

Se ha realizado una encuesta a 10 alumnos para estudiar cómo influyen las horas que pasan jugando con el ordenador a los resultados escolares

X (HORAS)	Y (SUSPENSOS)
0	0
0	1
3	5
2	3
1	0
1	2
1	1
2	2
0	0
1	2

En este estudio intervienen dos variables
 X: Horas diarias (de media) que juegan al ordenador
 Y: Número de suspensos en la segunda evaluación

Por esa razón, decimos que tenemos una
VARIABLE ESTADÍSTICA BIDIMENSIONAL

De cada uno de esos 10 alumnos, estamos estudiando dos variables, y resulta comprensible que queramos saber si influyen entre sí

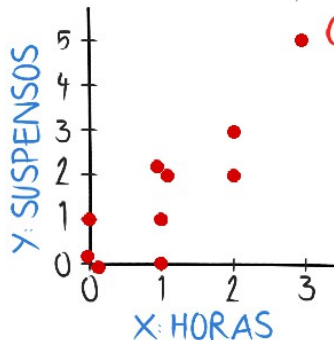
Para poder trabajar con esos datos, recurriremos a una tabla de doble entrada, en cuyas celdas reflejaremos las frecuencias

	X: HORAS				
	0	1	2	3	f _i
0	2	1			3
1	1	1			2
2		2	1		3
3			1		1
4					0
5				1	1
f _j	3	4	2	1	10

Acabado el recuento, ya tenemos la frecuencia absoluta de los pares de valores asociados a cada alumno concreto

Ya se aprecia una tendencia. Vamos a trasladarlo a una gráfica

Trasladamos la información a la NUBE DE PUNTOS o DIAGRAMA DE DISPERSIÓN



• Cuando tenemos datos repetidos ($f > 1$) pintamos varios puntos cercanos o trazamos un punto más grueso

Cuanto más alineados veamos los puntos, mayor dependencia lineal se aprecia

ESTADÍSTICA 4

		X: HORAS							
		0	1	2	3	f_i	$y_j \cdot f_i$	$y_j^2 \cdot f_i$	
Y: SUSPENSOS	0	2	1			3	0	0	
	1	1	1			2	2	2	
	2		2	1		3	6	12	
	3			1		1	3	9	
	4					0	0	0	
	5				1	1	5	25	
		f_i	3	4	2	1	10	16	48
		$x_i \cdot f_i$	0	4	4	3	11	TOTALES	
		$x_i^2 \cdot f_i$	0	4	8	9	21		

Recuperemos nuestra tabla para comenzar el estudio estadístico

Primero estudiamos las variables por separado

$$\bar{Y} = 16/10 = 1.6$$

$$V_Y = 48/10 - 1.6^2 = 2.24$$

$$S_Y = \sqrt{2.24} \approx 1.50$$

$$\bar{X} = 11/10 = 1.1$$

$$V_X = 21/10 - 1.1^2 = 0.89$$

$$S_X = \sqrt{0.89} \approx 0.94$$

Ahora pasemos al estudio con junto

Considerando que: $2 = f_{12}$

$$Y \text{ que } \sum_{i=1}^3 x_i = x_1 + x_2 + x_3$$

SUMATORIA
DESDE $i=1$ HASTA 3 DE...

$$\text{La COVARIANZA es } \text{Cov}(xy) = \frac{\sum_{i=1}^n \sum_{j=1}^m x_i \cdot y_j \cdot f_{ij}}{N} - \bar{X} \cdot \bar{Y}$$

Veamos cómo se calcula en la práctica:

$$\text{Cov}(xy) = \frac{2 \cdot 0 \cdot 0 + 1 \cdot 0 \cdot 1 + 1 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot 1 + 2 \cdot 1 \cdot 2 + 1 \cdot 2 \cdot 2 + 1 \cdot 2 \cdot 3 + 1 \cdot 3 \cdot 5}{10} - 1.1 \cdot 1.6$$

$$\text{Cov}(xy) = 3 - 1.76 = 1.24$$

Y el COEFICIENTE DE CORRELACIÓN es

$$r = \frac{\text{Cov}(xy)}{S_X \cdot S_Y} = \frac{\text{Cov}(xy)}{\sqrt{V_X \cdot V_Y}} = \frac{1.24}{\sqrt{1.9936}} \approx 0.878$$

El coeficiente de correlación siempre tiene que dar un valor entre -1 y 1. Si nos da un valor fuera de ese intervalo significa que hemos cometido algún error en los cálculos

Si "r" vale 1 o -1, eso indica que hay dependencia lineal entre las variables (una función lineal convierte una en la otra); una de las variables determina exactamente a la otra

Si r está cercano a 0, significa que la correlación (la medida de la dependencia lineal) es muy débil

Si r está cerca de 1, como en nuestro caso (0'878), la correlación es fuerte y directa (cuanto mayor es una variable la otra también lo es)

Si r está cerca de -1, la correlación lineal es fuerte e inversa

Todo ello lo vamos a ver más claro con las RECTAS DE REGRESIÓN

La recta de regresión de Y sobre X se calcula con la fórmula:

$$y - \bar{y} = \frac{\text{Cov}(xy)}{V_x} (x - \bar{x})$$

En nuestro caso tendríamos $y - 1'6 = \frac{1'24}{0'89} (x - 1'1)$

Despejando $y = 1'3933x + 0'06742$

Sirve para predecir resultados de la variable Y a partir de valores de X, siempre que la correlación sea fuerte

¿Cuántos suspensos podemos esperar de un alumno que jugase una hora y media diarias con el ordenador?

$y = 1'3933 \cdot 1'5 + 0'06742 = 2'15737$ Podemos esperar dos suspensos

Y la recta de regresión de X sobre Y será:

$$x - \bar{x} = \frac{\text{Cov}(xy)}{V_y} (y - \bar{y})$$

En nuestro caso tendremos $x - 1'1 = \frac{1'24}{2'24} (y - 1'6)$

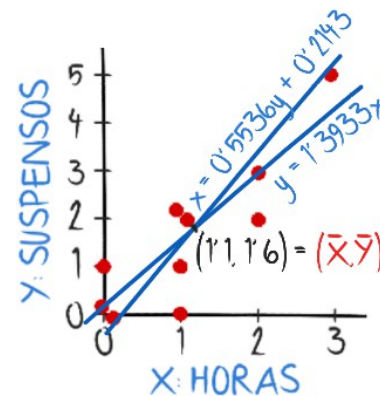
Despejando $x = 0'5536y + 0'2143$

Sirve para predecir resultados de la variable X a partir de valores de Y, siempre que la correlación sea fuerte

¿Cuántas horas diarias podemos esperar que esté jugando al ordenador un alumno con 4 suspensos?

$x = 0'5536 \cdot 4 + 0'2143 = 2'4287$ Podemos esperar que juegue casi dos horas y media diarias

Y si representamos las dos rectas obtenidas sobre el diagrama de dispersión, observamos que son muy parecidas y se ajustan bastante a la nube de puntos



Eso ocurre porque la correlación es muy fuerte. Si no es así, las rectas son casi perpendiculares. Las pendientes son positivas, ya que la correlación es directa. Si fuese inversa, como habría pasado al tratar los aprobados, las pendientes serían negativas como la tendencia de la nube

Q jQ : Como ya vimos en el vídeo anterior, los resultados estadísticos pueden resultar muy engañosos si no van acompañados de una profunda reflexión

Si realizásemos un estudio sobre las notas de Lengua y Matemáticas, seguramente descubriríamos que existe una correlación fuerte: casi todos los alumnos con buenas notas en Lengua tienen buenas notas en Matemáticas y viceversa

Resulta absurdo pensar que las notas de una asignatura influyan en la otra, es decir, que saber mucho de Matemáticas me garantice una buena nota en Lengua.

Realmente, estas dos variables vienen determinadas por una tercera, que es "las horas de estudio" que invierte cada alumno en sus exámenes

Por regla general, un buen estudiante dedicará bastantes horas a cada asignatura, mientras que otro alumno más perezoso dedicará pocas a ambas

Cuando haya terceras variables que influyan, los análisis bidimensionales no son los más apropiados. En un futuro, tendrás la oportunidad de descubrir cómo se mide la influencia de varias variables entre sí