Ejemplos de distribuciones bidimensionales

Ejemplo 1

Una compañía de seguros considera que el número de vehículos (y) que circulan por una determinada autopista a más de 120 km/h , puede ponerse en función del número de accidentes (x) que ocurren en ella. Durante 5 días obtuvo los siguientes resultados:

Accidentes xi	5	7	2	1	9
Número de vehículos yi	15	18	10	8	20

- Calcula el coeficiente de correlación lineal.
- Si ayer se produjeron 6 accidentes, ¿cuántos vehículos podemos suponer que circulaban por la autopista a más de 120 km / h?
- ¿Es buena la predicción?

Construimos una tabla, teniendo en cuenta que la frecuencia absoluta es uno. Debemos conocer la media aritmética de las dos variables, las varianzas, las desviaciones típicas y la covarianza.

		Media aritmética		**Varianza**		**Covarianza**
	fi	**xi**	**yi**	**xi^2**	**yi^2**	**xi . yi**
	1	**5**	**15**	25	225	75
	1	**7**	**18**	49	324	126
	1	**2**	**10**	4	100	20
	1	**1**	**8**	1	64	8
	1	**9**	**20**	81	400	180
$\sum$	5	24	71	160	1113	409

Medias aritméticas

$\bar{x} = \frac{\Sigma x_i}{N} = \frac{24}{5} = 4,8$ $\quad \bar{x} = 4,8$ $\qquad \bar{y} = \frac{\Sigma y_i}{N} = \frac{71}{5} = 14,2$ $\quad \bar{y} = 14,2$

Varianzas y desviaciones típicas

$\sigma_x^2 = \frac{\Sigma (x_i)^2}{N} - (\bar{x})^2 = \frac{160}{5} - (4,8)^2 = 8,96$ $\qquad \sigma_x = \sqrt{8,96} = 2,993$

$\sigma_y^2 = \frac{\Sigma (y_i)^2}{N} - (\bar{y})^2 = \frac{1113}{5} - (14,2)^2 = 20,96$ $\quad \sigma_y = \sqrt{20,96} = 4,578$

Covarianza $\sigma_{xy} \Rightarrow \sigma_{xy} = \frac{\Sigma xi \cdot yi}{N} - \bar{x} \cdot \bar{y} = \frac{409}{5} - (4,8 \cdot 14,2) = 13,64$

a) Correlación lineal de Pearson r $\Rightarrow r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} = \frac{13,64}{2,993 \cdot 4,578} = 0,995 \Rightarrow r = 0,995$

Comentarios:
La covarianza es positiva, correlación directa. Al aumentar la velocidad aumentará el número de accidentes.
El valor de r está muy próximo a 1, la estimación realizada estará muy cerca del valor real.
Dependencia funcional fuerte.

b) Recta de regresión de y sobre x $\Rightarrow y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

$y = 14,2 + \frac{13,64}{8,96}(x - 4,8) \Rightarrow y = 14,2 + 1,52(x - 4,8) \Rightarrow y = 1,52\,x + 6,9$

Para x = 6 accidentes el número de vehículos estimado es: y = 1,52 · 6 + 6,9 = 16
Podemos suponer que ayer circulaban 16 vehículos a más de 120 km/h

c) La predicción hecha es buena, el coeficiente de correlación está muy próximo a uno.

Ejemplo 2

Las calificaciones de 40 alumnos en psicología evolutiva y en estadística han sido las de la tabla adjunta.

Psicología xi	3	4	5	6	6	7	7	8	10
Estadística yi	2	5	5	6	7	6	7	9	10
Nº de alumnos fi	4	6	12	4	5	4	2	1	2

a) Obtener la ecuación de la recta de regresión de calificaciones de estadística respecto de las calificaciones de psicología.
b) ¿Cuál será la nota esperada en estadística para un alumno que obtuvo un 4,5 en psicología?

				Media aritmética		Varianza		Covarianza
	xi	**yi**	**fi**	**fi . xi**	**fi . yi**	**fi . xi^2**	**fi . yi^2**	**fi . xi . yi**
	3	**2**	**4**	12	8	36	16	24
	4	**5**	**6**	24	30	96	150	120
	5	**5**	**12**	60	60	300	300	300
	6	**6**	**4**	24	24	144	144	144
	6	**7**	**5**	30	35	180	245	210
	7	**6**	**4**	28	24	196	144	168
	7	**7**	**2**	14	14	98	98	98
	8	**9**	**1**	8	9	64	81	72
	9	**10**	**2**	20	20	200	200	200
Σ			40	220	224	1314	1378	1336

Medias aritméticas

$$\bar{x} = \frac{\Sigma x_i \cdot fi}{N} = \frac{220}{40} = 55 \qquad \bar{x} = 5,5 \qquad \bar{y} = \frac{\Sigma y_i \cdot fi}{N} = \frac{224}{40} = 5,6 \qquad \bar{y} = 5,6$$

Varianzas y desviaciones típicas

$$\sigma_x^2 = \frac{\Sigma fi \cdot (x_i)^2}{N} - (\bar{x})^2 = \frac{1314}{40} - (5,5)^2 = 2,60 \qquad \sigma_x = \sqrt{2,6} = 1,61$$

$$\sigma_y^2 = \frac{\Sigma fi \cdot (y_i)^2}{N} - (\bar{y})^2 = \frac{1378}{40} - (5,6)^2 = 3,09 \qquad \sigma_y = \sqrt{3,09} = 1,76$$

Covarianza $\sigma_{xy} \Rightarrow \sigma_{xy} = \frac{\Sigma fi \cdot xi \cdot yi}{N} - \bar{x} \cdot \bar{y} = \frac{1336}{40} - (5,5 \cdot 5,6) = 2,6$

Correlación lineal de Pearson $r \Rightarrow r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} = \frac{2,6}{1,61 \cdot 1,76} = 0,92 \Rightarrow r = 0,92$

La covarianza es positiva, correlación positiva fuerte.
El valor de r está muy próximo a 1, la estimación realizada estará muy cerca del valor real.

a) Recta de regresión de y sobre x $\Rightarrow y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

$$y = 5,6 + \frac{2,6}{2,6}(x - 5,5) \Rightarrow y = 5,6 + 1(x - 5,5) \Rightarrow y = x + 0,1$$

b) Nota esperada en estadística habiendo obtenido un 4,5 en psicología.

$y = 4,5 + 0,1 = 4,6$

Ejemplo 3

Las notas obtenidas por 10 alumnos en Matemáticas y en Música son:

Matemáticas	6	4	8	5	3,5	7	5	10	5	4
Música	6,5	4,5	7	5	4	8	7	10	6	5

- Calcula la covarianza y el coeficiente de correlación.
- ¿Existe correlación entre las dos variables?
- ¿Cuál será la nota esperada en Música para un alumno que hubiese obtenido un 8,3 en Matemáticas?

Solución:

a) Covarianza = 3,075. Coeficiente de correlación r = 0,92.

b) Existe una correlación positiva fuerte.

c) Recta de regresión: y = 1,6 + 0,817 x La nota esperada en Música = 8, 38

Ejemplo 4

Cinco niñas de 2, 3, 5, 7 y 8 años de edad pesan respectivamente 14, 20, 30, 42 y 44 Kg . Halla

Ecuación de la recta de regresión: x = 0,192 y - 0,76

Peso aproximado de una niña de 6 años: 35,2 kg

EJEMPLO 5 la ecuación de la recta de regresión de la edad sobre el peso. ¿Cuál sería el peso aproximado de una niña de 6 años?

Solución:

Una asociación dedicada a la protección de la infancia decide estudiar la relación entre la mortalidad infantil en cada país y el número de camas de hospitales por cada mil habitantes.. Datos

x	50	100	70	60	120	180	200	250	30	90
y	5	2	2,5	3,75	4	1	1,25	0,75	7	3

Donde **x** es el nº de camas por mil habitantes e **y** el tanto por ciento de mortalidad.

Se pide calcular las rectas de regresión y el coeficiente de correlación lineal.

¿ Si se dispusiese de 175 camas por mil habitantes que tanto por ciento de mortalidad cabria esperar?. ¿La estimación es fiable? Razona la respuesta.

Solución :

Para facilitar los cálculos de los parámetros se utiliza la siguiente tabla:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	50	5	2500	25	250
	100	2	10000	4	200
	70	2,5	4900	6,25	170
	60	3,75	3600	14,0625	225
	120	4	14400	16	480
	180	1	32400	1	180
	200	1,25	40000	1,5625	250
	250	0,75	62500	0,5625	187,5
	30	7	900	49	210
	90	3	8100	9	270
å =	1150	30,25	179300	126,4375	2422,5

x =115; **y**= 3,025%; $S_x = \sqrt{17930 - 13225} = 68,59$; $S_y = \sqrt{12,64375 - 9,150625} = 1,87$; $S_{xy} = 242,25 - (115)(3,025) = -105,625$

Las rectas de regresión serán por tanto:

y - 3,025 = -0,022449 (x - 115)

x - 115 = -30,2053 (y - 3,025)

El coeficiente de correlación lineal:

$$\mathbf{r} = \frac{-105,625}{(68,59)(1,87)} = -0,8235$$

es una correlación inversa alta .

Para la estimación que nos piden utilizaremos la recta de regresión de Y sobre X.

y= 3,025 - 0,022449(175- 115) = 1,6783 que sería fiable por ser alto el coeficiente de correlación.

EJEMPLO 6

Dada la distribución bidimensional:

X	1	2	1	2	3	2	2	2	3	1
Y	3	5	2	3	5	4	3	5	5	3

Encuentra el valor del coeficiente de correlación lineal usando una tabla de correlación.

Solución

Se usa la siguiente tabla de doble entrada que facilita los cálculos:

X / Y	1	2	3	n'_j	$n_j' y_j$	$n'_j y_j^2$	$n_{ij} x_i y_j$
2	1			1	2	4	2
3	2	2		4	12	36	18
4		1		1	4	16	8
5		2	2	4	20	100	50
n_i	3	5	2	10	å=38	å=156	å=78
$n_i x_i$	3	10	6	å=19			
$n_i x_i^2$	3	20	18	å=41			
$n_{ij} x_i y_j$	7	40	30	å=78			

De aquí se tiene:

x = 19/10= 1,9; **y** = 38/10= 3,8; S_x^2= 4,1 - $(1,9)^2$= 0,49, S_x =0,7 ; S_y^2=15,6 - $(3,8)^2$= 1,16,

S_y = 1,077; S_{xy} = 7,8 - (1,9)(3,8)= 0,58.

Luego **r** = $\frac{0,58}{(0,7)(1,077)}$ = 0,769

EJEMPLO 7

En la tabla siguiente se dan los valores y algunas frecuencias absolutas de un par de variables tratadas conjuntamente. Los valores de la primera fila corresponden a la variable Y, y los de la primera columna a la variable X. La última columna es la marginal de X y la última fila es la marginal de Y.

	1	2	4	7	9	11	
1	1	2		1	0	0	5
3	0			1	1	0	4
4	1	0	2	1	1	3	
5	1	1	3	2	4	0	
6		1	1		1	0	4
7	0	0	0	1	3	1	
	4	5	8	6	10	4	

a) Completar la tabla.

b) Calcular el coeficiente de correlación y las rectas de regresión.

c) ¿Sirven las rectas de regresión para hacer predicciones de una variable en función de la otra? ¿Por qué?

Solución

x y	1	2	4	7	9	11	
1	1	2	**1**	1	0	0	5
3	0	**1**	**1**	1	1	0	4
4	1	0	2	1	1	3	**8**
5	1	1	3	2	4	0	**11**
6	**1**	1	1	**0**	1	0	4
7	0	0	0	1	3	1	**5**
	4	5	8	6	10	4	**37**

b) **x** = $\frac{1.5+3.4+4.8+5.11+6.4+7.5}{37} = 4{,}405$; **y** = $\frac{1.4+2.5+4.8+7.6+9.10+11.4}{37} = 6$

M_{xy} [3] = $\frac{\sum_{ij} x_i y_j n_{ij}}{N}$ = 28,378, luego $S_{xy} = M_{xy}$ - **x.y** = 1,948

$S_x^2 = \frac{1.5+3^2.4+4^2.8+5^2.11+6^2.4+7^2.5}{37} - (4{,}405)^2 = 3{,}11$; S_x=1,764

S_y^2 = 47,027 - 36 = 11,027; S_y = 3,321

El coeficiente de correlación lineal **r** = $\frac{1{,}948}{(1{,}764)(3{,}321)}$ **= 0,3325 <0,40, correlación baja.**

m_{yx} = 1,948/3,11=0,626 y m_{xy} = 1,948/11,027=0,177 son los coeficientes de regresión.

Las rectas de regresión son:

y - 6 = 0,626 (**x** - 4,405) de Y sobre X, y x - 4,405 = 0,177(**y** - 6) de X sobre y

c) Las rectas de regresión no sirven para hacer predicciones, **fiables,** de una variable respecto de la otra ya que la correlación es baja. (*El módulo del coeficiente de correlación lineal está muy alejado de la unidad)*