{k} n_j x_j}{N} \qquad \text{o bien} \qquad \bar{y} = \frac{\sum_{j=1}^{q} n_{\bullet j} y_j}{N}$$

$$\text{Var}(X) = \sigma_X^2 = \frac{\sum_{i=1}^{p} n_{i\bullet} x_i^{\,2}}{N} - \bar{x}^2 \quad ; \quad \sigma_X = +\sqrt{\text{Var}(X)} = \sqrt{\sigma_X^2}$$

$$\text{Var}(Y) = \sigma_Y^2 = \frac{\sum_{j=1}^{q} n_{\bullet j} y_j^{\,2}}{N} - \bar{y}^2 \quad ; \quad \sigma_Y = +\sqrt{\text{Var}(Y)} = \sqrt{\sigma_Y^2}$$

1.3. 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:

$$\sigma_{XY} = \text{Cov}(X, Y) = \overline{(X - \bar{X}) \cdot (Y - \bar{Y})}$$

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

$$\sigma_{XY} = \text{Cov}(X, Y) = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum_{i=1}^{k} n_i x_i y_i}{N} - \bar{x}\,\bar{y}$$

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

$$\sigma_{XY} = \text{Cov}(X, Y) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} (x_i - \bar{x})(y_j - \bar{y})}{N} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} x_i y_j}{N} - \bar{x}\,\bar{y}$$

En ambos casos:

$$\sigma_{XY} = \text{Cov}(X, Y) = \overline{(x \cdot y)} - \bar{x} \cdot \bar{y}$$

Para el ejemplo 2 (página 150), tendríamos

x_i	y_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$	$n_i y_i$	y_i^2	$n_i y_i^2$	$x_i y_i$	$n_i x_i y_i$
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 \ Y	1	2	3	$n_{i\bullet}$	$n_{i\bullet}x_i$	x_i^2	$n_{i\bullet}x_i^2$	$\Sigma n_{ij}x_iy_j$
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
$n_{\bullet j}$	19	17	14	**N = 50**	**94**		**202**	**187**
$n_{\bullet j}y_j$	19	34	42	**95**				
y_j^2	1	4	9					
$n_{\bullet j}y_j^2$	19	68	126	**213**				
$\Sigma n_{ij}x_iy_j$	32	62	93	**187**				

$\Sigma n_{ij}x_iy_j$

Las medias, varianzas, desviaciones típicas y covarianza resultan:

$$\bar{x} = \frac{94}{50} = 1{,}88 \qquad \bar{y} = \frac{95}{50} = 1{,}9$$

$$\text{Var}(X) = \frac{202}{50} - 1{,}88^2 = 0{,}5056 \qquad \text{Var}(Y) = \frac{213}{50} - 1{,}9^2 = 0{,}65$$

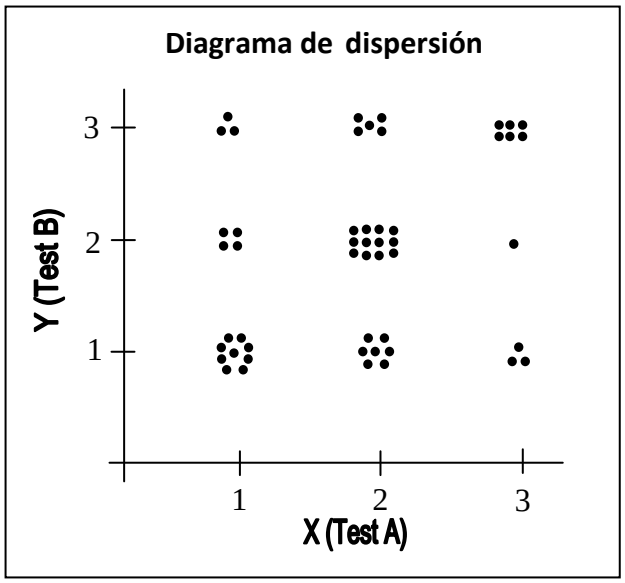
$$\sigma_X = +\sqrt{\text{Var}(X)} = +\sqrt{0{,}5056} = 0{,}711 \qquad \sigma_Y = +\sqrt{\text{Var}(Y)} = +\sqrt{0{,}65} = 0{,}806$$

$$\sigma_{XY} = \text{Cov}(X, Y) = \frac{187}{50} - 1{,}88 \cdot 1{,}9 = 0{,}168$$

2. Representaciones gráficas

2.1. Diagrama de dispersión

Como los valores de la variable estadística bidimensional (X, Y) son pares ordenados de números reales, de la forma (x_i, y_j), podemos representarlos en unos ejes cartesianos obteniendo un conjunto de puntos sobre el plano. Tal conjunto de puntos se llama *diagrama de dispersión* o *nube de puntos*.



Hay que tener en cuenta que como cada par (x_i, y_j) tiene una frecuencia absoluta n_{ij}, cuando ésta sea mayor que 1, el punto representa la concentración de n_{ij} puntos superpuestos.

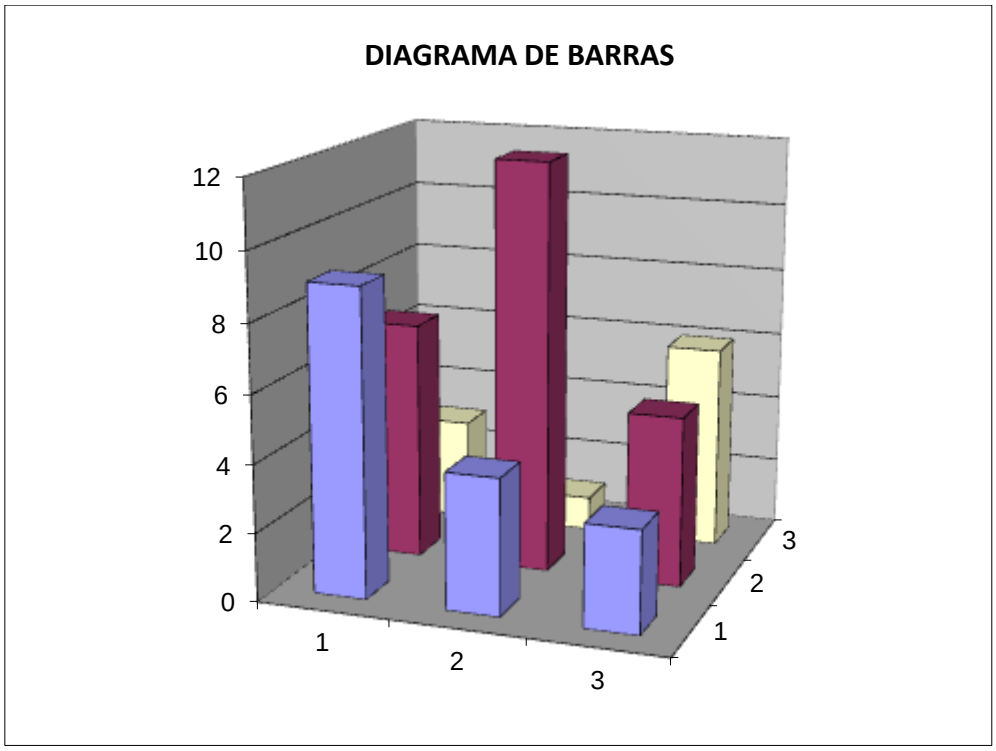
- ✓ Cuando las frecuencias no son muy altas se suelen representar

los n_{ij} puntos alrededor del punto (x_i, y_j). Observa en la figura el diagrama de dispersión del ejemplo anterior.

- ✓ A veces, lo que se hace es representar pequeños círculos cuyas áreas reflejan las frecuencias correspondientes.
- ✓ También se emplea un procedimiento mixto: se señalan con un punto los pares (x_i, y_j) con $n_{ij} = 1$, y cuando la frecuencia es mayor que 1, se escribe el valor de n_{ij}.

2.2. Diagramas de barras

Son una generalización de los diagramas de barras e histogramas de las variables unidimensionales. El diagrama de barras del ejemplo que estamos considerando es:



3. 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 $g \cong 9{,}81$ 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.

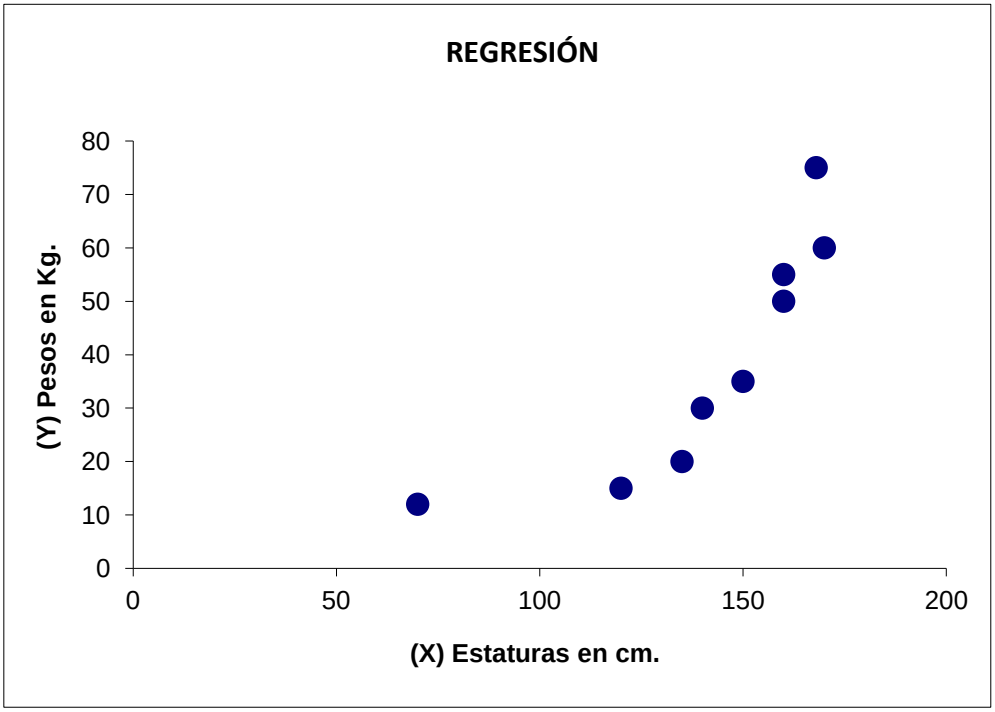
En un conjunto de personas a las que pesamos y tallamos se comprueba que, aproximadamente, a más estatura corresponde mayor peso. Pero no hay ninguna función matemática que ligue a esas dos variables. Se trata de una *relación estadística*.

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

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.

3.1. 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, $S_{Y/X} = \Sigma d_i^2 = \Sigma(y_{in} - y_{ir})^2$, 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 $y_o = \bar{y} + \frac{\sigma_{XY}}{\sigma_X^{\,2}}(x_o - \bar{x})$ kilogramos.

La pendiente de $r_{Y/X}$, $m_{Y/X} = \dfrac{\sigma_{XY}}{\sigma_X^{\,2}} = \dfrac{Cov(X,Y)}{Var(X)}$, se llama *coeficiente de regresión de* Y *sobre* X.

3.2. 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, $S_{X/Y} = \Sigma D_i^{\,2} = \Sigma(x_{in} - x_{ir})^2$, 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 estatura de $x_o = \bar{x} + \dfrac{\sigma_{XY}}{\sigma_Y^{\,2}}(y_o - \bar{y})$ cm.

$m_{X/Y} = \dfrac{\sigma_{XY}}{\sigma_Y^{\,2}} = \dfrac{Cov(X,Y)}{Var(Y)}$, es el *coeficiente de regresión de* X *sobre* Y (ahora $m_{X/Y}$ no es la pendiente de $r_{X/Y}$ sino su inversa).

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.

Sin embargo, 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).

Ejemplo 4:

Para el ejemplo anterior se tendría

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 resultan:

$$\bar{x} = \frac{1273}{9} = 141{,}44 \qquad \bar{y} = \frac{352}{9} = 39{,}11$$

$$\text{Var(X)} = \frac{187949}{9} - 141{,}44^2 = 876{,}69 \qquad \text{Var(Y)} = \frac{17644}{9} - 39{,}11^2 = 430{,}77$$

$$\sigma_X = +\sqrt{\text{Var(X)}} = +\sqrt{876{,}69} = 29{,}61 \qquad \sigma_Y = +\sqrt{\text{Var(Y)}} = +\sqrt{430{,}77} = 20{,}75$$

$$\sigma_{XY} = \text{Cov(X, Y)} = \frac{54390}{9} - 141{,}44 \cdot 39{,}11 = 511{,}28$$

Los coeficientes de regresión son:

$$m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{511{,}28}{876{,}69} = 0{,}58 \qquad m_{X/Y} = \frac{\sigma_{XY}}{\sigma_Y^2} = \frac{511{,}28}{430{,}77} = 1{,}19$$

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

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

Se emplea para predecir el valor de Y para un valor de X. 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:

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

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

Ejemplo 5:

Para el ejemplo 2 (páginas 150, 152 y 153) se tendría:

Los coeficientes de regresión de Y sobre X y de X sobre y son:

$$m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{0{,}168}{0{,}5056} = 0{,}332 \qquad m_{X/Y} = \frac{\sigma_{XY}}{\sigma_Y^2} = \frac{0{,}168}{0{,}65} = 0{,}258$$

Por tanto, la recta de regresión de Y sobre X ($r_{Y/X}$) tiene por ecuación:

$$y - 1{,}9 = 0{,}332 \cdot (x - 1{,}88) \Rightarrow y = 0{,}332x + 1{,}28$$

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

$$x - 1{,}88 = 0{,}258 \cdot (y - 1{,}9) \Rightarrow x = 0{,}258y + 1{,}39$$

4. 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.

4.1. Coeficiente de correlación lineal de Pearson

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

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

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:

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

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:

$$\sigma_{XY} = \text{Cov}(X, Y) > 0 \Rightarrow \rho > 0$$

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:

$$\sigma_{XY} = \text{Cov}(X, Y) < 0 \Rightarrow \rho < 0$$

Dos variables son *incorreladas* (*correlación nula*) cuando no existe ninguna correlación entre ellas. Esto ocurre cuando

$$\sigma_{XY} = \text{Cov}(X, Y) = 0 \Rightarrow \rho = 0$$

o bien cuando $\sigma_{XY} = \text{Cov}(X, Y)$ es muy próximo a cero: $\sigma_{XY} = \text{Cov}(X, Y) \approx 0$

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

Observemos que

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

de donde

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

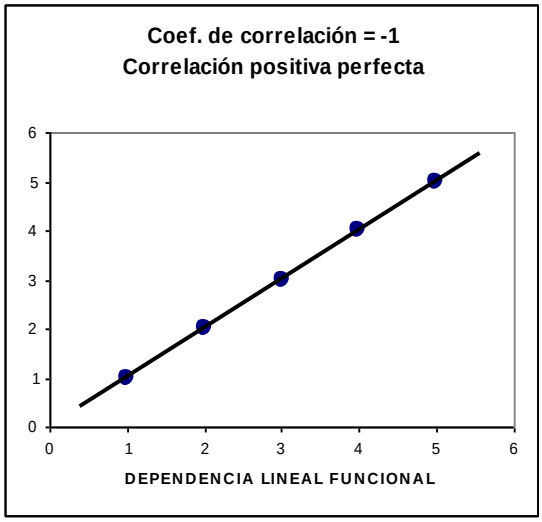
(se tomará el mismo signo que el de la covarianza, tal y como se ha visto anteriormente).

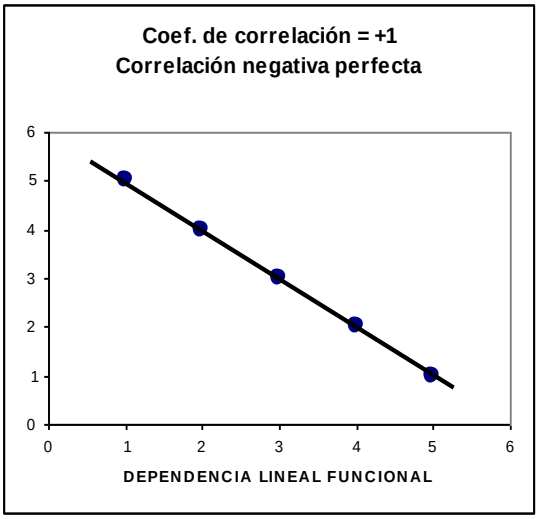
4.3. Interpretación de los posibles valores del coeficiente de correlación, ρ, y significado de la posición relativa de las rectas de regresión

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 $\rho^2 = m_{Y/X} \cdot m_{X/Y} = 1 \Rightarrow m_{Y/X} = \dfrac{1}{m_{X/Y}}$, luego las dos rectas de regresión tienen la misma pendiente. Como ambas pasan 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 $\rho^2 = m_{Y/X} \cdot m_{X/Y} \approx 1 \Rightarrow m_{Y/X} \approx \dfrac{1}{m_{X/Y}}$, 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 $\rho^2 = 0$ ($\rho = 0$), las variables X e Y son incorreladas (no existe correlación entre X e Y). La pendiente de la recta de regresión de Y sobre X es $m_{Y/X} = \dfrac{\sigma_{XY}}{\sigma_X^2} =$

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

$m_{X/Y} = \dfrac{\sigma_{XY}}{\sigma_Y^2} = \dfrac{\rho \cdot \sigma_X \cdot \sigma_Y}{\sigma_Y^2} = \rho \cdot \dfrac{\sigma_X}{\sigma_Y} = 0$, luego $r_{X/Y}$: $x = \bar{x}$ (recta vertical). Hemos demostrado que, en este caso, las dos rectas de regresión 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 $m_{Y/X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\rho \cdot \sigma_X \cdot \sigma_Y}{\sigma_X^2} = \rho \cdot \frac{\sigma_Y}{\sigma_X} \approx 0$, mientras que la pendiente de la recta de regresión de X sobre Y, $\frac{1}{m_{X/Y}}$, toma valores muy grandes en valor absoluto, es decir, las rectas de regresión están muy separadas.

Ejemplo 6:

En el ejemplo 2 (páginas 150, 152 y 153), se tenía: $\sigma_X = 0{,}711$, $\sigma_Y = 0{,}806$, $\sigma_{XY} = 0{,}168$. Entonces $\rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{0{,}168}{0{,}711 \cdot 0{,}806} = 0{,}293$. Como el coeficiente de correlación es relativamente próximo a 0, la correlación entre las dos variables es escasa.

Ejemplo 7:

Para el ejemplo 3 (página 155, talla y peso de los 9 miembros de una familia), se tenía $\sigma_X = 29{,}61$, $\sigma_Y = 20{,}75$, $\sigma_{XY} = 511{,}28$. Por tanto $\rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{511{,}28}{29{,}61 \cdot 20{,}75} = 0{,}832$.

Como el coeficiente de correlación es bastante próximo a 1, la correlación entre las dos variables es bastante fuerte y las predicciones que se realicen mediante las rectas de regresión son bastante fiables.

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

a) Estudiar la correlación entre ambas variables e interpretarla.

b) Hallar la recta de regresión que permite predecir la mortalidad conocido el consumo de cigarrillos.

c) ¿Qué mortalidad se puede esperar para un consumidor de 30 cigarrillos diarios?

2. Los valores de dos variables X e Y se distribuyen según la tabla adjunta. Determinar el coeficiente de correlación y la recta de regresión de Y sobre X. Comentar cómo de fiables son las predicciones basadas en esa recta.

X \ Y	**0**	**2**	**4**
1	2	1	3
2	1	4	2
3	2	5	0

3. Las estaturas y los pesos, en cm y Kg, respectivamente, 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

a) Si una persona pesa 71 Kg., ¿qué estatura se le puede predecir?

b) Si una persona mide 183 cm, ¿qué peso se le puede predecir?

c) ¿Cómo de fiables son las predicciones? Justificar la respuesta.

4. 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 \ 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.

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

	Pesos (Y)	**50–55**	**55–60**	**60–65**	**65–70**
Tallas (X)	x_i \ y_J	**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.

6. 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

a) Hallar la recta de regresión para predecir el número de gérmenes por cm^3 en función del tiempo transcurrido.

b) ¿Qué cantidad de gérmenes se puede predecir que habrá cuando pasen 6 horas?

c) ¿Cómo de buena es la predicción anterior?

7. La producción (en miles de quintales métricos) y la superficie (en miles de hectáreas) de lentejas en España, durante los años 1960 a 1964, según el Anuario Estadístico Español, fueron

Superficie	38	45	47	49	45
Producción	239	289	335	337	216

a) Hallar la recta de regresión de la producción de lentejas en función de la superficie cultivada.

b) Si se sabe que en un determinado año la producción fue de 420.000 quintales métricos, ¿cuál habrá sido previsiblemente la superficie cultivada durante ese año?