

Distribuciones Bidimensionales

Contenido

1	ESTADÍSTICA DESCRIPTIVA DE UNA VARIABLE.....	2
1.1	VARIABLES ESTADÍSTICAS.....	2
1.2	FRECUENCIA ABSOLUTA Y F. RELATIVA.....	2
1.3	MEDIDAS DE CENTRALIZACIÓN.....	2
1.4	MEDIDAS DE POSICIÓN.....	3
1.5	MEDIDAS DE DISPERSIÓN.....	4
2	VARIABLES BIDIMENSIONAL.....	7
2.1	DISTRIBUCIÓN CONJUNTA DE UNA VARIABLE BIDIMENSIONAL.....	7
2.2	COVARIANZA.....	8
3	DISTRIBUCCIONES CONDICIONADAS.....	14
3.1	INDEPENDENCIA EN LA TABLA DE CONTINGENCIA.....	14
4	MODELO DE REGRESIÓN LINEAL SIMPLE.....	18
4.1	MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS.....	18
4.2	ERROR CUADRÁTICO MEDIO O VARIANZA RESIDUAL.....	19
5	REGRESIÓN Y CORRELACIÓN.....	23

1 Estadística descriptiva de una variable

Este apartado con más detalle y ejercicios se puede seguir en el apartado Estadísticas del curso de 4º de la ESO

1.1 Variables estadísticas

Miden las características de una población:

Cualitativas: Nos indican una cualidad no numérica, raza, religión, color de pelo..

Cuantitativa: Son aquellas que se pueden medir,

Discretas :sus valores son limitados, número de hijos, número de alumnos..

Continuas: pueden tomar cualquier valor, altura de los alumnos.

1.2 Frecuencia absoluta y F. relativa

1. Ejercicio

Analicemos las notas obtenidas por 30 alumnos en el último examen

2 6 4 1 7 5 4 6 8 5 6 5 7 6 3 2 1 7 6 8 4 3 5 1 2 3 5 6 4 5

N Tamaño de la población

0 muestra (30)

X_i : variable o dato

estadistic. o Variando i de 1 a

n obtenemos todos los valores

f_i : Frecuencia absoluta

es el número de veces que

el dato aparece en la muestra

F_i : Frecuencia absoluta

acumulada ; Para cada dato

es igual a la suma de todas

las f_i inferiores a el

h_i : Frecuencia relativa : frecuencia relativa de un dato es igual al cociente de la frecuencia

absoluta de ese dato dividido por el numero total de datos $h_i = \frac{f_i}{N}$

H_i : Frecuencia relativa acumulada: para cada dato es igual a las h_i de todos los datos inferiores a él.

* las frecuencias acumuladas no se pueden calcular con variables cualitativas

X_i	f_i	F_i	h_i	h_i %	H_i	H_i %
1	3	3	0,1	10%	0,1	10%
2	3	6	0,1	10%	0,2	20%
3	3	9	0,1	10%	0,3	30%
4	4	13	0,13333	13%	0,43333	43%
5	6	19	0,2	20%	0,63333	63%
6	6	25	0,2	20%	0,83333	83%
7	3	28	0,1	10%	0,93333	93%
8	2	30	0,06667	7%	1	100%
	N=30		1			

1.3 Medidas de centralización

Analizando los datos del ejercicio 1

1 1 1 2 2 2 3 3 3 4 4 4 4 5 5 5 5

5 5 6 6 6 6 6 6 7 7 7 8 8

$\bar{X}$ media aritmética: Es igual a la suma, de multiplicar todos los datos por su frecuencia, y dividirlo

por en nuemro total de datos. $\bar{X} = \frac{\sum_{i=1}^n X_i * f_i}{N}$

$$\bar{X} = \frac{\sum_{i=1}^n X_i * f_i}{N} = \frac{1 * 3 + 2 * 3 + 3 * 3 + 4 * 4 + 5 * 6 + 6 * 6 + 7 * 3 + 8 * 2}{30} = 4,5666$$

M_o Moda. Es el dato que tiene mayor frecuencia . Si la variable es continua le llamamos intervalo modal

M_o. El dato que más se repite es el 5 y 6

M_e Mediana. Es el dato que ocupa el lugar central, después de ordenar los datos de menor a mayor .

Si el número total de datos es par la mediana es la media de los dos datos centrales

1 1 1 2 2 2 3 3 3 4 4 4 4 5 5 5 5
5 5 6 6 6 6 6 6 7 7 7 8 8

$$M_e \text{ Mediana} = \frac{5 + 5}{2} = 5$$

1.4 Medidas de posición

Las medidas de posición nos indican el lugar que ocupa un dato dentro del conjunto ordenado de los valores .

Cuartiles. Q₁, Q₂, Q₃ Son los tres valores que dividen al conjunto de datos ordenados en cuatro partes porcentualmente iguales. En cada tramo están el 25% de los datos del estudio

Percentil : Son puntos que dividen al conjunto de datos en 100 partes iguales .

Para hacer esta medida la variable debe ser cuantitativa y si la variable es continua tomaremos como valor la marca de clase

2. Ejercicio

Analicemos las notas obtenidas por 30 alumnos en el último examen

2 6 4 1 7 5 4 6 8 5 6 5 7 6 3 2 1 7 6 8 4 3 5 1 2 3 5 6 4 5

y calculemos los Cuartiles. Q₁, Q₂, Q₃ y el percentil P85

Solución

Q₁ : 25% de los valores están por debajo de este cuartil y el 75% por encima

Para calcularlo dividimos el número de datos por 4 ; $\frac{30}{4} = 7,5 \Rightarrow 25\%$

En la columna de las frecuencias absolutas acumuladas **F_i** el primer dato cuyo valor sea mayor o igual de 7,5 es el 9. Luego Q₁ corresponde al dato 3 (la cuarta parte de las notas son menores o igual a 3)

Q₂ : 50% de los valores están por debajo de este cuartil y el 50% por encima

Para calcularlo dividimos el número de datos por 2 ; $\frac{30}{2} = 15 \Rightarrow 50\%$

En la columna de las frecuencias absolutas acumuladas **F_i** el primer dato cuyo valor sea mayor de 15 es el 19. Luego Q₂ corresponde al dato 5 (la mitad de las notas son menores o igual a 5)

Q₃ : 75% de los valores están por debajo de este cuartil y el 25% por encima

Para calcularlo multiplicamos el número de datos por $\frac{3}{4}$; $\frac{30 \cdot 3}{4} = 22,5 \Rightarrow 75\%$

En la columna de las frecuencias absolutas acumuladas **F_i** el primer dato cuyo valor sea mayor o igual de 22,5 es el 25. Luego Q₃ corresponde al dato 6 (el 75% de las notas son menores o igual a 6)

Percentil 85: El 85% de los valores están por debajo de este percentil

Para calcularlo multiplico el numero de datos por $\frac{85}{100}$; $30 * \frac{85}{100} = 25,5 \Rightarrow 85\%$

En la columna de las frecuencias absolutas acumuladas F_i el primer dato cuyo valor sea mayor o igual de 22,5 es el 28. Luego P85 corresponde al dato 7 (el 85% de las notas son menores o igual a 7)

X_i	f_i	F_i	h_i	$h_i \%$	H_i	$H_i \%$			
1	3 $\rightarrow$	3	0,1	10%	0,1	10%			
2	$\leftarrow$ 3	6	0,1	10%	0,2	20%			
3	3	9	0,1	10%	0,3	30%	Q_1	7,5	25%
4	4	13	0,13333	13%	0,43333	43%			
5	6	19	0,2	20%	0,63333	63%	Q_2	15,0	50%
6	6	25	0,2	20%	0,83333	83%	Q_3	22,5	75%
7	3	28	0,1	10%	0,93333	93%	P85	25,5	85%
8	2	30	0,06667	7%	1	100%			
	N=30		1	100%					

1.5 Medidas de dispersión

También llamadas medidas de variabilidad, muestran la variabilidad de una distribución, indicando lo alejados que están los datos respecto a un valor central (media aritmética).

Rango o recorrido: $R = \text{Maximo valor de la variable} - \text{minimo valor de la variable}$

Desviacion media DM : Es la media de las desviaciones absolutas

Se calcula en valor absoluto, pues el sumatorio siempre dará 0

$$DM = \frac{\sum_{i=1}^n |X_i - \bar{X}| f_i}{N}$$

Varianza σ^2 : Es la media de los cuadrados de las desviaciones

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$$

Desviacion Tipica $\sigma = \sqrt{\sigma^2}$: Es la raiza cuadrada positiva de la varianza

Coefficiente de Variación $CV = \frac{\sigma}{\bar{X}}$: Es el cociente entre la desviacion Tipica y la media aritmetica

A mayor valor del CV mayor heterogeneidad de los valores de la variable. A menor CV mayor homogeneidad en los valores de la variable

3. Ejercicio

Analicemos las notas obtenidas por 30 alumnos en el último examen y calcular todas las medidas de dispersión

2 6 4 1 7 5 4 6 8 5 6 5 7 6 3 2 1 7 6 8 4 3 5 1 2 3 5 6 4 5

Solucion

$$\bar{X} = \frac{\sum_{i=1}^n X_i * f_i}{N} = \frac{1*3 + 2*3 + 3*3 + 4*4 + 5*6 + 6*6 + 7*3 + 8*2}{30} = 4,5666;$$

X_i	f_i	$\bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})f_i$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
1	3	4,56666	3,56666	10,69998	12,72106	38,16319
2	3	4,56666	2,56666	7,69998	6,587744	19,76323
3	3	4,56666	1,56666	4,69998	2,454424	7,36327
4	4	4,56666	0,56666	2,26664	0,321104	1,28441
5	6	4,56666	0,43334	2,60004	0,187784	1,12670
6	6	4,56666	1,43334	8,60004	2,054464	12,32678
7	3	4,56666	2,43334	7,30002	5,921144	17,76343
8	2	4,56666	3,43334	6,86668	11,78782	23,57565
	N=30		1	50,73336		121,36667
				$\sum_{i=1}^n X_i - \bar{X} f_i$	$\sum_{i=1}^n (X_i - \bar{X})^2 f_i$	

$$DM = \frac{\sum_{i=1}^n |X_i - \bar{X}| f_i}{N} = \frac{50,73336}{30} = 1,691112$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{121,36667}{30} = 4,0455556$$

$$\text{Desviación Típica } \sigma = \sqrt{\sigma^2}, \sigma = 2,0113566$$

$$\text{Coeficiente de Variación } CV = \frac{\sigma}{\bar{X}}; CV = 0,440443$$

4. Ejercicio

En un periodo de fuertes lluvias se midió durante 48 horas el caudal de río Ebro en $\frac{m^3}{s}$.

- Hallar la media , mediana y clase modal
- Calcular la varianza y desviación típica

X_i	(26-38)	(38-50)	(50-62)	(62-74)	(74-86)	(86-98)	(98-110)
f_i	6	2	8	10	10	4	8

Solucion

$$\bar{X} = \frac{\sum_{i=1}^n X_i * f_i}{N} = \frac{32 * 6 + 44 * 2 + 56 * 8 + 68 * 10 + 80 * 10 + 92 * 4 + 104 * 8}{48} = \frac{3408}{48} = 71 \frac{m^3}{s};$$

	X_i (marca de clase)	f_i	$\bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})f_i$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
(26-38)	32	6	71	39	234	1521	9126
(38-50)	44	2	71	27	54	729	1458
(50-62)	56	8	71	15	120	225	1800
(62-74)	68	10	71	3	30	9	90
(74-86)	80	10	71	9	90	81	810
(86-98)	92	4	71	21	84	441	1764
(98-110)	104	8	71	33	264	1089	8712
		N=48		147	876	4095	23760

$$\sum_{i=1}^n |X_i - \bar{X}| f_i \quad \sum_{i=1}^n (X_i - \bar{X})^2 f_i$$

Clase modal = (62 – 74) y (74 – 86)

Mediana, M; Es el valor que ocupa el punto medio del conjunto ordenado, luego será el 2M = 2

$$DM = \frac{\sum_{i=1}^n |X_i - \bar{X}| f_i}{N} = \frac{876}{48} = 18,25$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{23760}{48} = 495$$

$$\text{Desviación Típica } \sigma = \sqrt{\sigma^2}, \sigma = 22,248595 \frac{m}{s}$$

$$\text{Coeficiente de Variación } CV = \frac{\sigma}{\bar{X}}; CV = \frac{22,248595}{71} = 0,3133604$$

5. Ejercicio

En una muestra de 150 familias. El número de hijos por familia se recoge en la tabla.

- Hallar la media, mediana y clase modal
- Calcular los cuartiles
- Calcular la varianza y desviación típica

X_i	0	1	2	3	4	5
f_i	20	32	59	28	8	3

Solución

$$\bar{X} = \frac{\sum_{i=1}^n X_i * f_i}{N} = \frac{0 * 20 + 1 * 32 + 2 * 59 + 3 * 28 + 4 * 8 + 5 * 3}{150} = \frac{281}{150} = 1,873;$$

X_i	f_i	$\bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})f_i$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$	h_i	H_i	$H_i \%$	
0	20	1,873	1,873	37,46	3,508	70,16258	0,1333	0,1333	13%	
1	32	1,873	0,873	27,936	0,762	24,388128	0,2133	0,3467	35%	Q_1
2	59	1,873	0,127	7,493	0,016	0,951611	0,3933	0,7400	74%	Q_2
3	28	1,873	1,127	31,556	1,270	35,563612	0,1867	0,9267	93%	Q_3
4	8	1,873	2,127	17,016	4,524	36,193032	0,0533	0,9800	98%	
5	3	1,873	3,127	9,381	9,778	29,334387	0,0200	1,0000	100%	
	N=150		9,254	130,842	19,859	196,59335	1			

$$\sum_{i=1}^n |X_i - \bar{X}| f_i \quad \sum_{i=1}^n (X_i - \bar{X})^2 f_i$$

$$M_o \text{Moda} = 2$$

M_e Mediana; La mitad del conjunto de datos es 75. Si ordenamos los datos de menor a mayor el que ocupa el lugar central es el 2; **M_e = 2**

$$Q_2 = 1; Q_2 = 2, Q_3 = 3$$

$$DM = \frac{\sum_{i=1}^n |X_i - \bar{X}| f_i}{N} = \frac{130,842}{150} = 0,872$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{196,59335}{150} = 1,310622$$

$$\text{Desviación Típica } \sigma = \sqrt{\sigma^2}, \sigma = 1,14482$$

$$\text{Coeficiente de Variación } CV = \frac{\sigma}{\bar{X}}; CV = \frac{0,36142}{1,873} = 0,112248$$

2 Variables bidimensional

Una variable estadística bidimensional (x, y) representa dos características de una población, en la que cada elemento de la muestra se le asocian dos valores, uno por cada variable.

2.1 Distribución conjunta de una variable bidimensional

Analizamos los valores de un test realizado a 91 niños de edades comprendidas entre 5 y 8 años

Marginal de X: expresa cómo se distribuye X en la población total, al margen de la otra variable

Marginal de Y: expresa cómo se distribuye Y en la población total, al margen de la otra variable

Edad (x)	Test (y)			Marginal (x)
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25
Marginal (y)	20	30	41	91

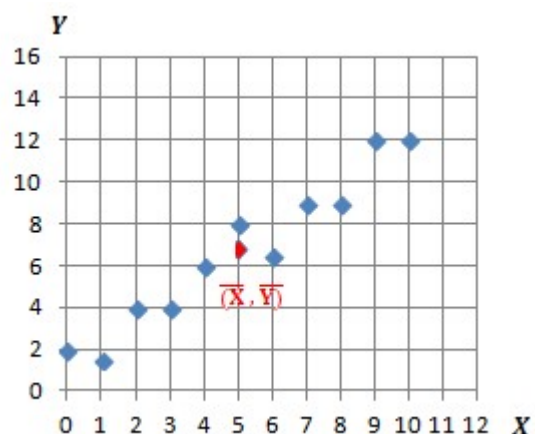
Diagrama de dispersión

Este tipo de gráfico indica de forma visual la relación entre las dos variables. Mostrando su dependencia.

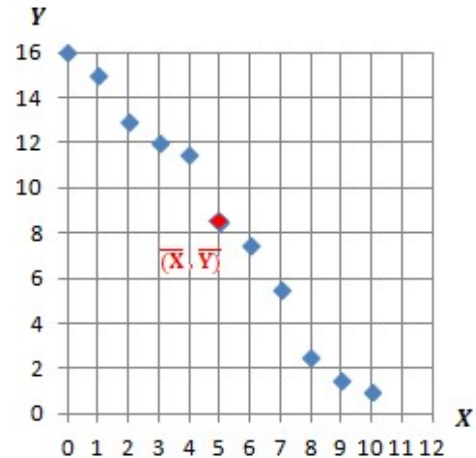
Centro de gravedad :

$$\text{Centro de gravedad} = (\bar{X}, \bar{Y})$$

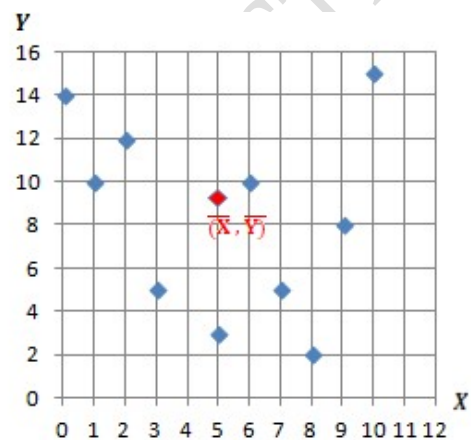
Dependencia Directa: a mayor valor de X mayor valor de Y.



Dependencia Inversa: a mayor valor de X menos valor de Y.



Independencia no hay relación entre X e Y



2.2 Covarianza

Es la variación conjunta de las dos variables

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

Covarianza Positiva: relación directa

Covarianza Negativa: relación inversa

Covarianza próxima a 0: no hay relación entre las variables

6. Ejercicio

Analizamos los valores de un test realizado a 91 niños de edades comprendidas entre 5 y 8 años del ejercicio anterior y hallar la covarianza

Edad (x)	Test (y)			Marginal (x)
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25
Marginal (y)	20	30	41	91

Solución

X/Y	120	125	130		X_i	Y_j	f_{ij}	$f_{ij}(X_i Y_j)$
5	10	8	2		5	120	10	6000
6	7	8	6		5	125	8	5000
7	2	10	13		5	130	2	1300
8	1	4	20		6	120	7	5040
					6	125	8	6000
					6	130	6	4680
					7	120	2	1680
					7	125	10	8750
					7	130	13	11830
					8	120	1	960
					8	125	4	4000
					8	130	20	20800
								76040

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_i^k \sum_j^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n} \sum_i^k \sum_j^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

$$\bar{X} = \frac{\sum X_i f_i}{n} = \frac{5 * 20 + 6 * 21 + 7 * 25 + 8 * 25}{91} = \frac{100 + 126 + 175 + 200}{91} = \frac{601}{91} = 6,6044 ;$$

$$\bar{Y} = \frac{\sum Y_j f_j}{n} = \frac{120 * 20 + 125 * 30 + 130 * 41}{91} = \frac{2400 + 3750 + 5330}{91} = \frac{11480}{91} = 126,154$$

$$S_{xy} = \frac{76040}{91} - (6,6044 * 126,154) = 2,433 \text{ La relación es directa}$$

7. Ejercicio

Con el fin de relacionar el índice de fluor en el agua (X en ppm) con la tasa de caries (Y en %) se han tomado muestras en 10 ciudades, los datos obtenidos son:

X	1,6	7,8	7,1	2,3	5,8	4,2	7,6	9,8	6,4	7,6
Y	2,0	4,0	5,0	5,5	6,0	6,5	7,1	7,6	8,4	9,3

- Calcular las medias y varianza de las distribuciones marginales
- Representar la nube de puntos, indicar la relación entre las variables y su tendencia
- Hallar la covarianza

Solución

- Calcular las medias y varianza de las distribuciones marginales

$$\bar{X} = \frac{60,2}{10} = 6,02; \quad \bar{Y} = \frac{61,4}{10} = 6,14$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$$

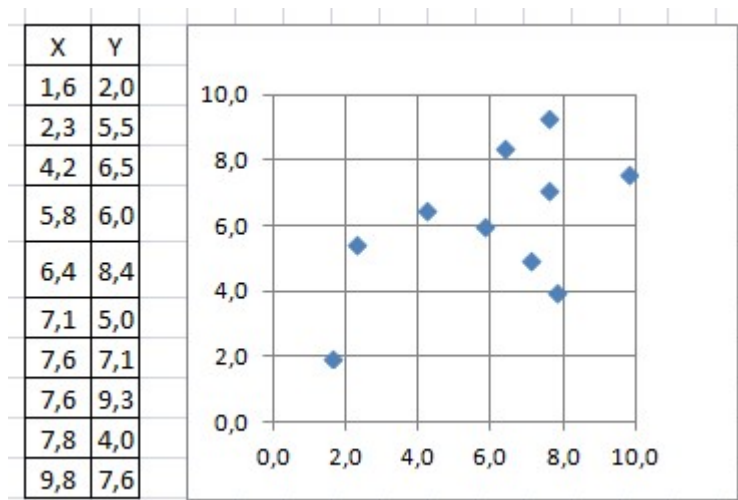
$(X_i - \bar{X})^2 f_i$	19,71	3,10	1,12	13,99	0,06	3,39	2,43	14,14	0,13	2,43
$(Y_i - \bar{Y})^2 f_i$	17,14	4,58	1,30	0,41	0,02	0,13	0,92	2,13	5,11	9,99

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{60,5}{10} = 6,05$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{41,72}{10} = 4,172$$

b) Representar la nube de puntos, indicar la relación entre las variables y su tendencia.

Las variables son independientes



c) Hallar la covarianza

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_i^k \sum_j^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n} \sum_i^k \sum_j^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

X	1,6	7,8	7,1	2,3	5,8	4,2	7,6	9,8	6,4	7,6
Y	2,0	4,0	5,0	5,5	6,0	6,5	7,1	7,6	8,4	9,3
$f_{ij}(X_i Y_j)$	3,2	31,2	35,5	12,7	34,8	27,3	54,0	74,5	53,8	70,7

$$S_{xy} = \frac{397,53}{10} - (6,02 * 6,14) = 39,753 - 36,96 = 2,793$$

8. Ejercicio

El número de libros de consultas solicitadas(X) en una biblioteca municipal, junto con el número de revistas (Y) se recoge en la tabla siguiente

a) Calcular las distribuciones

marginales

b) hallar las medias y varianzas

marginales

c) Cuantifica relación entre variables

X\Y	[10,15)	[15,20)	[20,25)	[25,30)
[20,30)	23	62	163	28
[30,40)	24	55	159	22
[40,50)	33	65	127	12

Solución

a) Calcular las distribuciones marginales

	X\Y	[10,15)	[15,20)	[20,25)	[25,30)	Marginal X
	[20,30)	23	62	163	28	276
	[30,40)	24	55	159	22	260
	[40,50)	33	65	127	12	237
Marginal Y		80	182	449	62	

b) Hallar marcas de clase, las medias y varianzas marginales

$$\text{Para calcular la marca de clase de cada intervalo} = \frac{L_{\text{superior}} + L_{\text{inferior}}}{2};$$

Ejemplo $\frac{30 + 20}{2} = 25$; $\frac{10 + 15}{2} = 12,5$

	X\Y	12,5	17,5	22,5	27,5	Marginal X
	25	23	62	163	28	276
	35	24	55	159	22	260
	45	33	65	127	12	237
Marginal Y		80	182	449	62	n=773

$$\bar{X} = \frac{25 * 276 + 35 * 260 + 45 * 237}{773} = 34,5;$$

$$\bar{Y} = \frac{12,5 * 80 + 17,5 * 182 + 22,5 * 449 + 27,5 * 62}{773} = 20,69$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$$

X_i	f_i	$\bar{X}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
25	276	34,5	9,5	90,250	24909
35	260	34,5	-0,5	0,250	65
45	237	34,5	10,5	110,250	26129,25
	N=773		19,5	200,750	51103,25

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{51103,25}{773} = 66,11$$

(Y_i)	f_i	$\bar{Y}$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2 f_i$
12,5	80	20,69	8,19	67,076	5366,09
17,5	182	20,69	3,19	10,176	1852,05
22,5	449	20,69	-1,81	3,276	1470,97
27,5	62	20,69	-6,81	46,376	2875,32
	N=773		2,76	126,904	11564,425

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{11464,425}{773} = 14,96$$

c) Cuantifica relacion entre variables

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_i^k \sum_j^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y})$$

$$= \frac{1}{n} \sum_i^k \sum_j^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

$$S_{xy} = \frac{f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y};$$

$$S_{xy} = \frac{549012,50}{773} - 34,5 * 20,69 =$$

$$S_{xy} = 710,23609 - 713,805 =$$

$$S_{xy} = -3,56891$$

x	y	fi	$f_{ij}(X_i Y_j)$
25	12,5	23	7187,5
25	17,5	62	27125
25	22,5	163	91687,5
25	27,5	28	19250
35	12,5	24	10500
35	17,5	55	33687,5
35	22,5	159	125212,5
35	27,5	22	21175
45	12,5	33	18562,5
45	17,5	65	51187,5
45	22,5	127	128587,5
45	27,5	12	14850
			549012,50

9. Ejercicio

En una muestra de 10 viviendas de una urbanización se ha contado el número de personas que viven en cada vivienda (X) y el número de habitaciones que tienen.

a) dibuja el diagrama de dispersion. Comprobar si existe relacion lineal.

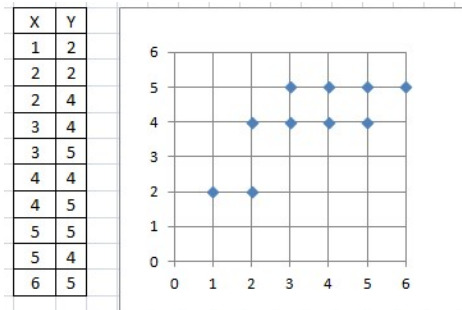
b) Calcular covarianza y varianza

Solución

a) dibuja el diagrama de dispersion.

Comprobar si existe relacion lineal

No hay relacion lineal



b) Calcular covarianza y varianza

$$\bar{X} = \frac{35}{10} = 3,5 ; \quad \bar{Y} = \frac{40}{10} = 4$$

X	5	3	2	4	1	3	6	2	5	4	35	$\sum_{i=1}^n X_i f_i$
Y	5	4	2	4	2	5	5	4	4	5	40	$\sum_{i=1}^n Y_i f_i$
$f_{ij}(X_i Y_j)$	25,0	12,0	4,0	16,0	2,0	15,0	30,0	8,0	20,0	20,0	152	$\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j$
$(X_i - \bar{X})^2 f_i$	2,25	0,25	2,25	0,25	6,25	0,25	6,25	2,25	2,25	0,25	22,50	$\sum_{i=1}^n (X_i - \bar{X})^2 f_i$
$(Y_j - \bar{Y})^2 f_j$	1,00	0,00	4,00	0,00	4,00	1,00	1,00	0,00	0,00	1,00	12,00	$\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j$

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{22,5}{10} = 2,25$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j}{N} = \frac{12}{10} = 1,2$$

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

$$S_{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y} ; S_{xy} = \frac{152}{10} - (3,5 * 4) = 15,2 - 14,0 = 1,2$$

10. Ejercicio

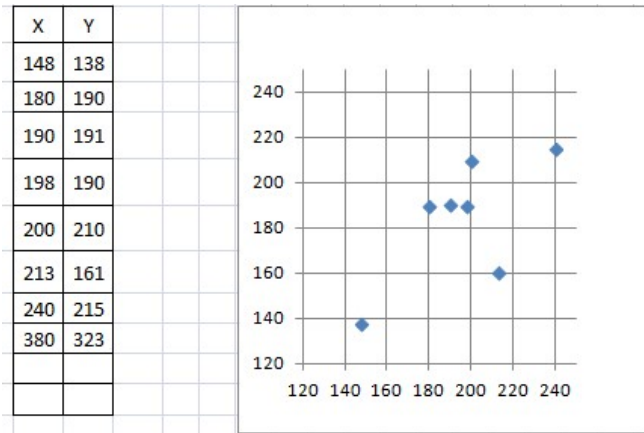
En una fábrica se quiere probar la resistencia al calor de una determinada clase de cerámica. Se elije una muestra de 8 pares de piezas idénticas (de la misma hornada). De cada pareja una fue sometida a la prueba de dureza antes del proceso térmico y se anoto su punto de rotura X y la otra después del horneado Y

a) Calcula la media y la mediana de la resistencia antes y despues del proceso termico

b) En que caso hay mas variabilidad, antes o despues del proceso termico?

b) ¿ Se puede afirmar que existe relacion linial entre las dos variable?

Solución



a) Calcula la media y la mediana de la resistencia antes y después del proceso termico

X	148	213	380	180	200	190	240	198	1749	$\sum_{i=1}^n X_i f_i$
Y	138	161	323	190	210	191	215	190	1618	$\sum_{j=1}^n Y_j f_j$
$f_{ij}(X_i Y_j)$	20424,0	34293,0	122740,0	34200,0	42000,0	36290,0	51600,0	37620,0	379167	$\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j$
$(X_i - \bar{X})^2 f_i$	20880,25	43890,25	141752,25	31152,25	38612,25	34782,25	55932,25	37830,25	404832,00	$\sum_{i=1}^n (X_i - \bar{X})^2 f_i$
$(Y_j - \bar{Y})^2 f_j$	17956,00	24649,00	101761,00	34596,00	42436,00	34969,00	44521,00	34596,00	335484,00	$\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j$

$$\bar{X} = \frac{1749}{8} = 218,625 ; \quad \bar{Y} = \frac{1618}{8} = 202,25$$

$$\text{Mediana } X = \frac{180+200}{2} = 190;$$

$$\text{Mediana } y = \frac{190 + 191}{2} = 190,5$$

X	148	180	190	198	200	213	240	380
Y	138	161	190	190	191	210	215	323

b) En que caso hay mas variabilidad, antes o después del proceso termico?

$$\text{Varianza } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{404832}{8} = 50604$$

$$\text{Varianza } \sigma^2 = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j}{N} = \frac{335484}{8} = 41935,5$$

Hay más variabilidad antes del proceso

b) ¿Se puede afirmar que existe relacion linial entre las dos variable?

$$\text{Covarianza} = S_{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j - \bar{X}\bar{Y}$$

$$S_{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y}; \quad S_{xy} = \frac{379167}{8} - (218,625 * 202,25) =$$

$$47395,8975 - 44216,90625 = 3178,99125 \text{ Existe una relacion directa}$$

3 Distribuciones condicionadas

Analizamos los valores de un test realizado a 91 niños de edades comprendidas entre 5 y 8 años del ejercicio 6

Podremos elegir la distribución de una variable Y en función de una variable X

Ejemplo distribución de Y en función de la variable X

De esta tabla podemos deducir:

Que 20 alumnos tienen 5 años, 21 tienen 6; 25 tienen 7 y 25 tienen 8

Que 20 alumnos sacaron un 120, 30 sacaron 125 y 41 obtuvieron 130

Distribución condicionada de Y en función de los alumnos de edad 6 años

Se representa como $Y|X = X_2$

La distribución de la frecuencia absoluta y relativa es la que se indica para los alumnos de 6 años

Edad (x)	Test (y)			Marginal (x)
	120 Y_1	125 Y_2	130 Y_3	
5 X_1	10	8	2	20
6 X_2	7	8	6	21
7 X_3	2	10	13	25
8 X_4	1	4	20	25
Marginal (y)	20	30	41	91

$Y X = X_2$	f_{2j}	$h_{j 2}$
120 Y_1	7	0,33
125 Y_2	8	0,38
130 Y_3	6	0,29
	21	1,00

La frecuencia relativa de se representa como: $y_j | X = x_i$ es $h_{j|i} = \frac{f_{ij}}{f_{i.}}$ para $j = 1, 2, 3.. i = 1, 2, 3.$

De igual manera la distribución condicionada de x a los valores que tome Y es:

$x_i | Y = y_j$ es $h_{i|j} = \frac{f_{ij}}{f_{.j}}$ para $j = 1, 2, 3.. i = 1, 2, 3.$

$$h_{i|j} = h_{i.} * h_{j|i} = h_{.j} * h_{i|j}$$

Las frecuencias relativas conjuntas nos indican la frecuencia de una variable en relación con otra variable

En el ejemplo

En el ejemplo $y_j | X = x_2$ es $h_{j|2} = \frac{f_{2j}}{f_{2.}}$

Frecuencia relativa conjunta de alumnos de 6 años que hayan obtenido 125 = $\frac{8}{21} = 0,381 \Rightarrow 38,1\%$

Frecuencia relativa conjunta de alumnos hayan obtenido 130 y tengan 5 años $25 = \frac{2}{41} = 0,049 = 4,9\%$

3.1 Independencia en la tabla de contingencia

La variable X es independiente de la variable Y es decir, la variable Y no modifica la frecuencia marginal de X

$$h_{i|j} = h_{i.} \text{ para cualquier } j \Rightarrow h_{i|j} = h_{i.} * h_{.j}$$

11. Ejercicio

De la tabla siguiente

	3	4	5	6	Marginal (x)
5	3	12	6	15	36
10	5	20	10	25	60
Marginal (y)	8	32	16	40	96

Hallar la distribución de X condicionada por cada valor de Y y el valor de Y condicionada por cada valor de X.

Comprobar si las variables son independientes

Solución

1.-Tabla de distribución de X condicionada a Y

Para i = 1

$$1^{\text{a}} \text{ celda} = \frac{3}{8} = 0,375$$

$$2^{\text{a}} \text{ celda} = \frac{12}{32} = 0,375$$

$$3^{\text{a}} \text{ celda} = \frac{6}{16} = 0,375$$

	X y _i			
X	h _{i 1}	h _{i 2}	h _{i 3}	h _{i 4}
X ₁ = 5	0,375	0,375	0,375	0,375
X ₂ = 10	0,625	0,625	0,625	0,625
	1	1	1	1

todas las celdas dan el mismo valor independientes del valor de j

Para i = 2

$$1^{\text{a}} \text{ celda} = \frac{5}{8} = 0,625 ; 2^{\text{a}} \text{ celda} = \frac{20}{32} = 0,625$$

todas las celdas dan el mismo valor independientes del valor de j

2.- Tabla de distribución de Y condicionada a X

Para j = 1

$$1^{\text{a}} \text{ celda} = \frac{3}{36} = 0,083$$

$$2^{\text{a}} \text{ celda} = \frac{5}{60} = 0,083$$

Para j = 2

$$1^{\text{a}} \text{ celda} = \frac{12}{36} = 0,333$$

$$2^{\text{a}} \text{ celda} = \frac{20}{60} = 0,333$$

todas las celdas dan el mismo valor independientes del valor de i

	3	4	5	6
h _{1 j}	0,083	0,3333	0,16667	0,4166667
h _{2 j}	0,083	0,3333	0,16667	0,4166667

3.- Comprobar si las variables son independientes

	3	4	5	6	
h _{ij}					h _{i.}
5	0,031	0,125	0,0625	0,15625	0,375
10	0,052	0,2083	0,10417	0,2604167	0,625
h _{.j}	0,083	0,3333	0,16667	0,4166667	1

Las variables hemos visto que son independientes pero para comprobarlo veamos si

$$h_{ij} = h_{i.} * h_{.j}$$

$$h_{1|1} = 0,375 * 0,083 = 0,031$$

$$h_{1|2} = 0,375 * 0,3333 = 0,125$$

$$h_{2|4} = 0,625 * 0,41666 = 0,2604167$$

Se cumple son independientes

12. Ejercicio

La distribución de fumadores según el sexo X y el rango de edad entre 15 y 65 años Y

- Escribe la distribución marginal
- Halla la distribución de frecuencia relativa de Y condicionada a X
- Halla la media y varianza de Y|mujeres
- ¿Son dependientes estas variables?

Solución

- Distribución marginal

	[15,24]	[25,34]	[35,44]	[45,54]	Marginal (x)
Hombres	112	178	164	172	626
Mujeres	105	141	141	150	537
Marginal (y)	217	319	305	322	1163

- Halla la distribución de frecuencia relativa de Y condicionada a X

	[15,24]	[25,34]	[35,44]	[45,54]
h_{1j}	0,178914	0,284345	0,26198	0,27476
h_{2j}	0,195531	0,2625698	0,26257	0,27933

- Halla la media y varianza de Y|mujeres

	19,5	29,5	39,5	49,5	Marginal (x)
Hombres	112	178	164	172	626
Mujeres	105	141	141	150	537
Marginal (y)	217	319	305	322	1163

$$\bar{Y}_{x=mujeres} = \frac{19,5 * 105 + 29,5 * 141 + 39,5 * 141 + 49,5 * 150}{537} =$$

$$= \frac{2047,5 + 4159,5 + 5569,5 + 7425}{537} = \frac{19201,5}{537} = 35,76$$

$$\text{Varianza } \sigma^2_{x=mujeres} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{(19,5 - 35,76)^2 * 105}{537} = \frac{(29,5 - 35,76)^2 * 141}{537} +$$

$$\frac{(39,5 - 35,76)^2 * 141}{537} + \frac{(49,5 - 35,76)^2 * 150}{537} = 51,68 + 10,28 + 3,68 + 52,76 = 118,39$$

- Son independientes las variables. SI

h_{ij}	[15,24]	[25,34]	[35,44]	[45,54]	h_i
Hombres	0,092343	0,1586628	0,14087	0,14677	0,53826311
Mujeres	0,094612	0,1160575	0,12138	0,13012	0,46173689
h_j	0,186586	0,2742906	0,26225	0,27687	1

$$h_{ij} = h_i * h_j$$

$$h_{1|1} = 0,53826311 * 0,186586 = 0,10043 \neq 0,092343$$

$$h_{1|2} = 0,53826311 * 0,2742906 = 0,147640527 \neq 0,1586628$$

$$h_{2|4} = 0,46173689 * 0,27687 = 0,127841167 \neq 0,13012$$

No se cumplen luego son dependientes

13. Ejercicio

La tabla recoge las puntuaciones de 25 trabajadores en un test (Y) y las ausencias al trabajo en un mes

- Hallar la media y varianza de las distribuciones marginales
- Halla la media y varianza de de la distribución de Y| x=0
- Hallar la covarianza

Solución

- Distribución marginal

	[40,60]	[60,80]	[80,100]
0	2	4	5
1	2	4	2
2	1	3	2

	[40,60]	[60,80]	[80,100]	Marginal (x)
0	2	4	5	11
1	2	4	2	8
2	1	3	2	6
Marginal (y)	5	11	9	25

	50	70	90	Marginal (x)
0	2	4	5	11
1	2	4	2	8
2	1	3	2	6
Marginal (y)	5	11	9	25

$$\bar{Y}|_{x=i} = \frac{0 \cdot 11 + 1 \cdot 8 + 2 \cdot 6}{25} = \frac{20}{25} = 0,8$$

$$\text{Varianza } \sigma^2_{x=i} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{(0 - 0,8)^2 \cdot 11}{25} + \frac{(1 - 0,8)^2 \cdot 8}{25} + \frac{(2 - 0,8)^2 \cdot 6}{25} =$$

$$= 0,20485 + 0,32 + 0,3456 = 0,87$$

$$\bar{X}|_{y=j} = \frac{50 \cdot 5 + 70 \cdot 11 + 90 \cdot 9}{25} = 73$$

$$\text{Varianza } \sigma^2_{x=i} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{(50 - 73)^2 \cdot 5}{25} + \frac{(70 - 73)^2 \cdot 11}{25} + \frac{(90 - 73)^2 \cdot 9}{25} = 213,8$$

$$= 105,8 + 3,96 + 104,04 = 213,8$$

- Halla la media y varianza de de la distribución de Y| x=0

$$\bar{Y}|_{x=0} = \frac{50 \cdot 2 + 70 \cdot 4 + 90 \cdot 5}{11} = \frac{830}{11} = 75$$

$$\text{Varianza } \sigma^2_{x=0} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{(50 - 75)^2 \cdot 2}{11} + \frac{(70 - 75)^2 \cdot 4}{11} + \frac{(90 - 75)^2 \cdot 5}{11} =$$

$$= 113,636 + 9,0909091 + 102,27 = 225$$

- Hallar la covarianza

$$S_{xy} = \frac{\sum_i^k \sum_j^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y};$$

$$S_{xy} = \frac{36600}{25} - 0,8 \cdot 73$$

X	0	8	12	20	$\sum_{i=1}^n X_i f_i$
Y	250	770	810	1830	$\sum_{i=1}^n Y_i f_i$
$f_{ij}(X_i, Y_j)$	0,0	6160,0	9720,0	36600,0	$\sum_i^k \sum_j^h f_{ij} X_i Y_j$

$$S_{xy} = 1464 - 58,4 = 1405,6 \text{ Existe una relacion directa}$$

4 Modelo de regresión lineal simple

El diagrama de dispersión indica de forma visual la relación entre las dos variables.

El modelo de regresión lineal consiste en construir una recta que se encuentre lo mas cerca posible de la verdadera respuesta.

4.1 Método de mínimos cuadrados ordinarios

Es un procedimiento de análisis numérico en la que, dados un conjunto de datos, se intenta determinar la función continua que mejor se aproxime a los datos, proporcionando una demostración visual de la relación entre los puntos de los mismos.

Se llama recta de regresión a la recta $y = mx + b$ que minimiza el error cuadrático medio ECM

$ECM = \frac{1}{N} \sum_{i=1}^n (y_i - mx_i - b)^2$ los valores de m y b que minimizan ECM se halla derivando la función EMC

$$m = \frac{n \sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j - \sum_{i=1}^n X_i * \sum_{j=1}^n Y_j}{n \sum_{i=1}^n (X_i)^2 - |\sum_{i=1}^n (X_i)|^2} = \bar{Y} - \frac{S_{xy}}{S_x^2} \bar{X}$$

$$b = \frac{S_{xy}}{S_x^2}$$

La recta de regresión será $y = \left(\bar{Y} + \frac{S_{xy}}{S_x^2} \bar{X} \right) x - \frac{S_{xy}}{S_x^2}$; $y = \bar{Y} + \frac{S_{xy}}{S_x^2} (X - \bar{X})$;

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right)$$

14. Ejercicio

Con el fin de relacionar el índice de fluor en el agua (X en ppm) con la tasa de caries (Y en %) se han tomado muestras en 10 ciudades, los datos obtenidos son:

X	1,6	7,8	7,1	2,3	5,8	4,2	7,6	9,8	6,4	7,6
Y	2,0	4,0	5,0	5,5	6,0	6,5	7,1	7,6	8,4	9,3

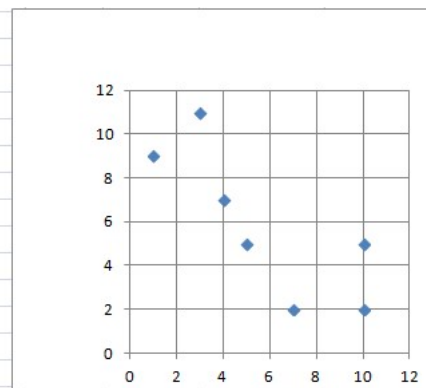
Encontrar la recta que mejor se ajusta a los datos

Solución

$$\bar{X} = \frac{55}{9} = 6,11;$$

$$\bar{Y} = \frac{57}{9} = 6,33$$

X	Y
7	2
1	9
10	2
5	5
4	7
3	11
13	2
10	5
2	14
55,0	57,0
$\bar{X}$ 6,11	$\bar{Y}$ 6,33



La recta de regresión será $y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X})$;

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{137}{9} = 15,22$$

$$S_{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y} = \frac{233}{9} - 6,11 * 6,33 = 25,89 - 38,6763 = -12,7863$$

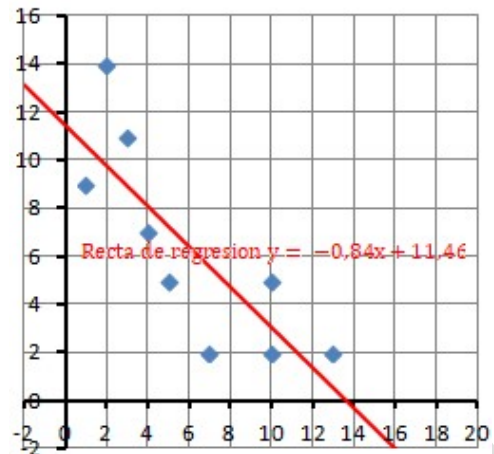
18(25)

X	Y	$(X_i - \bar{X})^2$	$X_i * Y_i$
7	2	1	14
1	9	26	9
10	2	15	20
5	5	1	25
4	7	4	28
3	11	10	33
13	2	47	26
10	5	15	50
2	14	17	28
55,0	57,0	137	233
$\bar{X}$ 6,11	$\bar{Y}$ 6,33	$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$	$\frac{\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j}{N}$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2}(x - \bar{X});$$

$$y = 6,33 + \frac{(-12,7863)}{15,22}(x - 6,11);$$

$$y = 6,33 - 0,84x + 5,13; \bar{y} = -0,84x + 11,46$$



4.2 Error cuadrático medio o varianza residual

Residuo es la diferencia entre el valor observado y el estimado según la recta de regresión $\hat{y}_i$

ECM error cuadrático medio es la media aritmética de los cuadrados de los residuos

$$ECM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

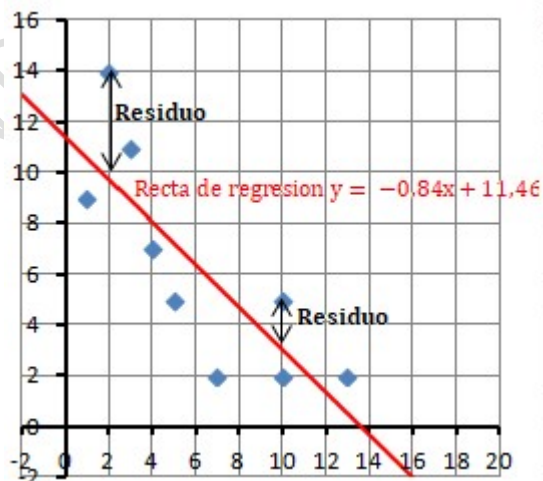
R² Coeficiente de determinación

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} \text{ cuanto mas cercano a uno este}$$

el coeficiente de determinación mas precisa serán las predicciones para cualquier valor de x.

Si R² fuera igual 1 la predicción coincidiría con el valor real y el error sería nulo.

Si R² esta por encima del 0,9, el 90% podremos considerar valido el valor de la predicción



15. Ejercicio

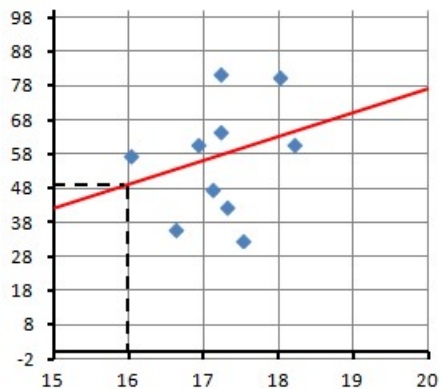
En la tabla que recogen para el periodo 2004-2013 la temperatura media invernal X y el número de días en que el nivel de ozono supero los 0,20ppm (partes por millón):

X	16,00	17,20	18,00	17,20	16,90	17,10	18,20	17,30	17,50	16,60
Y	58	82	81	65	61	48	61	43	33	36

Encontrar la recta que mejor se ajusta a los datos

Solución

X	Y	x^2	y^2	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$X \cdot Y$	y
16,00	58	256	3364	1,44	1,44	928	4
16,60	36	276	1296	0,36	432,64	598	3
16,90	61	286	3721	0,09	17,64	1031	3
17,10	48	292	2304	0,01	77,44	821	3
17,20	82	296	6724	0,00	635,04	1410	3
17,20	65	296	4225	0,00	67,24	1118	3
17,30	43	299	1849	0,01	190,44	744	3
17,50	33	306	1089	0,09	566,44	578	3
18,00	81	324	6561	0,64	585,64	1458	3
18,20	61	331	3721	1,00	17,64	1110	3
172,00	568,0	2962,0	34854,0	3,64	2591,6	9795	-71
		$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$		$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N}$		$\frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N}$	
$\bar{X} 17,20$	$\bar{Y} 56,80$						



$$\bar{X} = \frac{172}{10} = 17,2; \quad \bar{Y} = \frac{568}{10} = 56,8$$

La recta de regresion será $y = \bar{Y} - \frac{S_{xy}}{S_x^2} (x - \bar{X})$; $S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{3,64}{10} = 0,364$

$$S_{xy} = \frac{\sum_i^k \sum_j^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y} = \frac{9795}{10} - 17,20 * 56,8 = 979,5 - 976,96 = 2,54$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X}); y = 56,8 + \frac{2,54}{0,364} (x - 17,2); y = 56,8 + 7x - 120,4;$$

$$y = 7x - 63,6$$

Sustituyo en la recta para un valor de 16 grados $y = 7(16) - 63,6 = 48,4$ días supero el 20ppm

$$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{2591,6}{10} = 259,16$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); ECM = 259,16 \left(1 - \frac{(2,54)^2}{0,364 * 259,16} \right) = 241,436$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; R^2 = \frac{(2,54)^2}{0,364 * 259,16} = 0,06839; R^2 = 6,8\% \text{ La relacion entre variables es muy debil}$$

16. Ejercicio

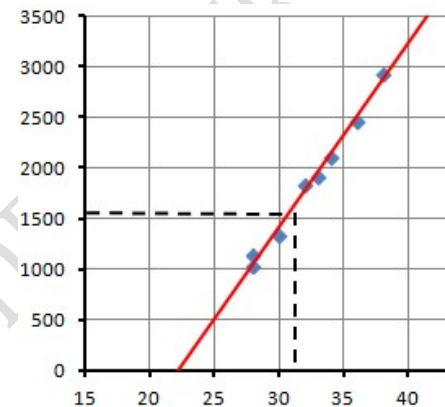
Para analizar la relación entre las semanas de gestación X y el peso del feto Y se ha tomado una muestra de 8 embarazadas:

X	28	30	33	32	28	34	38	36
Y	1150	1350	1925	1850	1040	2125	2950	2475

- Representa los datos en un grafico
- Encontrar la recta de regresion de Y sobre X
- Calcula EMC y en coeficiente de determinación y determina la precision del ajuste
- Que peso tendria un feto de 31 semanas

Solución

X	Y	X^2	Y^2	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$X \cdot Y$	Y
28	1150	784	1322500	19,18	501448,10	32200	-2
30	1350	900	1822500	5,66	258196,10	40500	-3
33	1925	1089	3705625	0,38	4471,60	63525	-4
32	1850	1024	3422500	0,14	66,10	59200	-4
28	1040	784	1081600	19,18	669336,70	29120	-2
34	2125	1156	4515625	2,62	71219,60	72250	-5
38	2950	1444	8702500	31,58	1192180,10	112100	-7
36	2475	1296	6125625	13,10	380528,60	89100	-6
259,00	14865,0	8477,0	30698475,0	91,88	3077446,9	497995	-113
$\bar{X} = 32,38$ $\bar{Y} = 1858,13$		$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{91,88}{8} = 11,485$ $S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{3077446,9}{8} = 384680,86$					
		$S_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y} = \frac{497995}{8} - 32,375 \cdot 1858,13 = 62249,375 - 60156,8 = 2092,575$					



$$\bar{X} = \frac{259}{8} = 32,375; \quad \bar{Y} = \frac{14865}{8} = 1858,13$$

La recta de regresion será $y = \bar{Y} - \frac{S_{xy}}{S_x^2} (x - \bar{X})$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{91,88}{8} = 11,485; \quad S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{3077446,9}{8} = 384680,86$$

$$S_{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y} = \frac{497995}{8} - 32,375 \cdot 1858,13 = 62249,375 - 60156,8 = 2092,575$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X}); y = 1858,13 + \frac{2092,575}{11,485} (x - 32,375); y = 1858,13 + 182,2x - 5898,75;$$

$$y = 182,2x - 4040,62$$

Sustituyo en la recta para un valor de 31 semanas $y = 182,2 \cdot 31 - 4040,62 = 1607,58$

$$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{3077446,9}{8} = 384680,86$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); ECM = 384680,86 \left(1 - \frac{(2092,575)^2}{11,485 \cdot 384680,86} \right) = 3411,97$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; R^2 = \frac{(2092,575)^2}{11,485 \cdot 384680,86} = 0,9911; R^2 = 99,11\%$$

17. Ejercicio

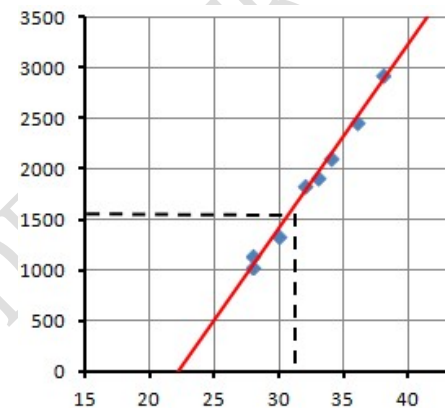
Para analizar la relación entre las semanas de gestación X y el peso del feto Y se ha tomado una muestra de 8 embarazadas:

X	28	30	33	32	28	34	38	36
Y	1150	1350	1925	1850	1040	2125	2950	2475

- Representa los datos en un grafico
- Encontrar la recta de regresion de Y sobre X
- Calcula EMC y en coeficiente de determinación y determina la precision del ajuste
- Que peso tendria un feto de 31 semanas

Solución

X	Y	X ²	Y ²	(X _i - $\bar{X}$) ²	(Y _i - $\bar{Y}$) ²	X*Y	y
28	1150	784	1322500	19,18	501448,10	32200	-2
30	1350	900	1822500	5,66	258196,10	40500	-3
33	1925	1089	3705625	0,38	4471,60	63525	-4
32	1850	1024	3422500	0,14	66,10	59200	-4
28	1040	784	1081600	19,18	669336,70	29120	-2
34	2125	1156	4515625	2,62	71219,60	72250	-5
38	2950	1444	8702500	31,58	1192180,10	112100	-7
36	2475	1296	6125625	13,10	380528,60	89100	-6
259,00	14865,0	8477,0	30698475,0	91,88	3077446,9	497995	-113
$\bar{X} = 32,38$ $\bar{Y} = 1858,13$		$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$ $S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N}$ $S_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) f_i}{N}$					



$$\bar{X} = \frac{259}{8} = 32,375; \quad \bar{Y} = \frac{14865}{8} = 1858,13$$

La recta de regresion será $y = \bar{Y} + \frac{S_{xy}}{S_x^2}(x - \bar{X})$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{91,88}{8} = 11,485; \quad S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{3077446,9}{8} = 384680,59$$

$$S_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y} = \frac{497995}{8} - 32,375 * 1858,13 = 62249,375 - 60156,8 = 2092,575$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2}(x - \bar{X}); y = 1858,13 + \frac{2092,575}{11,485}(x - 32,375); y = 1858,13 + 182,2x - 5898,75;$$

$$y = 182,2x - 4040,62$$

Sustituyo en la recta para un valor de 31 semanas $y = 182,2 * 31 - 4040,62 = 1607,58$

$$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{3077446,9}{8} = 384680,86$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); ECM = 384680,86 \left(1 - \frac{(2092,575)^2}{11,485 * 384680,86} \right) = 3411,97$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; R^2 = \frac{(2092,575)^2}{11,485 * 384680,86} = 0,9911; R^2 = 99,11\%$$

5 Regresión y correlación

El coeficiente de correlación entre las variable X, Y es $r = \sqrt{R^2} = \frac{S_{xy}}{S_x S_y}$

- Su valor estará comprendido entre 1 y -1
- Cuando está próximo a 0 no hay relación lineal entre las variables
- Si es positiva la relación entre las variables es directa (para mayor valor de X mayor Y)
- Si es negativa la relación entre las variables es inversa (para mayor valor X menor de Y)

18. Ejercicio

La productividad de un cultivo Y viene determinada por las horas de regadío X. Para analizar la relación se ha tomado una muestra de 6 fincas.

a) Representa los datos en un grafico

b) Encontrar la recta de regresion de Y sobre X

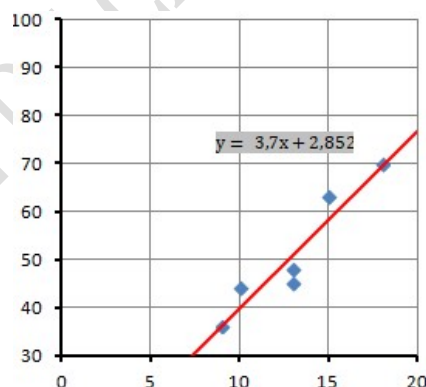
c) Calcula R^2 y r

d) Es fiable el ajuste linea

X	9	10	13	15	18	13
Y	36	44	48	63	70	45

Solución

X	Y	X^2	Y^2	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$X \cdot Y$	y
9	36	81	1296	16,00	225,00	324	36
10	44	100	1936	9,00	49,00	440	40
13	48	169	2304	0,00	9,00	624	51
15	63	225	3969	4,00	144,00	945	58
18	70	324	4900	25,00	361,00	1260	69
13	45	169	2025	0,00	36,00	585	51
		0	0				
		0	0				
78,00	306,0	1068,0	16430,0	54,00	824,0	4178	
		$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$		$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N}$		$\frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N}$	
$\bar{X}$ 13,00	$\bar{Y}$ 51,00						



$$\bar{X} = \frac{78}{6} = 13; \quad \bar{Y} = \frac{306}{6} = 51$$

La recta de regresion será $y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X})$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{54}{6} = 9; \quad S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{824}{6} = 137,33$$

$$S_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X} \bar{Y} = \frac{4178}{6} - 13 \cdot 51 = 696,33 - 663 = 33,33$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X}); y = 51 + \frac{33,33}{9} (x - 13); y = 51 + 3,7x - 48,148;$$

$$y = 3,7x + 2,852$$

$$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{824}{6} = 137,333$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); ECM = 137,333 \left(1 - \frac{(33,33)^2}{9 \cdot 137,33} \right) = 13,8787$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; R^2 = \frac{(33,33)^2}{9 \cdot 137,33} = 0,8989; R^2 = 89,89\%$$

$$R^2 = 0,8989; r = \sqrt{R^2} = \frac{S_{xy}}{S_x S_y} = \frac{33,33}{3 \cdot 11,72}; r = \frac{33,33}{35,157} = 0,95 \text{ El ajuste lineal es muy bueno}$$

19. Ejercicio

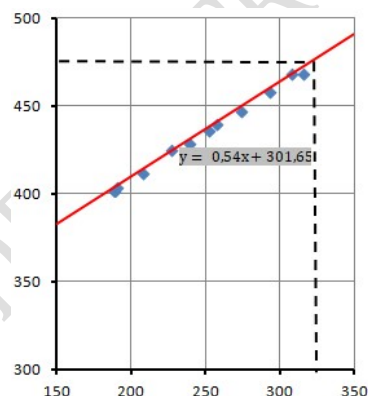
Se desea hacer una previsión de ventas Y (en miles de euros) en un país en función de la renta nacional X (en millones de euros). Para hacer el estudio parte de los datos obtenidos en otro mercado.

X	189	190	208	227	239	252	257	274	293	308	316
Y	402	404	412	425	429	436	440	447	458	469	469

- Representa los datos en un gráfico
- Encontrar la recta de regresión de Y sobre X
- Calcula R^2 y r
- si en el 2001 la renta nacional del país va a ser de 325 millones ¿cual será la predicción de ventas?

Solución

	X	Y	X^2	Y^2	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$X \cdot Y$	Y
1990	189	402	35721	161604	3754,01	1125,60	75978	702
1991	190	404	36100	163216	3632,47	995,40	76760	706
1992	208	412	43264	169744	1786,75	554,60	85696	
1993	227	425	51529	180625	541,49	111,30	96475	
1994	239	429	57121	184041	127,01	42,90	102531	
1995	252	436	63504	190096	2,99	0,20	109872	
1996	257	440	66049	193600	45,29	19,80	113080	954
1997	274	447	75076	199809	563,11	131,10	122478	1017
1998	293	458	85849	209764	1825,85	504,00	134194	1087
1999	308	469	94864	219961	3332,75	1118,90	144452	1142
2000	316	469	99856	219961	4320,43	1118,90	148204	1172
	2753,00	4791,0	708933,0	2092421,0	19932,18	5722,7	1209720	
	$\bar{X} = 250,27$	$\bar{Y} = 435,55$						



$$\bar{X} = \frac{2753}{11} = 250,27; \quad \bar{Y} = \frac{4791}{11} = 435,55$$

La recta de regresión será $y = \bar{Y} + \frac{S_{xy}}{S_x^2}(x - \bar{X})$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{19932,18}{11} = 1812,02; \quad S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{5722,7}{11} = 520,25$$

$$S_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^h f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y} = \frac{1209720}{11} - 250,27 \cdot 435,55 = 109974,55 - 109005,099 = 969,4515$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2}(x - \bar{X}); \quad y = 435,55 + \frac{969,4515}{1812,02}(x - 250,27); \quad y = 435,55 + 0,54x - 133,9;$$

$$y = 0,54x + 301,65$$

$$S_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 f_i}{N} = \frac{520,25}{11} = 47,3$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); \quad ECM = 520,25 \left(1 - \frac{(969,4515)^2}{1812,02 \cdot 520,25} \right) = 1,58$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; \quad R^2 = \frac{(969,4515)^2}{1812,02 \cdot 520,25} = 0,996958; \quad R^2 = 99,7\%$$

$$R^2 = 0,996958; \quad r = \sqrt{R^2} = \frac{S_{xy}}{S_x S_y} = \frac{969,4515}{42,57 \cdot 22,81};$$

$$r = 0,9983 = 0,95 \text{ El ajuste lineal es muy bueno}$$

$$y = 0,54x + 301,65; \text{ para 2001 y una renta de 325 millones; } y = 0,54 \cdot 325 + 301,65;$$

$$y = 477,15 \text{ miles de euros en ventas}$$

20. Ejercicio

La productividad de un cultivo Y viene determinada por las horas de regadío X. Para analizar la relación se ha tomado una muestra de varias fincas.

- Representa los datos en un gráfico
- Encontrar la recta de regresión de Y sobre X
- Calcula R^2 y r
- Es fiable el ajuste lineal

[25,35)	[35,45)	[45,55)	[55,65)
30	40	50	60
6	3	0	0
4	6	1	0
0	1	5	2
0	0	3	7

Solución

X_i	Y_j	f_i	$X_i f_i$	$Y_j f_j$	$X^2 f$	$Y^2 f$	$(X_i - \bar{X})^2 f_i$	$(Y_j - \bar{Y})^2 f_j$	$X_i f_i * Y_j f_j$
20	30	6	120	180,00	2400	5400	1305,375	1350	3600
20	40	3	60	120,00	1200	4800	652,6875	75	2400
20	50	0	0	0,00	0	0	0	0	0
20	60	0	0	0,00	0	0	0	0	0
30	30	4	120	120,00	3600	3600	90,25	900	3600
30	40	6	180	240,00	5400	9600	135,375	150	7200
30	50	2	60	100,00	1800	5000	45,125	50	3000
30	60	1	30	60,00	900	3600	22,5625	225	1800
40	30	0	0	0,00	0	0	0	0	0
40	40	1	40	40,00	1600	1600	27,5625	25	1600
40	50	5	200	250,00	8000	12500	137,8125	125	10000
40	60	2	80	120,00	3200	7200	55,125	450	4800
50	30	0	0	0,00	0	0	0	0	0
50	40	0	0	0,00	0	0	0	0	0
50	50	3	150	150,00	7500	7500	697,6875	75	7500
50	60	7	350	420,00	17500	25200	1627,9375	1575	21000
		40	1390	1800	53100	86000	4797,5	5000	66500
$\bar{X}$	34,75	$\bar{Y}$	45				$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N}$		$S_y^2 = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j}{N}$
									$\frac{\sum_i^k \sum_j^h f_{ij} X_i Y_j}{N}$

$$\bar{X} = \frac{1390}{40} = 34,75; \quad \bar{Y} = \frac{1800}{40} = 45$$

La recta de regresión será $y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X})$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{N} = \frac{4797,5}{40} = 119,938;$$

$$S_y^2 = \frac{\sum_{j=1}^n (Y_j - \bar{Y})^2 f_j}{N} = \frac{5000}{40} = 125$$

$$S_{xy} = \frac{\sum_i^k \sum_j^h f_{ij} X_i Y_j}{N} - \bar{X}\bar{Y} = \frac{66500}{40} - 34,75 * 45 = 1662,5 - 1563,75 = 98,75$$

$$y = \bar{Y} + \frac{S_{xy}}{S_x^2} (x - \bar{X}); y = 45 + \frac{98,75}{119,938} (x - 34,75); y = 45 + 0,8233x - 28,611;$$

$$y = 0,8233x + 16,389$$

$$ECM = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right); ECM = 125 \left(1 - \frac{(98,75)^2}{119,938 * 125} \right) = 43,695$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}; R^2 = \frac{(98,75)^2}{119,938 * 125} = 0,65044; R^2 = 65,04\%$$

$$R^2 = 0,65044; r = \sqrt{R^2} = \frac{S_{xy}}{S_x S_y} = \frac{98,75}{10,95 * 11,18};$$

$r = 0,8066 = 0,81$ existe una correlación positiva

