ESTADÍSTICA BIDIMENSIONAL

ESTUDIO CONJUNTO DE DOS VARIABLES ESTADÍSTICAS.
(Dependencia lineal)

En muchas ocasiones no sólo se desea estudiar una característica en particular si no que se podría estar interesado en analizar como dos características pueden estar relacionadas como por ejemplo horas de sueño y resultados académicos; consumo de un fármaco y horas de vigilia; tiempo de dedicación a redes sociales y calificación en lengua; dinero destinado a publicidad de un producto y número de ventas, etc.

Para ello tomamos los datos como pares de valores, así llamando X a la primera característica e Y a la segunda, obtendremos pares de valores.

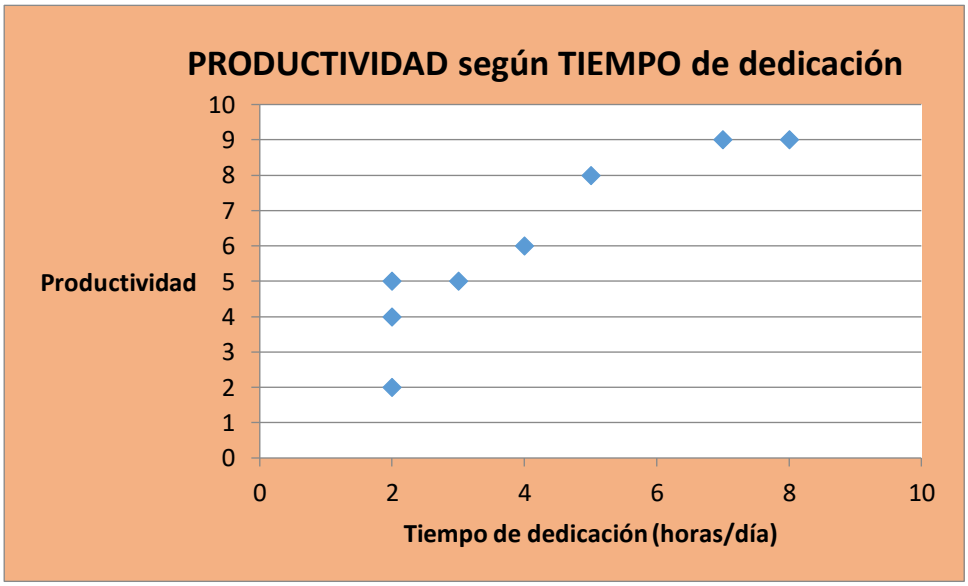
Ejemplo 1:

En una empresa, con motivo de establecer turnos, se desea conocer si existe relación entre el tiempo dedicado a una tarea y la efectividad/productividad del trabajo efectuado. Para ello se han tomado datos en 8 empleados obteniendo las siguientes respuestas (2,2),(2,4),(2,5),(3,5), (4,6),(5,8),(7,9),(8,9).

Una vez obtenidos los datos, procedemos a colocarlos en una tabla y representarlos gráficamente, realizamos un diagrama de dispersión y obtenemos una "nube de puntos"[1].

A partir de los datos del ejemplo, siendo X el tiempo trabajado en horas e Y la productividad medida en una escala de 0 a 10, los datos obtenidos y el diagrama de dispersión correspondiente, obtenemos la nube de puntos

x	y
2	2
2	4
2	5
3	5
4	6
5	8
7	9
8	9



Observando la nube de puntos sería posible hacer conjeturas sobre la posible dependencia funcional[2]. Pero nosotros vamos a analizar la existencia de dependencia lineal mediante el COEFICIENTE DE CORELACIÓN DE PEARSON[4].

Para los cálculos necesarios elaboramos una tabla que nos simplificará la tarea.

	x	y	X^2	Y^2	x·y
	2	2	4	4	4
	2	4	4	16	8
	2	5	4	25	10
	3	5	9	25	15
	4	6	16	36	24
	5	8	25	64	40
	7	9	49	81	63
	8	9	64	81	72
∑	33	48	175	332	236

Hallamos los parámetros necesarios:

	Tiempo dedicación	Productividad
Media	4,13	6
Varianza	4,86	5,5
D Típica	2,2	2,345
C.V.	0,53	0,391

Covarianza[(3)]	4,75
Coeficiente de correlación de Pearson	0,192

Obtenido el coeficiente de correlación lineal, al ser r=0,192< 0,75, concluimos que no existe dependencia lineal entre las variables y daríamos la conclusión pertinente

El problema del ejemplo y su presentación, quedaría así:

Enunciado:

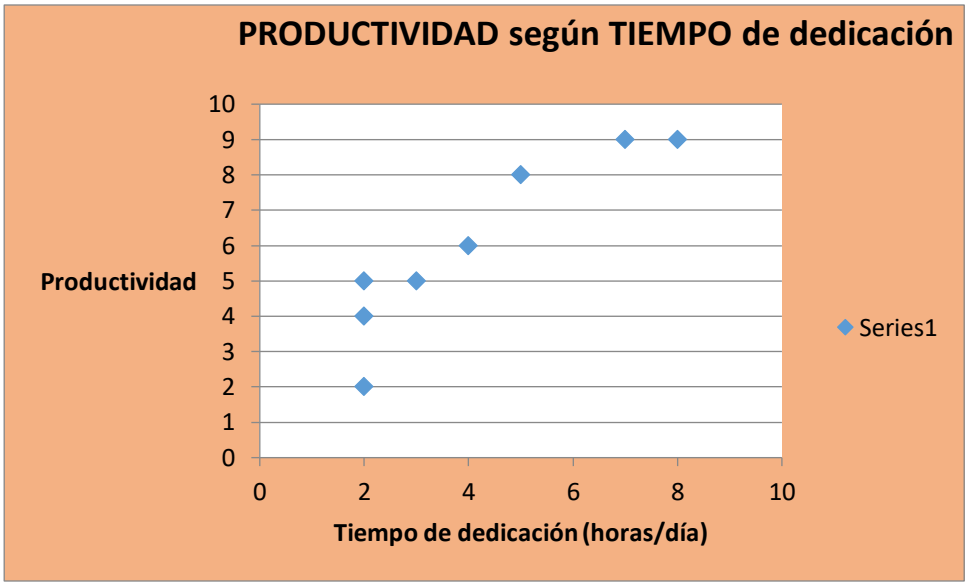
En una empresa, con motivo de establecer turnos, se desea conocer si existe relación entre el tiempo dedicado a una tarea y la efectividad/productividad del trabajo efectuado. Para ello se han tomado datos en 8 empleados obteniendo las siguientes respuestas (2,2),(2,4),(2,5),(3,5), (4,6),(5,8),(7,9),(8,9).

Estudia la correlación entre ambas variables e interpreta el resultado.

Solución:

Siendo X el tiempo trabajado en horas e Y la productividad medida en una escala de 0 a 10, los datos obtenidos y la nube de puntos correspondientes son:

x	y
2	2
2	4
2	5
3	5
4	6
5	8
7	9
8	9



Hallamos los parámetros:

	x	y	x^2	y^2	x·y
	2	2	4	4	4
	2	4	4	16	8
	2	5	4	25	10
	3	5	9	25	15
	4	6	16	36	24
	5	8	25	64	40
	7	9	49	81	63
	8	9	64	81	72
Sumas	33	48	175	332	236

	Tiempo dedicación	Productividad
Media	4,13	6
Varianza	4,86	5,5
D Típica	2,2	2,345
C.V.	0,53	0,391

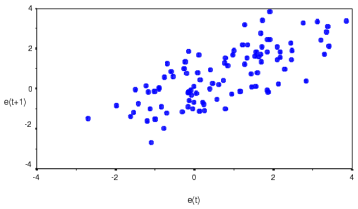
Covarianza 4,75

Coeficiente de correlación de Pearson 0,192

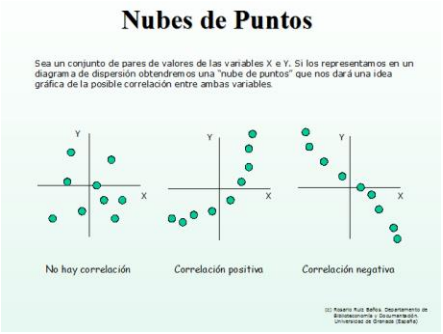
CONCLUSIÓN:
El coeficiente de correlación r=0,192 <0,75, indica que no hay correlación lineal.
De los datos obtenidos y de la nube de puntos podría interpretarse que por encima de 7 horas de trabajo no hay indicios de aumento en la productividad.

Aclaraciones.

(1) **"nube de puntos"**.- Es la imagen obtenida al representar los pares de valores en los ejes cartesianos. Cuando el número de observaciones es alto semeja a una nube.



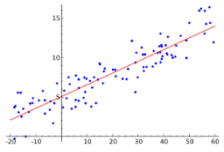
A veces podemos hacernos una idea de la existencia de correlación con tan solo observar la nube de puntos.



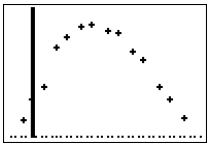
En esta imagen (de Rosario Ruíz) observamos cómo en el primer caso no hay evidencia de correlación. En el segundo caso parece que las variables si están relacionadas, correlación positiva puesto que al crecer X crece Y. En el tercer caso también parece que hay correlación, esta vez negativa puesto que al crecer X decrece Y.

Para hacer la gráfica en la hoja de cálculo:
1º seleccionar los datos de la tabla (las dos columnas, X e Y)
2º Click en Insertar/gráfico/dispersión
3º Entrando en el menú contextual del gráfico (click en el gráfico) puedes poner el color de fondo, los rótulos de los ejes, título, etc.; es decir, modificarlo a tu gusto.
Después es fácil copiar y pegar en el documento de texto.

(2) Dependencia funcional entre dos variables.- se puede ajustar una expresión algebraica que liga ambas variables. Es muy útil cuando esto puede hacerse pues permite predecir el valor que tomara una de las variables al variar la otra. La relación funcional entre las dos variables puede ser de muchos tipos: lineal, cuadrática, logarítmica, exponencial.



Por ejemplo, la nube de puntos correspondiente a estas dos variables parece aproximarse a una recta.



En este otro caso parece que la relación es cuadrática (la forma de la nube se parece más a una parábola)

(3) **Covarianza**.- Este valor da una idea del grado de variación conjunta de las variables con respecto de sus medias. Es la media aritmética de los productos de las desviaciones de cada una de las variables con respecto de la media. Se simboliza mediante COV(X,Y) ó S_{xy} (covarianza muestral) o σ_{xy} (covarianza poblacional) y se calcula mediante:

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})n_i}{n} = \frac{\sum x_i y_i n_i}{n} - \bar{x}\bar{y}$$

Puede tener signo positivo o negativo según la correlación sea positiva o negativa.

(4) **Coeficiente de correlación de Pearson**.- Este coeficiente mide el grado de dependencia lineal entre dos variables. Se designa por r (para las muestras) ó por ρ (se lee ro, para las poblaciones). Se calcula mediante:

$$r = \frac{S_{xy}}{S_x S_y}$$ se cumple que -1 < r < 1.

Generalmente para |r|<0,75 se descarta la correlación lineal; cuanto más se acerque el valor a 1 ó a -1, la dependencia lineal es mayor.

Una vez constatada la correlación se procedería a hallar la recta de regresión de Y sobre X que nos permitirá predecir valores de Y para diferentes valores de X

En el caso de correlación lineal,

la recta de regresión de Y sobre X, $y - \bar{y} = \frac{\sigma_{XY}}{\sigma^2_X}(x - \bar{x})$

la recta de regresión de X sobre Y sería $x - \bar{x} = \frac{\sigma_{XY}}{\sigma^2_Y}(y - \bar{y})$

Ejemplo 2:

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
Σ	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

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

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

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

- Coeficiente de correlación lineal de Pearson:
  r= $\frac{\sigma_{xy}}{\sigma_x \sigma_y}$ ≅ 0 ,9953
- Calculando la recta de regresión de nº de vehículos sobre nº de accidentes (recta de regresión de y sobre x)

$$y - \bar{y} = \frac{\sigma_{XY}}{\sigma^2_X}(x - \bar{x}),$$

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

  para x=6, y = 1,52·6+6,9= 16,02

  Podemos suponer que ayer circulaban unos 16 coches a más de 120 km/h

- El coeficiente de correlación de Pearson, r≅ 0,995, próximo a 1. La correlación lineal es fuerte por tanto la predicción se considera buena.