Unidad 4
RELACIÓN ENTRE DOS VARIABLES

Al realizar el estudio de una población o muestra puede que nos interese observar más de un carácter o variable, dando lugar a las variables estadísticas bidimensionales o multidimensionales. En esta unidad se estudiará la relación de tipo estadístico entre dos variables. Como consecuencia, cada individuo tendrá asociado un par de valores (x, y).

Ejemplo 1: Se ha solicitado a un grupo de 50 individuos información sobre el número de horas dedicadas diariamente a dormir y a ver la televisión. Con las respuestas obtenidas se ha elaborado la siguiente tabla:

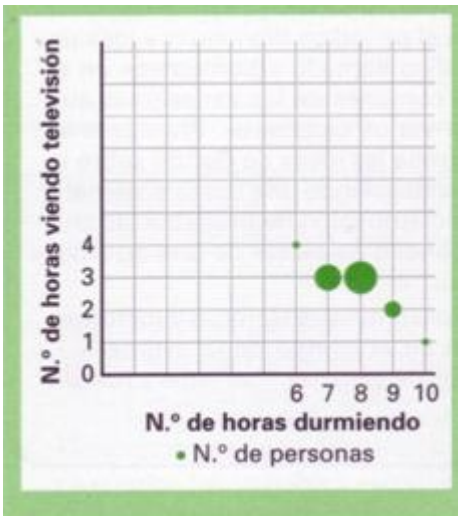
Número de horas durmiendo	**Número de horas viendo televisión**	**Número de personas**
6	4	3
7	3	16
8	3	20
9	2	10
10	1	1

En este ejemplo estamos estudiando dos variables, "número de horas durmiendo" y "número de horas viendo la televisión", de una muestra de 50 individuos. Dicha información aparece recogida en forma de tabla.

1.- DIAGRAMAS DE DISPERSIÓN O NUBE DE PUNTOS

Un **diagrama de dispersión o nube puntos** consiste en un sistema de ejes coordenados que representan los posibles valores de una variable, en el que se reflejan los distintos pares de puntos (x_i , y_i) observados. Estos diagramas son muy útiles para indicarnos si existe o no relación entre las dos variables y medir el sentido y la intensidad de dicha relación. Fundamentalmente hay dos formas de hacer la representación, en la primera a los puntos se les da un grosor proporcional a la

frecuencia con que aparecen y en la segunda se colocan tantos puntos como indica la frecuencia.



(A la hora de representar los datos en los ejes, en principio, no es fundamental el orden en las variables).

2.- TABLAS DE FRECUENCIAS. DISTRIBUCIONES MARGINALES Y CONDICIONADAS.

Para la organización de los datos en el estudio de dos variables estadísticas X e Y con sus correspondientes modalidades o valores $x_1, x_2, x_3, \ldots, x_k$ e $y_1, y_2, y_3, \ldots, y_m$, utilizamos unas **tablas de doble entrada**.

		Y							
		y_1	y_2	y_3	...	y_p	...	y_m	
X	x_1	n_{11}	n_{12}	n_{13}	...	n_{1p}	...	n_{1m}	n_{1*}
	x_2	n_{21}	n_{22}	n_{23}	...	n_{2p}	...	n_{2m}	n_{2*}
	...	...	...	...	...	...	...	...	...
	x_s	n_{s1}	n_{s2}	n_{s3}	...	n_{sp}	...	n_{sm}	n_{s*}
	...	...	...	...	...	...	...	...	...
	x_k	n_{k1}	n_{k2}	n_{k3}	...	n_{kp}	...	n_{km}	n_{k*}
		n_{*1}	n_{*2}	n_{*3}	...	n_{*p}	...	n_{*m}	n

donde:

$\mathbf{n_{ij}}$: indica el número de individuos que presentan la modalidad x_i de la variable X y la modalidad y_j de la variable Y.

$\mathbf{n_{i*}}$: indica el número de individuos que presentan la modalidad x_i .

$\mathbf{n_{*j}}$: indica el número de individuos que presentan la modalidad y_j .

n : indica el número de individuos de la población o muestra.

En el caso de que alguna variable sea **agrupada en intervalos**, aparecerán las distintas clases o intervalos en los que se haya agrupado, y las frecuencias corresponden al número de observaciones que hay en cada intervalo

En el **ejemplo 1,** si llamamos X a la variable "número de horas viendo la televisión" e Y a la variable "número de horas durmiendo", la tabla de doble entrada sería:

		Y					
		6	7	8	9	10	
X	1					1	1
	2				10		10
	3		16	20			36
	4	3					3
		3	16	20	10	1	50

En este caso es más útil presentar los datos en una tabla simple, tal como se hizo al principio, ya que aparecen pocos pares de valores distintos.

Para la organización de los datos en el estudio de dos caracteres cualitativos se utilizan tablas de doble entrada que se llaman **tablas de contigencia.**

Ejemplo 2: En el curso de 2º de Bachillerato hay un total de 100 alumnos de los que conocemos:

	Usan gafas	**No usan gafas**	
Alumnos	12	28	40
Alumnas	18	42	60
	30	70	100

Dicha tabla nos muestra la siguiente información:

- De un total de 40 alumnos, 12 usan gafas y 28 no.
- De un total de 60 alumnas, 18 usan gafas y 42 no.

Ejemplo 3:

Consideremos los datos que se han obtenido al estudiar las variables X= número de goles marcados e Y= número de goles recibidos, en 40 partidos jugados por el equipo campeón de la liga de fútbol sala:

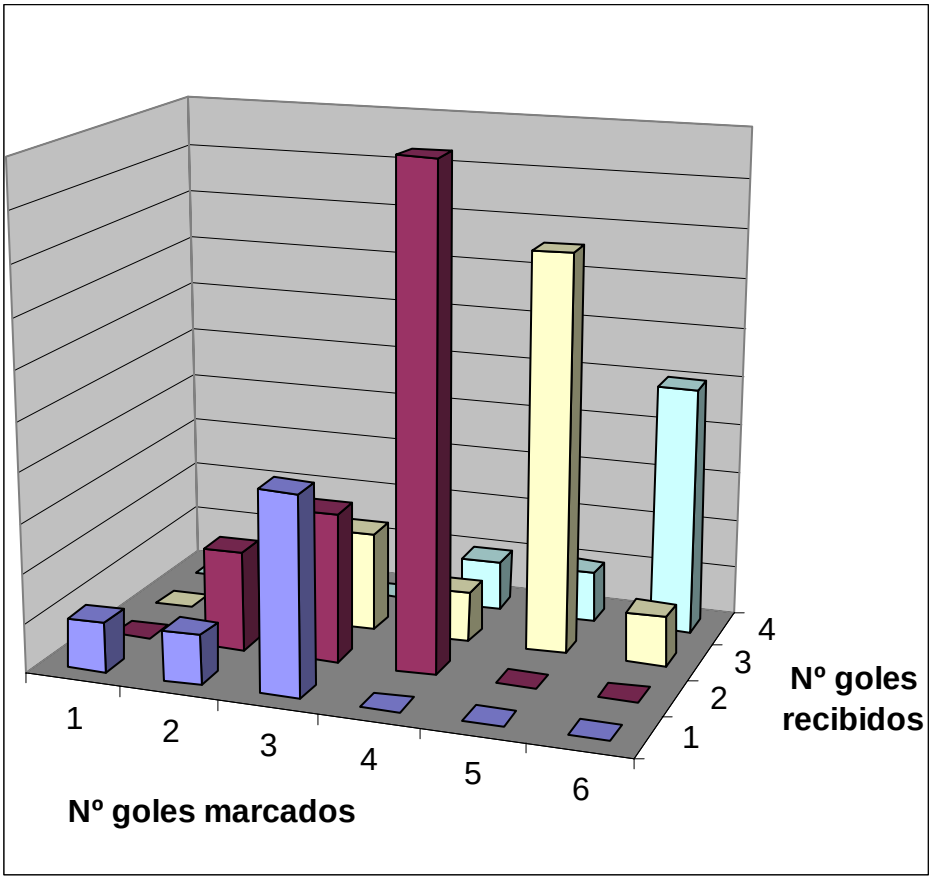
(5, 4)	(4, 2)	(6, 3)	(4, 4)	(3, 2)
(6, 4)	(3, 1)	(4, 2)	(4, 2)	(6, 4)
(4, 2)	(5, 3)	(3, 1)	(2, 2)	(4, 3)
(3, 1)	(4, 2)	(5, 3)	(5, 3)	(4, 2)
(3,3)	(1, 1)	(4, 2)	(5, 3)	(3, 2)
(5, 3)	(6, 4)	(4, 2)	(5, 3)	(2, 1)
(3, 2)	(6, 4)	(5, 3)	(4, 2)	(4, 2)
(3, 3)	(3, 1)	(2, 2)	(6, 4)	(5, 3)

Vamos a organizar los datos.

- Construimos una tabla con tantas columnas coma valores tome X y con tantas filas como valores tome Y.
- Si observamos los datos, X toma los valores 1, 2, 3, 4, 5 y 6, e Y toma los valores 1, 2, 3 y 4. En este caso, la tabla constará de 6 columnas y 4 filas.
- Contamos el número de veces que se repite cada par de valores en la distribución (**frecuencia absoluta**) y lo anotamos en la casilla correspondiente. Así, por ejemplo, se observa que el (5, 4) aparece una sola vez; el (4, 2) diez veces, y el (6, 1) ninguna.

		X						
Y		1	2	3	4	5	6	
	1	1	1	4	0	0	0	6
	2	0	2	3	10	0	0	15
	3	0	0	2	1	8	1	12
	4	0	0	0	1	1	5	**7**
		1	3	9	12	9	6	**40**

Para distribuciones bidimensionales de variables discretas como ésta se suelen utilizar diagramas de barras tridimensionales. Para su elaboración se levantan barras de altura proporcional a la frecuencia absoluta de los pares que puede tomar la variable tridimensional (X, Y)



La construcción de tablas de doble entrada en el caso de **variables continuas** es similar, con la salvedad de que previamente debemos agrupar los datos de la distribución en intervalos de clase.

Ejercicio 1.-

Los datos obtenidos al estudiar en 25 alumnos las variables X: nota final de matemáticas, e Y: nota final de lengua son los siguientes:

(7'3, 8'2)	(5'1, 4'8)	(3'0, 3'0)	(0'5, 1'6)	(1'0, 1'2)
(9'9, 9'2)	(8'3, 9'8)	(4'0, 5'3)	(2'1, 3'0)	(6'5, 5'0)
(5'4, 3'8)	(5'0, 6'2)	(3'9, 4'8)	(2'1, 2'0)	(7'0, 7'0)
(8'2, 5'4)	(6'9, 4'3)	(3'5, 6'1)	(1'9, 2'2)	(6'7, 7'3)
(9'5, 8'4)	(6'4, 5'8)	(6'1, 7'2)	(5'5, 5'0)	(7'8, 8'7)

Agrupa los datos en 5 intervalos de clase y construye una tabla de doble entrada.

Distribuciones marginales.-

Las **distribuciones marginales** de las variables estadísticas X e Y se obtienen a partir de la tabla de doble entrada considerando una sola variable. Representan las frecuencias de los valores de una variable independientemente de los valores de la otra.

Para la distribución marginal de X tomamos las siguientes columnas de la tabla de doble entrada:

X	x1	x2	x3	…	xk
ni*	n1*	n2*	n3*	…	nk*

Para la distribución marginal de Y tomamos la primera y última filas de la tabla de doble entrada:

Y	y1	y2	y3	…	ym
n*j	n*1	n*2	n*3	…	n*m

Así, las distribuciones marginales del **ejemplo 1** son:

X	**1**	**2**	**3**	**4**
n_{i*}	1	10	36	3

Y	**6**	**7**	**8**	**9**	**10**
n_{*j}	3	16	20	10	1

A partir de las distribuciones marginales podemos calcular las medidas estadísticas, **media** y **desviación típica** de cada una de ellas estudiándolas como una variable unidimensional. En el **ejemplo 1** quedaría:

$$\bar{x} = 2{,}82 \qquad S_x = 0{,}554$$
$$\bar{y} = 7{,}8 \qquad S_y = 0{,}894$$

Ejercicio 2.- Halla las distribuciones marginales de la variable del ejemplo 3, y halla las medias y desviaciones estudiándolas como una variable unidimensional.

Ejercicio 3.- Halla las distribuciones marginales de la variable del ejercicio 1, y halla las medias y desviaciones estudiándolas como una variable unidimensional.

Distribuciones condicionadas.-

La **distribución condicionada** de la variable X fijado un valor de $Y = y_p$ se obtiene de la tabla de doble entrada considerando la primera columna y la correspondiente al valor de $Y = y_p$:

X	x_1	x_2	x_3	…	x_k	
$Y = y_p$	n_{1p}	n_{2p}	n_{3p}	…	n_{kp}	n_{*p}

Análogamente, la **distribución condicionada** de la variable Y fijado un valor de $X = x_s$ se obtiene de la tabla de doble entrada considerando la primera fila y la correspondiente al valor de $X = x_s$:

Y	y_1	y_2	y_3	…	y_m	
$X = x_s$	n_{s1}	n_{s2}	n_{s3}	…	n_{sm}	n_{s*}

Al tratarse la distribución condicionada de una variable unidimensional, tiene sentido calcular las medidas de centralización y dispersión estudiadas en la unidad anterior.

Ejemplo 4: Consideremos la distribución dada por la tabla conjunta:

		X				
		10	15	17	20	
Y	2	3	5	2	4	14
	4	6	10	4	8	28
	7	12	20	8	16	56
		21	35	14	28	98

Las distribuciones condicionadas de X para Y= 2, para Y= 4 y para Y= 7 serían:

X	Frecuencia absoluta condicionada a Y=2	Frecuencia absoluta condicionada a Y=4	Frecuencia absoluta condicionada a Y=7
10	3	6	12
15	5	10	20
17	2	4	8
20	4	8	16
	14	28	56

Las distribuciones condicionadas de Y para X=10, para X=15, para X=17 y para Y= 20 serían:

Y	Frecuencia absoluta condicionada a X=10	Frecuencia absoluta condicionada a X=15	Frecuencia absoluta condicionada a X=17	Frecuencia absoluta condicionada a X=20
2	3	5	2	4
4	6	10	4	8
7	12	20	8	16
	21	35	14	28

Ejercicio 4.- Halla las distribuciones condicionadas de X para Y=1, Y=2, Y=3 y Y=4 de la distribución del ejemplo 3.

Ejercicio 5.- Halla las distribuciones condicionadas de Y para X=1, X=2, X=3, X=4, X=5 y X=6 de la distribución del ejemplo 3.

3.- RELACIÓN ENTRE VARIABLES. DEPENDENCIA FUNCIONAL Y DEPENDENCIA ESTADÍSTICA.

La etapa final de un estudio estadístico es el ***análisis*** de los datos con el fin de extraer conclusiones que puedan ser de interés.

En especial, puede interesarnos estudiar si las dos variables unidimensionales que forman una variable bidimensional presentan algún tipo de relación entre ellas y cuáles son las características de esta relación.

Relación entre variables. En una muestra de familias formadas por padre, madre y dos hijos, hemos estudiado las siguientes variables:

W = ingresos familiares anuales (€)

X = estatura del padre (cm.)

Y = gasto anual en energía eléctrica (€)

Z = consumo anual de energía eléctrica ($kW \cdot h$)

Los valores de Y pueden determinarse exactamente a partir de los valores de Z si conocemos las tarifas de la compañía eléctrica.

Decimos que entre dos variables estadísticas existe **dependencia funcional** si están relacionadas de forma que sea posible determinar con exactitud los valores que toma una de ellas a partir de los que toma la otra.

Consideremos ahora las variables W y Z. Los valores de Z no pueden calcularse exactamente sólo conociendo los de W.

Sin embargo, podemos suponer que consumirán menos energía eléctrica las familias con ingresos más modestos y, por el contrario, que consumirán más las familias con mayores recursos.

Decimos que entre dos variables estadísticas existe **dependencia estadística** o **correlación** cuando los valores que toma una de ellas están relacionadas con los valores que toma la otra, pero no de manera exacta.

Finalmente, parece razonable pensar que no existe ninguna relación entre los valores de W y los de X.

Decimos que dos variables estadísticas son **independientes** si no puede establecerse ninguna relación entre los valores que toma una de ellas y los que toma la otra.

Ejercicio 6.- Determina si entre los siguientes pares de variables existe dependencia funcional o estadística, o bien, si son independientes.

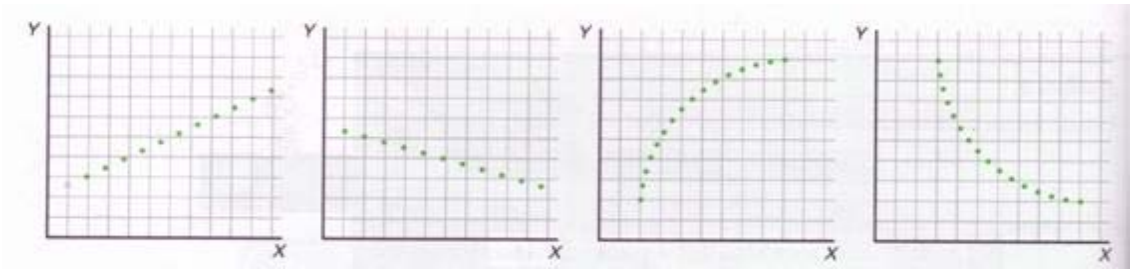
a) Talla de zapatos y estatura. b) Color de cabello y profesión.

c) Radio y longitud de la circunferencia d) Cociente intelectual y peso.

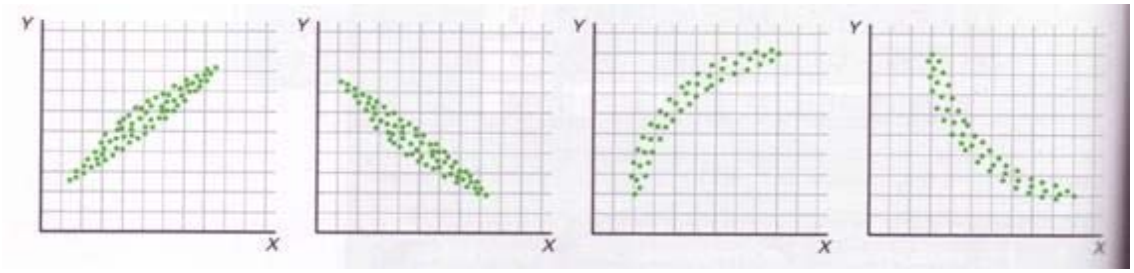
Interpretación gráfica de la relación entre dos variables. Hemos visto que al estudiar la relación entre dos variables pueden darse tres casos: **independencia**, **dependencia funcional** y una situación intermedia a la que llamamos **dependencia estadística** o **correlación.**

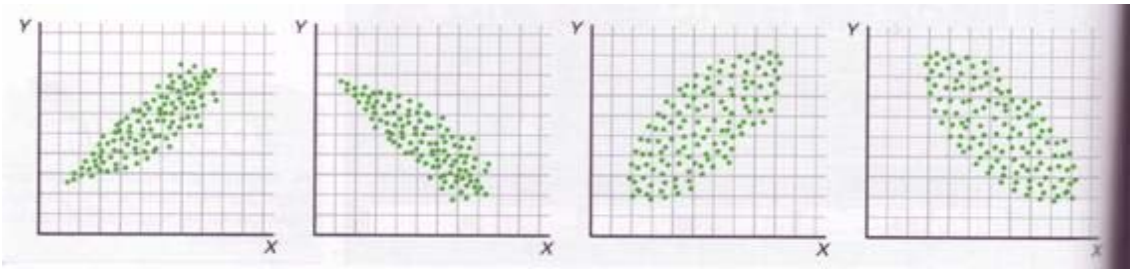
La relación existente entre dos variables queda reflejada en los diagramas de dispersión o nubes de puntos.

- Los puntos de la nube se sitúan sobre una curva cuya expresión matemática podríamos determinar. En este caso corresponde a una dependencia funcional entre las variables X e Y.

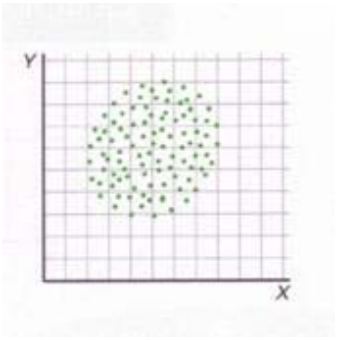


- Los puntos de la nube se agrupan en torno a una posible recta o curva, no muy definida, pero reconocible. Este caso corresponde al de una dependencia estadística o correlación entre las variables X e Y



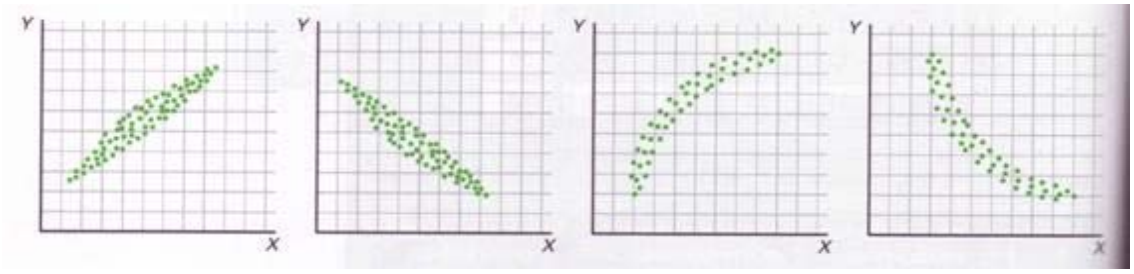


- Los puntos de la curva no se agrupan en tono a ninguna curva, están totalmente en desorden. Este caso corresponde al de interdependencia entre las variables X e Y.

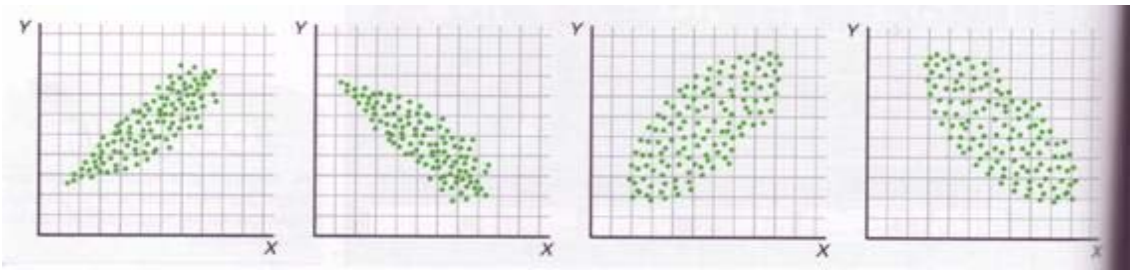


Grado, sentido y tipo de la correlación. Entre los casos extremos de dependencia funcional e independencia hay una amplia gama de situaciones en que se da dependencia estadística o correlación.

- La nube de puntos se ajusta bastante bien a una curva reconocible. En este caso existe **correlación fuerte** entre las variables X e Y.



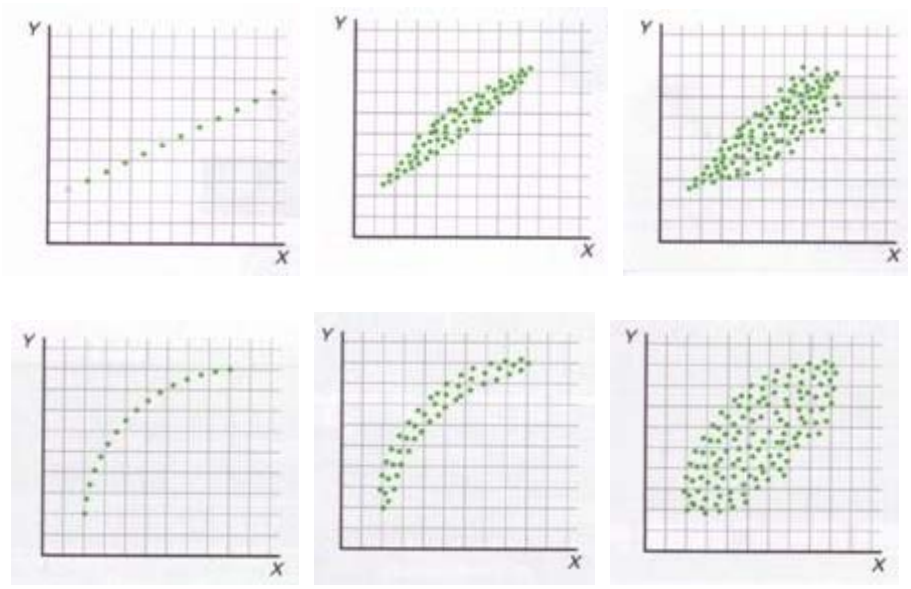
- La nube de puntos se ajusta a una curva, aunque muy ligeramente.



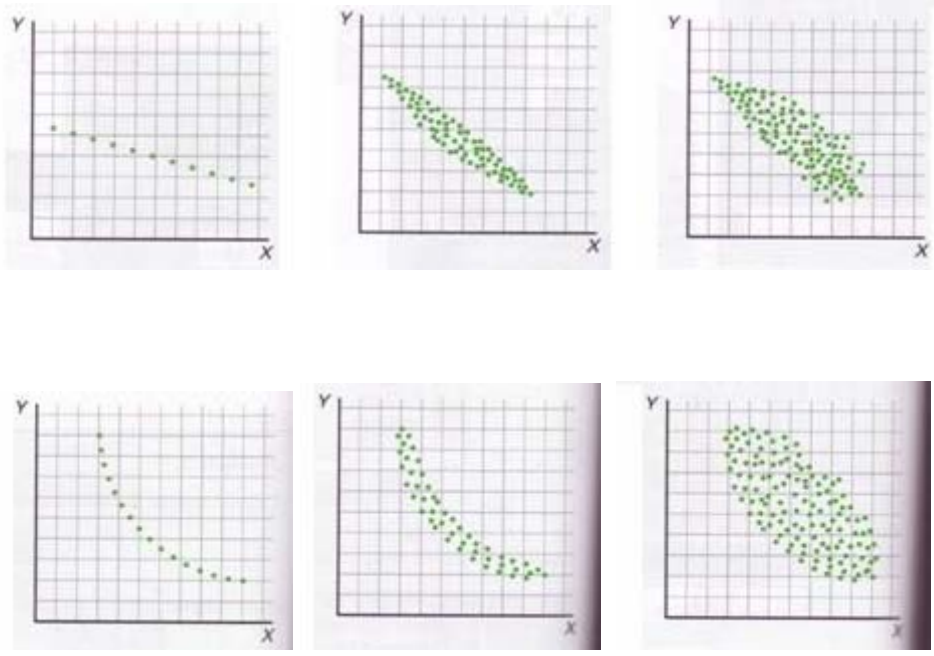
Resumiendo: Decimos que el **grado de la correlación** entre dos variables estadísticas es **fuerte** si la relación entre ambas se acerca a la dependencia funcional, y **débil** si se acerca a la independencia.

En un par de variables estadísticas relacionadas puede ocurrir que, al aumentar los valores de una, aumenten los de la otra o, por el contrario, que disminuyan.

- Al aumentar los valores de X también aumentan los valores de Y. Decimos que existe **correlación positiva** entre las variables e Y.



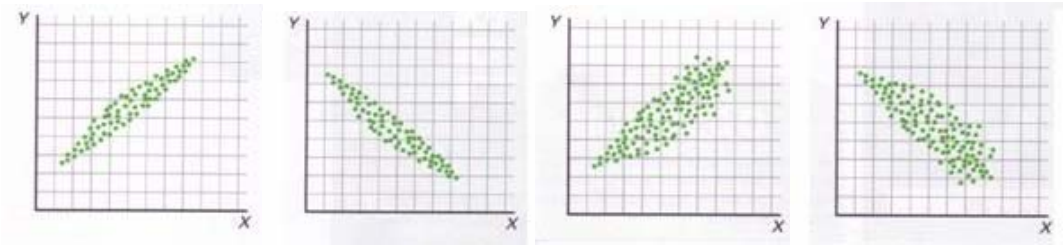
- Al aumentar los valores de X disminuyen los valores de Y. Decimos que existe **correlación negativa** entre las variables X e Y.



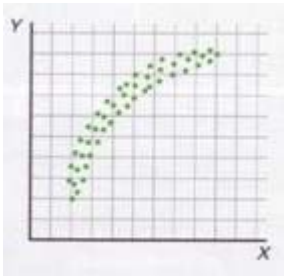
Resumiendo: Entre dos variables estadísticas existe una **correlación de sentido positivo** cuando ambas aumentan conjuntamente, y una **correlación de sentido negativo** cuando una de ellas disminuye al aumentar la otra.

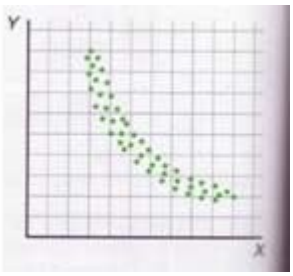
Finalmente, distinguimos entre distribuciones que se ajustan a una recta y distribuciones que se ajustan a otro tipo de curva.

- La nube de puntos se distribuye en torno a una línea recta. Decimos que existe **correlación lineal** entre las variables X e Y.



- La nube de puntos se distribuye en torno a una curva no lineal. Decimos que existe **correlación curvilínea** entre las variables X e Y.





Resumiendo: Cuando los puntos del diagrama de dispersión tienden a agruparse en torno a una línea recta, decimos que existe una **correlación de tipo lineal**. Si los puntos se agrupan en torno a cualquier otro tipo de curva, decimos que existe una **correlación de tipo curvilíneo**.

Ejercicio 7.- Representa la nube de puntos del ejemplo 3 y describe el grado, el sentido y el tipo de correlación que se observa en dicho diagrama de dispersión.

Ejercicio 8.- Dibuja un ejemplo de nube de puntos que corresponda a cada uno de los siguientes casos de correlación:

a) Lineal negativa débil. c) Lineal positiva débil. e) Curvilínea positiva débil.

b) Lineal negativa fuerte. d) Curvilínea negativa débil. f) Curvilínea positiva fuerte.

4.- COVARIANZA. COEFICIENTE DE CORRELACIÓN LINEAL.

Veamos como obtener una medida cuantitativa de la correlación existente entre dos variables estadísticas, para no tener que estar representando continuamente la nube de puntos.

La **covarianza** es una medida que nos permite saber si la relación entre las variables es directa o inversa, y si dicha relación puede ser lineal o no. Se le conoce como varianza conjunta y se calcula mediante la fórmula:

$$S_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \overline{x})\cdot(y_j - \overline{y})\cdot n_{ij}}{n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i \cdot y_j \cdot n_{ij}}{n} - \overline{x}\cdot\overline{y}$$

Si la covarianza es positiva, la relación entre las dos variables es directa, e inversa si la covarianza es negativa.

Vamos a calcular la covarianza del **ejemplo 1:**

$$S_{xy} = \frac{6\cdot4\cdot3 + 7\cdot3\cdot16 + 8\cdot3\cdot20 + 9\cdot2\cdot10 + 10\cdot1\cdot1}{50} - 2'82\cdot7'8 = -0'436$$

Como ya habíamos visto anteriormente, la relación entre las dos variables en estudio es inversa y por eso la covarianza es negativa.

Ejercicio 9.- Halla la covarianza de la distribución del ejemplo 3.

El problema que presenta la covarianza es que depende de las unidades que se emplean para medir las variables. Por ello necesitamos un número que no dependa de las unidades de medida de las variables, es decir, sea adimensional.

Si tenemos en cuenta únicamente el caso de la correlación lineal, se define el **coeficiente de correlación lineal**, que no depende de las unidades de medida y nos da el grado de intensidad de la relación lineal, como:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

El valor de este coeficiente siempre está comprendido entre -1 y 1.

Si el valor de r está próximo a -1 o a 1, entonces la dependencia lineal entre las dos variables es fuerte, siendo ésta directa si r es positivo e inversa si r es negativo.

Si el valor de r está próximo a 0, la dependencia lineal es débil, si es que la hay.

Si el valor es 1 o -1, la dependencia es lineal y todos los puntos de la nube están sobre una recta.

Vamos a ver el grado de intensidad de la relación del **ejemplo 1.**

$$S_{xy} = -0'436 \qquad S_x = 0'554 \qquad S_y = 0'894$$

$$r = \frac{-0'436}{0'554\cdot0'894} = -0'88$$

Se puede considerar que la relación entre las variables es fuerte, ya que el coeficiente de correlación está próximo a -1. Podemos decir que, entre las personas entrevistadas, las que dedican muchas horas a ver la televisión duermen menos tiempo.

Ejercicio 10.- Halla el coeficiente de correlación lineal de la distribución del ejemplo 3.

El coeficiente de correlación con la hoja de cálculo.

Se observaron las edades de cinco niños y sus pesos respectivos, y se consiguieron los resultados siguientes:

Edad (Años)	2	4'5	6	7'2	8
Peso (Kg)	15	19	25	33	34

a) Hallar las medias y desviaciones marginales

b) Calcular el coeficiente de correlación lineal.

Introducimos los datos en las celdas elegidas: en la columna A, la edad, y en la B, el peso.

Edad	Peso		
2	15		
4,5	19		
6	25		
7,2	33		
8	34		
		Media Edad	5,54
		Media Peso	25,2
		Desv. Típica Edad	2,12753378
		Desv. Típica Peso	7,4939976
		Coeficiente de correlación	0,9666494

Las funciones utilizadas en la hoja de cálculo (idénticas en Microsoft Excel y OpenOffice Calc) han sido:

=PROMEDIO(A2:A6)

=PROMEDIO(B2:B6)

=DESVESTP(A2:A6)

=DESVESTP(B2:B6)

=COEF.DE.CORREL(A2:A6;B2:B6)

Ejercicio 11.- Considera la distribución de la siguiente tabla:

X	-2	-2	2	2	2	-2	-1	1	0	1	0	-1
Y	2	0	2	-2	0	-2	1	-1	-2	1	2	-1

a) Dibuja el diagrama de dispersión de la distribución y describe el grado, el sentido y el tipo de la correlación que se observa.

b) Calcula el coeficiente de correlación.

c) Relaciona los resultados obtenidos en los apartados anteriores.

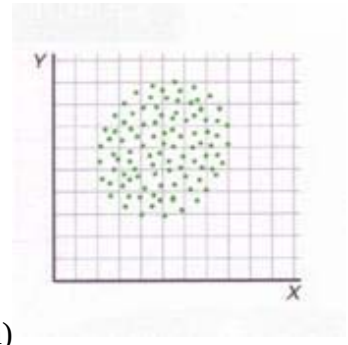
Ejercicio 12.- En la siguiente tabla se expone el crecimiento en centímetros (Y) de una muestra de 10 rosales en un periodo de tiempo, según los gramos de abono aplicados (X):

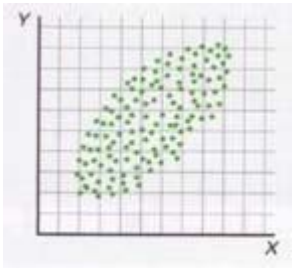
X	1	2	3	4	5	6	7	8	9	10
Y	4	4	6	7	9	6	7	8	9	10

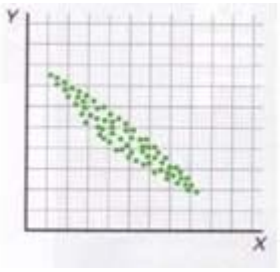
Dibuja la nube de puntos asociada a los datos y averigua, sin hacer cálculos, cuál de los siguientes números corresponde al coeficiente de correlación de la distribución: 0’15, 0’99, 0’03, -0’95.

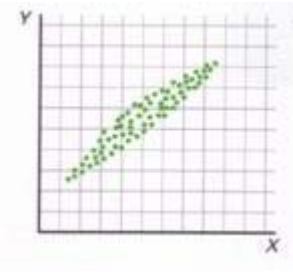
Ejercicio 13.- En cinco estudios estadísticos se han obtenido los siguientes coeficientes de correlación lineal: -0’98, 093, 0’05, 0’71 y -0’62.

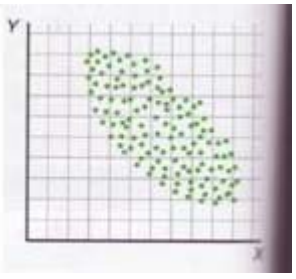
Identifica, justificando la respuesta, la nube de puntos correspondiente a cada una de ellas.

a) 

b) 

c) 

d) 

e) 

Ejercicio 14.- En una competición de patinaje artístico por parejas se otorgan dos notas: una a los ejercicios obligatorios (X) y otra a los ejercicios libres (Y). Las seis parejas que disputan la final han obtenido los siguientes resultados:

X: obligatorios	5	5	6	7	7	7
Y: libres	5	7	7	7	7	8

1. Dibuja el diagrama de dispersión.
2. Calcula el coeficiente de correlación lineal e interpreta el resultado.
3. Coincide los datos del apartado 2) con la información que se puede obtener del gráfico.

Ejercicio 15.- El número de trasplantes de corazón realizados en España entre 1966 y 1992 se recoge en la tabla siguiente, donde X es el año e Y, el número de trasplantes anuales:

X	1986	1987	1988	1989	1990	1991	1992
Y	45	53	73	97	164	232	254

1. Representa el diagrama de dispersión de la distribución y describe el tipo de relación que existe entre las variables X e Y.
2. Calcula el coeficiente de correlación. ¿Es coherente el resultado con el apartado anterior?

Ejercicio 16.- Los siguientes datos se han obtenido después de medir el volumen que ocupa una masa gaseosa, a temperatura fija, y aplicándole 8 valores de presión diferentes:

Presión (P) (atm)	2	5	10	20	25	40	50	100
Volumen (V) (dm^3)	100	40	20	10	8	5	4	2

1. Dibuja la nube de puntos y estudia la relación entre las variables a partir de ella. Si existe correlación, indica su grado, su sentido y su tipo.
2. Calcula el coeficiente de correlación correspondiente a la distribución dada y razona si es contradictorio el resultado obtenido con el apartado anterior.

5.- REGRESIÓN LINEAL. RECTAS DE REGRESIÓN. PREDICCIONES

Uno de los objetivos que se persiguen, al estudiar conjuntamente dos variables X e Y, es encontrar alguna manera de predecir los valores de una conocidos los de la otra.

En este sentido, es lógico pensar que, si hay una curva sobre la que se agrupan aproximadamente los puntos de un diagrama de dispersión, ésta ha ser dar una aproximación de los valores reales.

El análisis que pretende determinar la curva que mejor se aproxima al diagrama de dispersión recibe el nombre de **regresión.**

Nosotros estudiaremos el caso de la regresión lineal, es decir dada una nube de puntos, trataremos de construir una recta que se aproxime de la mejor forma posible a sus puntos, es decir, se trata de sustituir la nube de puntos por una recta de manera que conocido un valor de una variable, podamos estimar el valor de la otra variable.

Es fácil hallar una recta que se ajuste aproximadamente a una distribución. Basta con dibujar la que a simple vista nos parezca más representativa de la nube de puntos. Sin embargo, éste es un método subjetivo.

Para evitar este problema se considera algún criterio que permita determinar objetivamente la recta que se ajusta mejor a la distribución. El más utilizado es el llamado criterio de *los mínimos cuadrados*.

Dadas dos variables X e Y, se define la **recta de regresión** como la recta que hace mínima la suma de los cuadrados de las distancias de los puntos observados a los puntos estimados.

La recta de regresión de Y sobre X (se utiliza para predecir el valor de Y conocido el de X) viene dada por la ecuación:

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

O bien, la recta $y = ax + b$, donde $a = \frac{S_{xy}}{S_x^2}$, $b = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$

La recta de regresión de X sobre Y (se utiliza para predecir el valor de X conocido el de Y) viene dada por la ecuación:

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

O bien, la recta $x = cy + d$, donde $c = \frac{S_{xy}}{S_y^2}$, $d = \bar{x} - \frac{S_{xy}}{S_y^2}\bar{y}$

El cálculo de las rectas de regresión y las predicciones sólo tienen sentido en el caso de que la correlación sea fuerte y se estimen valores cercanos a los datos observados.

Por tanto, cuando el coeficiente de correlación lineal es cercano a 0, no es fiable realizara estimaciones.

El punto de corte de las dos rectas de regresión es $(\bar{x}, \bar{y})$, salvo en el caso de que r valga -1 o 1, en el que ambas rectas coinciden.

Las rectas de regresión del **ejemplo 1** son:

$\bar{x} = 2'82$ $\bar{y} = 7'8$ $S_{xy} = -0'436$ $S_x^2 = 0'3069$ $S_y^2 = 0'799$

De Y sobre X $y - 7'8 = -1'417(x - 2'82)$

De X sobre Y $x - 2'82 = -0'545(y - 7'8)$

O bien, $a = -1'417, b = 11'79, c = -0'545, d = 7'071$

Ejercicio 17.- a) Halla las rectas de regresión de la distribución del ejemplo 3.
b) Representa en un mismo gráfico la nube de puntos y las rectas de regresión.

Ejercicio 18.- a) Halla las rectas de regresión de la distribución del ejercicio 16.
b) Representa en un mismo gráfico la nube de puntos y las rectas de regresión.

La recta de regresión con la hoja de cálculo.-

Se observaron las edades de cinco niños y sus pesos respectivos, y se consiguieron los resultados siguientes:

Edad (Años)	2	4'5	6	7'2	8
Peso (Kg)	15	19	25	33	34

Hallar la recta de regresión de Y sobre X

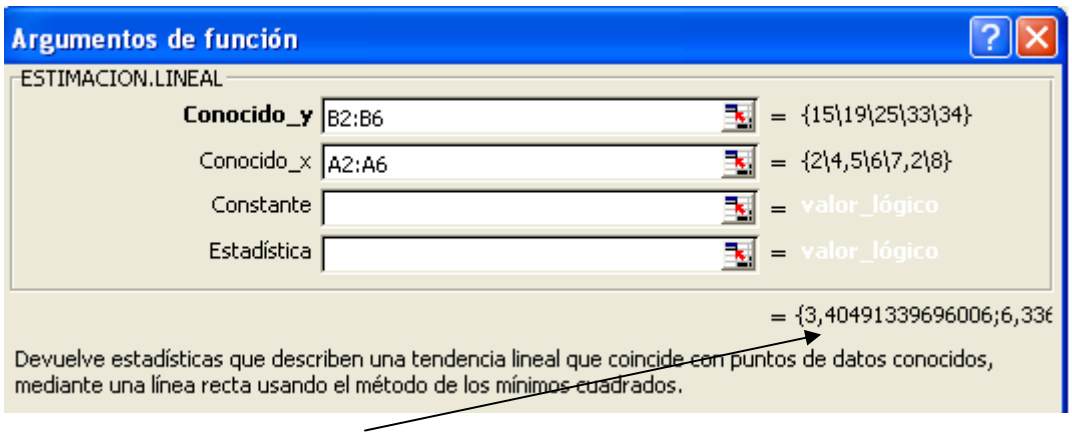
Introducimos los datos en las celdas elegidas: en la columna A, la edad, y en la B, el peso.

Edad	Peso		
2	15		
4,5	19		
6	25		
7,2	33		
8	34		
		Media Edad	5,54
		Media Peso	25,2
		Desv. Típica Edad	2,12753378
		Desv. Típica Peso	7,4939976
		Coeficiente de correlación	0,9666494
		Pendiente recta regresión	3,4049134

La función utilizada (tanto en Excel como en Open Calc) en la celda D14 ha sido: =PENDIENTE(B2:B6;A2:A6). Por tanto la recta de regresión es:

$$y - 25'2 = 3'4049\ (x - 5'54).$$

También podemos calcular la recta de regresión seleccionando la función ESTIMACIÓN.LINEAL de las funciones estadísticas.

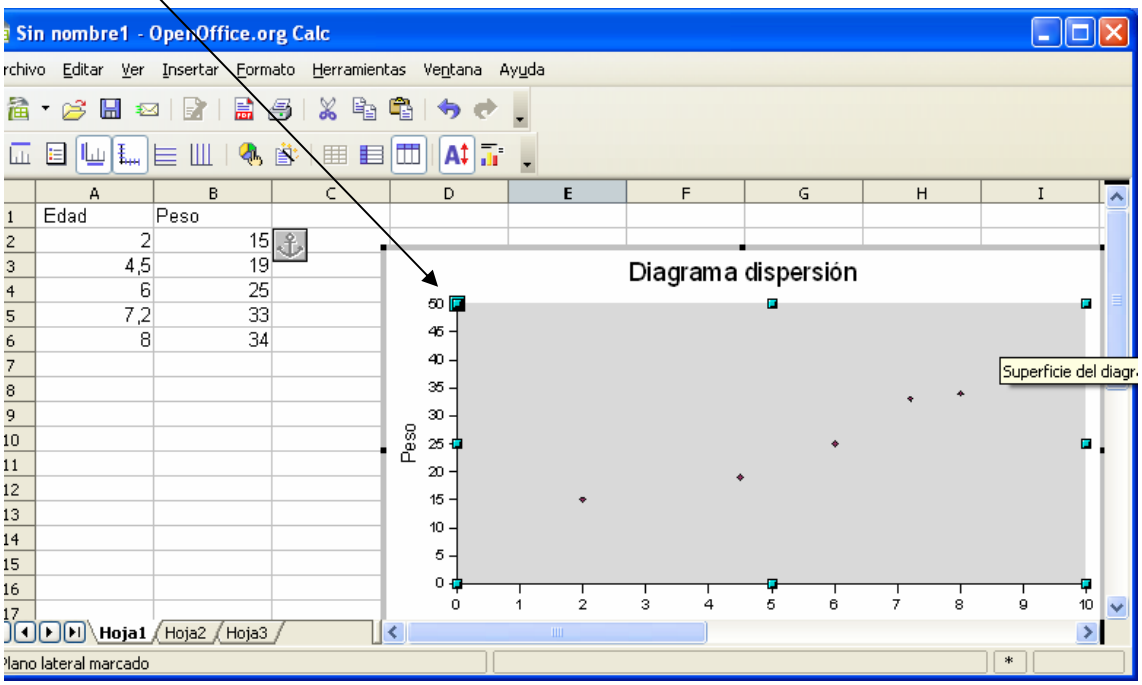


En la pantalla aparecen los valores *a* y *b*, siendo $y = ax+b$. Por tanto la recta de regresión es: y = 3'4049x + 6'336.

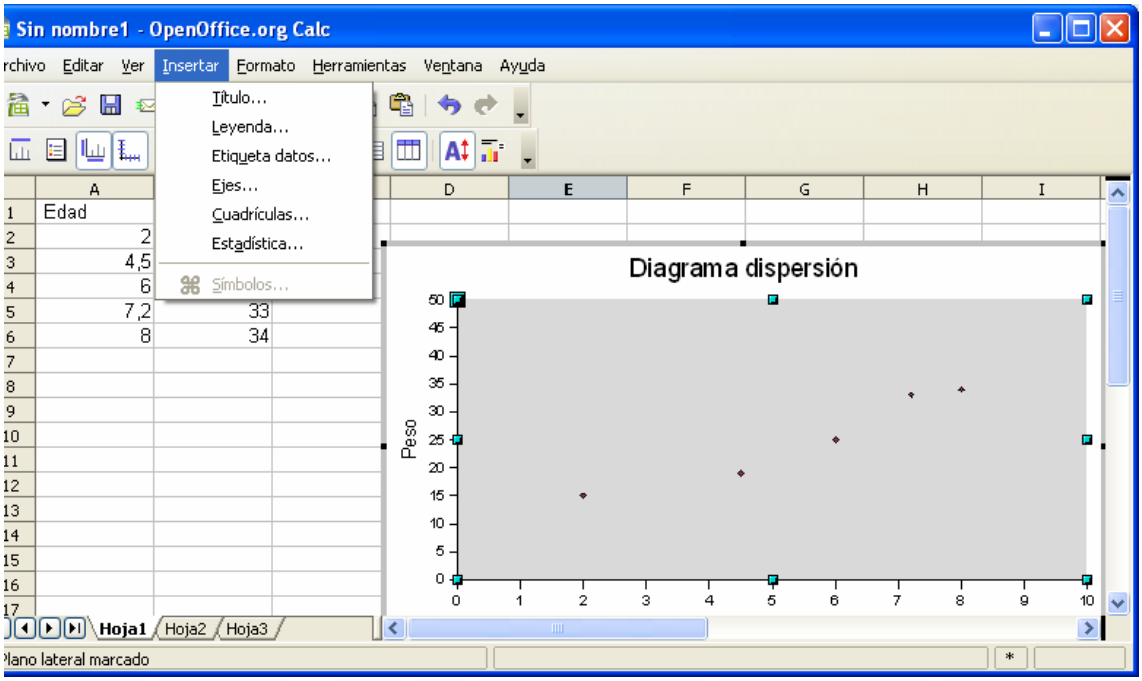
La nube de puntos y la recta de regresión en un mismo gráfico con la hoja de cálculo.-

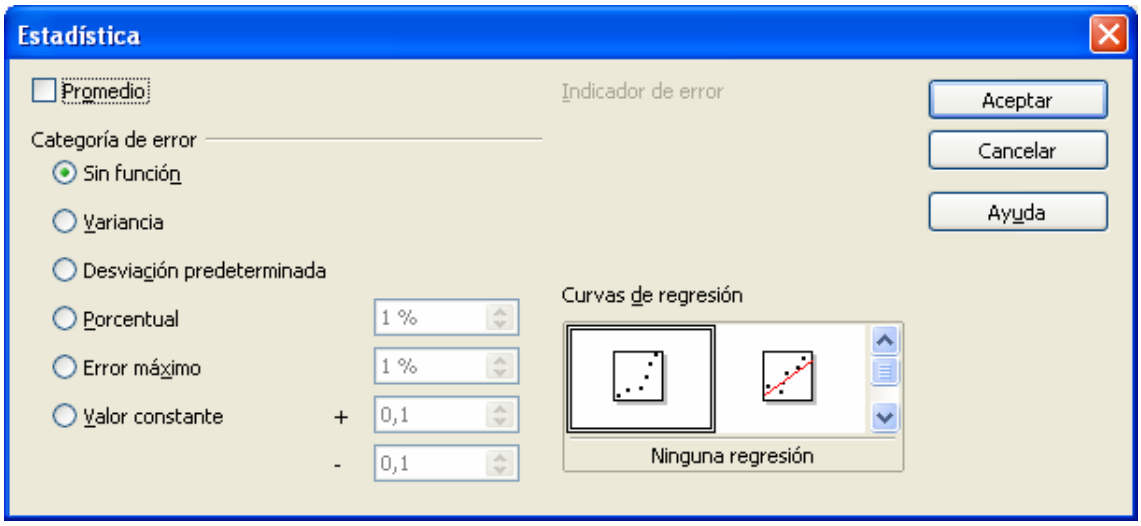
En OpenOffice Calc:

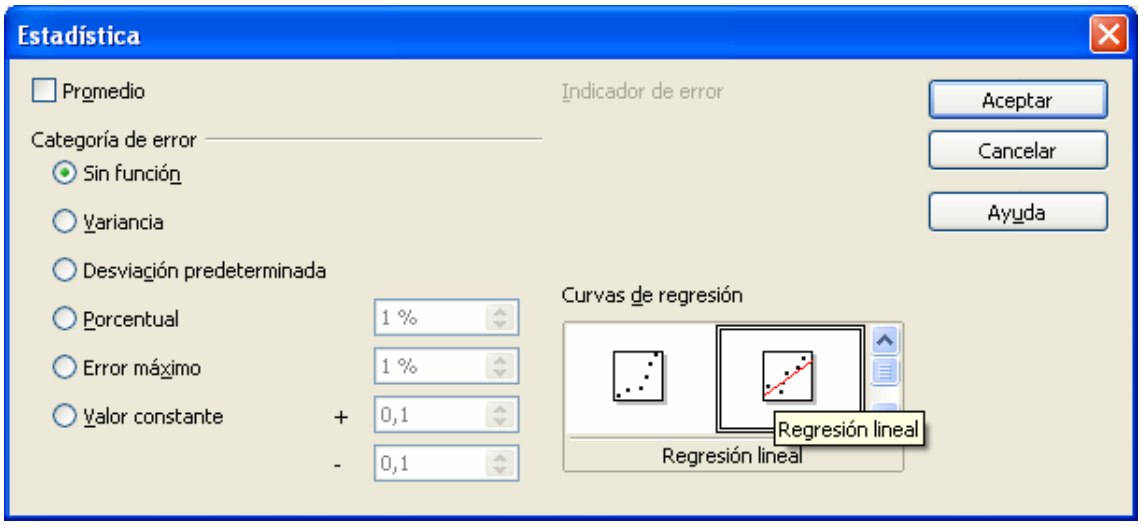
Se señala el gráfico como se indica en el siguiente gráfico:

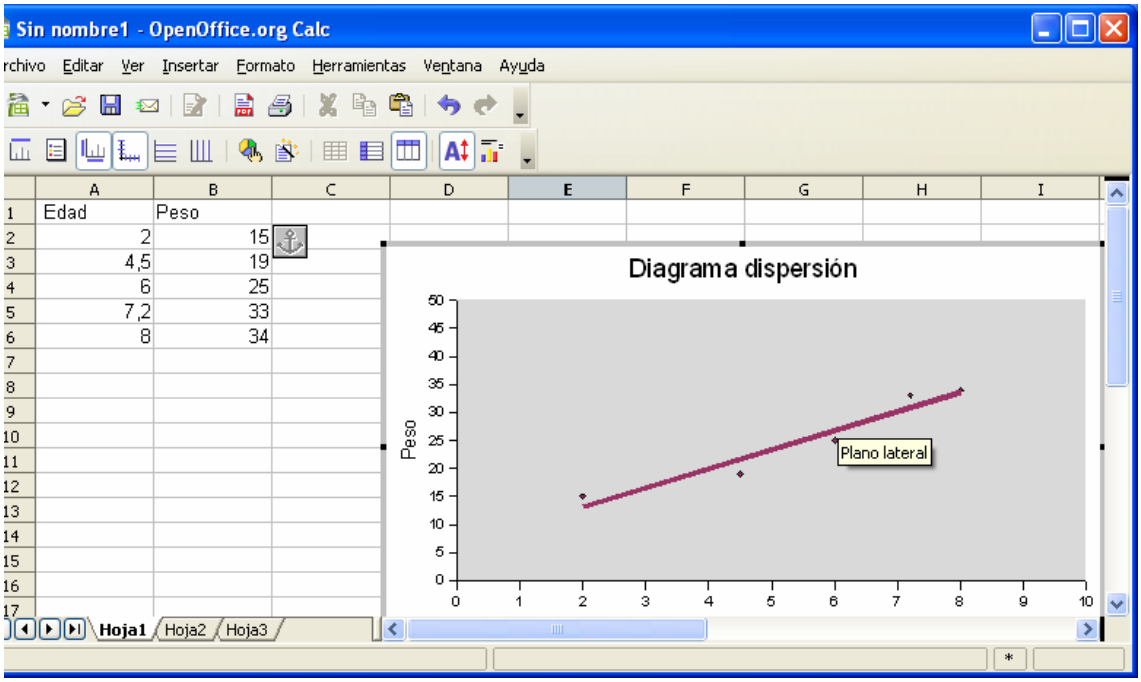


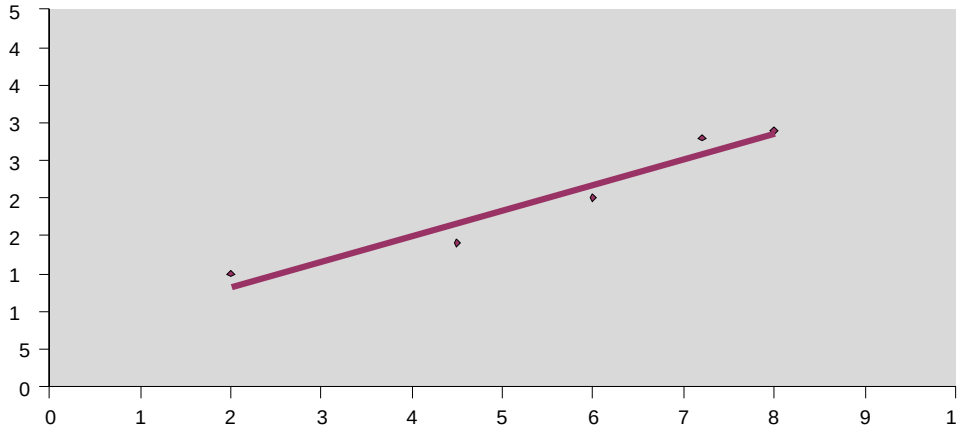
Se siguen las siguientes pantallas:




Estadística
Promedio
Indicador de error
Aceptar
Categoría de error
Cancelar
Sin función
Ayuda
Varianza
Desviación predeterminada
Curvas de regresión
Porcentual
1 %
Error máximo
1 %
Valor constante
+
0,1
-
0,1
Ninguna regresión

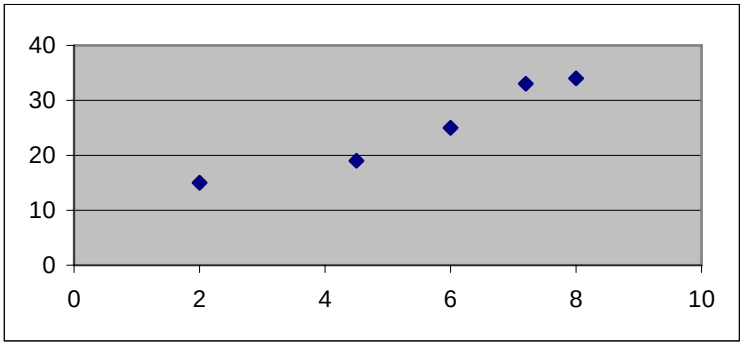

Estadística
Promedio
Indicador de error
Aceptar
Categoría de error
Cancelar
Sin función
Ayuda
Varianza
Desviación predeterminada
Curvas de regresión
Porcentual
1 %
Error máximo
1 %
Valor constante
+
0,1
-
0,1
Regresión lineal
Regresión lineal


Sin nombre1 - OpenOffice.org Calc
Archivo Editar Ver Insertar Formato Herramientas Ventana Ayuda
Edad
Peso
2
15
4,5
19
6
25
7,2
33
8
34
Diagrama dispersión
Peso
Plano lateral
Hoja1
Hoja2
Hoja3

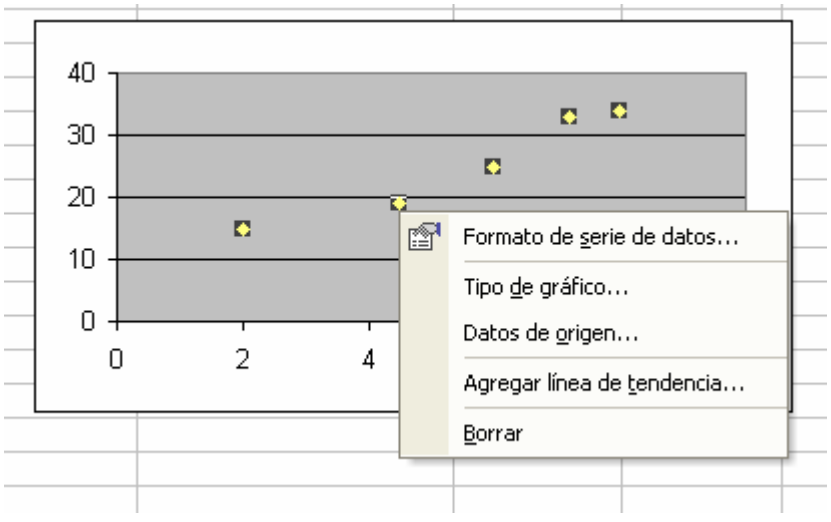


El gráfico confirma lo dicho en los apartados anteriores.

En Excel, seleccionamos el gráfico **XY (Dispersión)** y lo dibujamos:

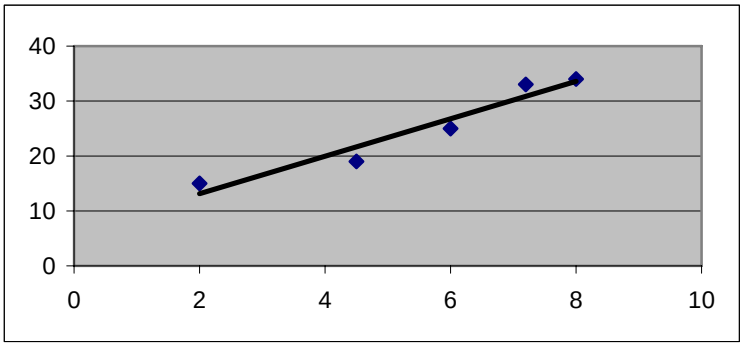


A continuación se elige un punto cualquiera con el botón izquierdo y hacemos clic con el derecho:



y seleccionamos “Agregar línea de tendencia”:

Por último, eligiendo “Lineal”, obtenemos:



Ejercicio 19.- Realiza con la hoja de cálculo el ejercicio 17 (sólo la recta de regresión de Y sobre X).

Ejercicio 20.- Realiza con la hoja de cálculo el ejercicio 18 (sólo la recta de regresión de Y sobre X).

Estimaciones o predicciones. A la hora de realizar estimaciones de los valores de una variable, tenemos que comprobar:

- Si a la vista del diagrama de dispersión podemos inducir una posible dependencia lineal.
- El coeficiente de correlación esté próximo a 1 o -1.
- El sentido común nos hace pensar que hay alguna relación entre las variables.

En el **ejemplo 1,** el número de horas que ve la televisión una persona que duerme siete horas y cuarto, sustituimos en la recta de regresión de X sobre Y la variable *y* por 7’25 y nos queda:

$$x - 2'82 = -0'545(7'25 - 7'8) = 2'82 + 0'29975 = 3'11975 \approx 3 \text{ horas}$$

Valoración de las predicciones. La recta de regresión nos permite predecir valores de una variable a partir de los de otra. No obstante, hay que tener siempre presente que existen las siguientes limitaciones:

- Las predicciones realizadas a partir de una recta de regresión no son fiables si entre *X* e *Y* no hay un alto grado de correlación lineal. Volvemos a decir que *r* debe estar, en valor absoluto, cercano a 1.
- La fiabilidad de una recta de regresión es mayor cuanto mayor sea el número de datos considerados para calcularla.
- Las predicciones obtenidas para valores próximos al punto medio de la distribución son más fiables que las obtenidas para valores muy alejados.

Ejemplo: A partir de la distribución del **ejercicio 14** (competición de patinaje artístico), calcula la recta de regresión de Y sobre X, y de la de X sobre Y, y responde a las siguientes cuestiones:

a) Una pareja de patinadores ha obtenido un 8 en los ejercicios obligatorios. ¿Qué nota cabe esperar que haya obtenido en los ejercicios libres?
b) Otra pareja ha obtenido un 9 en los ejercicios libres. ¿Qué nota cabe esperar que haya obtenido en los ejercicios obligatorios?
c) Valora las predicciones efectuadas.

La recta de regresión de Y sobre X es $y = 0'659x + 2'768$

La recta de regresión de X sobre Y es $x = 0'652y + 1'713$

Por tanto:

a) Sustituyendo $x = 8$ en la recta de regresión de Y sobre X, obtenemos $y = 8$

b) Sustituyendo $y = 9$ en la recta de regresión de X sobre Y, obtenemos $x = 8$

c) Las predicciones se han obtenido para valores no muy alejados del punto medio de la distribución. Además el coeficiente de correlación era $r = 0'655$, que indica una correlación relativamente fuerte. Sin embargo, el número de puntos utilizados para hallar la recta de regresión es pequeño.

Así pues, esto nos hace pensar que son predicciones fiables, aunque con reservas.

Estimaciones o predicciones con la hoja de cálculo.-

Se observaron las edades de cinco niños y sus pesos respectivos, y se consiguieron los resultados siguientes:

Edad (Años)	2	4'5	6	7'2	8
Peso (Kg)	15	19	25	33	34

¿Qué peso se espera que tenga un niño de 5 años?

Edad	Peso		
2	15		
4,5	19		
6	25		
7,2	33		
8	34		
		Peso esperado para una edad de cinco años	23,36134677

La función utilizada (tanto en Excel como en Open Calc) en la celda D8 ha sido:
=TENDENCIA(B1:B5, A1:A5 ,5)

Resolución completa de un problema.-

Ejemplo: En un hospital se está experimentando un medicamento que regula la temperatura corporal. Para ello, se administran diferentes dosis del producto a 10 pacientes con fiebre alta, y se observa cuánto tiempo tarda en normalizarse completamente su temperatura. Se obtienen los siguientes resultados:

Dosis (mg)	2	4	6	8	10	12	14	16	18	20
Tiempo (min)	136	126	115	98	75	60	55	42	38	31

¿Cuánto tiempo cabe esperar que tarde en normalizarse la temperatura de un paciente al que se le han administrado 11’5 mg del medicamento? ¿Y si toma una dosis de 25 mg?

Fases:

Comprensión del enunciado.

A partir de una distribución de datos hemos de realizar una predicción del efecto que tendrá sobre la temperatura corporal la prescripción de una determinada dosis del medicamento que se está experimentando.

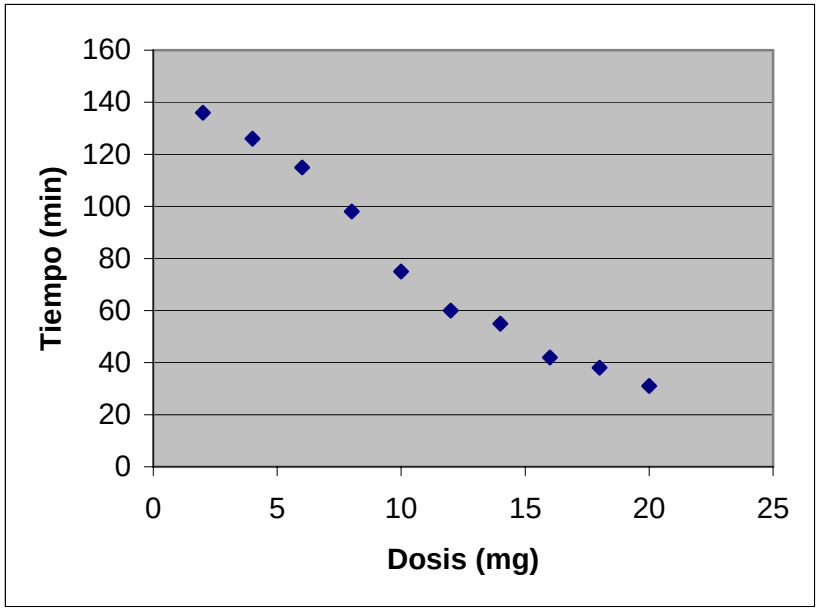
Los datos corresponden a dos variables: la dosis de medicamento (X) y el tiempo que tarda en normalizarse la temperatura (Y).

Planificación

- Representaremos un diagrama de dispersión y observaremos la relación que existe entre las variables y, en caso de dependencia estadística, el grado, el sentido y el tipo de correlación.
- Calcularemos el coeficiente de correlación lineal y valoraremos un posible ajuste mediante una recta de regresión.
- Hallaremos la recta de regresión de Y sobre X.
- Realizaremos una predicción del tiempo que tardará en normalizarse la temperatura para las dosis del medicamento del enunciado.

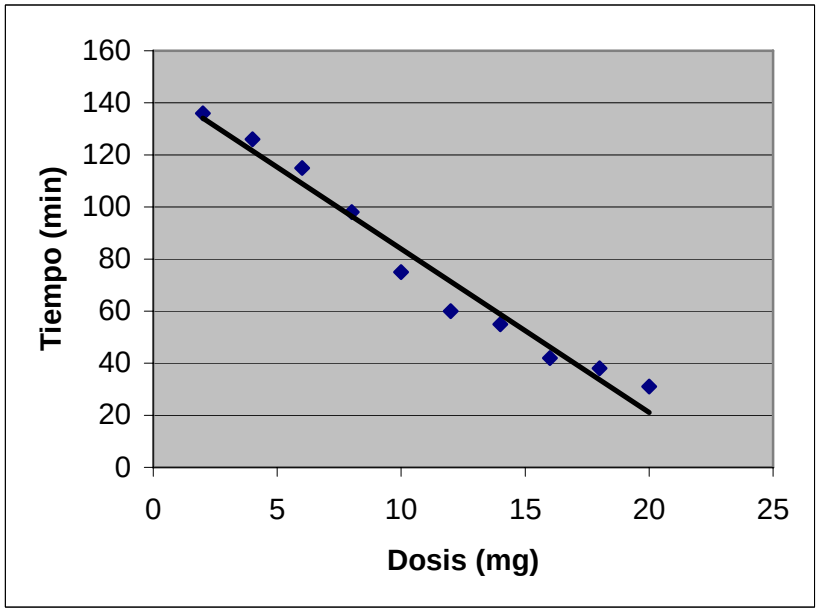
Ejecución

Con una aplicación informática adecuada (en este caso Excel) hacemos los cuatro puntos anteriores. Empezamos con el diagrama de dispersión.



A simple vista, podemos concluir que existe una fuerte correlación lineal negativa.

Incluimos a continuación la recta de tendencia para confirmar nuestras previsiones.

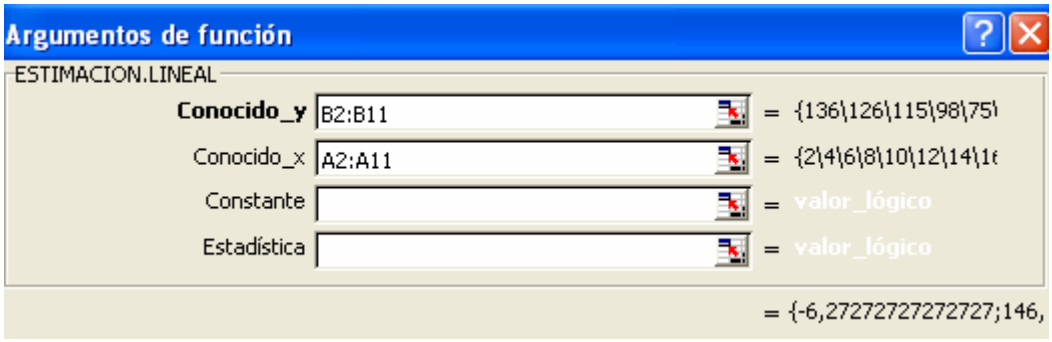


Calculamos el coeficiente de correlación.

X (Dosis)	Y (Tiempo)	
2	136	
4	126	
6	115	
8	98	
10	75	
12	60	
14	55	
16	42	
18	38	
20	31	
Coeficiente de correlación		-0,98428777

Comprobamos que existe una correlación lineal negativa muy acusada, puesto que r está cercano a -1.

Determinamos la recta de regresión de Y sobre X.



La recta buscada es: $y = -6'272x + 146'6$

Estimamos los tiempos correspondientes a las dosis de 11’5 mg y 25 mg.

X (Dosis)	Y (Tiempo)			
2	136			
4	126			
6	115			
8	98			
10	75			
12	60			
14	55			
16	42			
18	38			
20	31			
Tiempo estimado para una dosis de 11'5 gramos				74,4636364
Tiempo estimado para una dosis de 25 gramos				-10,2181818

O sea, $\hat{y}_{11'5mg} = 74'47 \text{ min}$ $\hat{y}_{25mg} = -10'2 \text{ min}$

Respuesta.

Si administramos a un enfermo 11’5 mg del medicamento, cabe esperar que su temperatura se normalice al cabo de 74’47 minutos. En cambio, si le administramos 25 mg, su temperatura se normalizará al cabo de -10’2 minutos… ¡antes de que se lo tome!

A pesar de que el grado de correlación es alto, la segunda predicción no es fiable porque está muy alejada del punto medio de la distribución ($\bar{x} = 11$)

Ejercicio 21.- Un centro comercial sabe los clientes que lo pueden visitar en función de la distancia, en kilómetros, a la que se sitúe de un núcleo de población, según los datos de la siguiente tabla:

Nº de clientes (en cientos)	8	7	6	4	2	1
Distancia (en kilómetros)	15	19	21	23	34	40

a) Si el centro comercial se sitúa a 2 km, ¿cuántos clientes puede esperar?
b) Si desea recibir a 500 clientes, ¿a qué distancia del núcleo de población debe situarse?
c) ¿Consideras fiables los resultados obtenidos?

Ejercicio 22.- La producción de largometrajes en España durante el periodo 1950 – 1960 se indica en la tabla siguiente, donde X representa el año e Y, el número de largometrajes producidos.

X	1950	1952	1954	1956	1958	1960
Y	49	41	69	75	75	73

a) ¿Qué producción de largometrajes cabe esperar que hubiese en 1957? ¿Y en 1966?
b) Sabiendo que en el año 1957 se produjeron realmente 72 largometrajes y que en 1966 fueron 160, interpreta los resultados del apartado anterior.

Ejercicio 23.- La siguiente tabla recoge las calificaciones de 40 alumnos en Matemáticas y en Física:

Matemáticas X	3	4	5	6	6	7	7	8	10
Física Y	2	5	5	6	7	6	7	9	10
Nº de alumnos	4	6	12	4	5	4	2	1	2

a) Escribe las distribuciones marginales de X e Y. Halla las respectivas medias y desviaciones típicas.
b) Dibuja el diagrama de dispersión.
c) Calcula el coeficiente de correlación lineal.
d) ¿Qué nota de Matemáticas se puede esperar que saque un alumno que ha obtenido 7’5 en Física? ¿Es fiable el resultado?

Ejercicio 24.- La talla media de una muestra de padres es de 1’68 m, con una desviación típica de 5 cm, y la talla media de una muestra de sus hijos es de 1’70 m, con una desviación típica de 7’5 cm. El coeficiente de correlación entre las tallas de hijos y padres es de 0’7. Estimar la talla de dos hijos si la talla de sus padres fuera de 1’80 y 1’60 m, respectivamente.

Ejercicio 25.- Para gestionar mejor su almacén, una empresa decide estimar sus necesidades de materias primas mediante el volumen de negocio y el total de salarios. Durante los seis últimos meses han obtenido los datos siguientes:

Mes	***1***	***2***	***3***	***4***	***5***	***6***
X	0'9	1'2	0'6	0'5	1'4	1
Y	220	240	200	200	246	210
Z	23	22	19	20	21	22

donde X representa, en toneladas, la cantidad de materias primas; Y representa el volumen de negocios, en miles de euros; y Z, el total de salarios, en miles de euros.

a) Calcula los coeficientes de correlación lineal entre X e Y y entre X y Z.

b) Utiliza la recta de regresión de Y sobre X para estimar las necesidades de materias primas cuando el volumen de negocios sea de 234000 euros.

c) ¿Crees que es fiable el resultado obtenido en el apartado anterior?

d) ¿Existe relación entre la cantidad de materias primas y el total del salario?

Ejercicio 26.- Una compañía discográfica ha recopilado la siguiente información sobre el número de conciertos dados por 15 grupos musicales y las ventas de discos de estos grupos (expresados en miles), obteniendo los datos siguientes:

		Conciertos		
		(0,30]	(30,40]	(40,80]
Discos	(1,5]	3	0	0
	(5,10]	1	4	1
	(10,20]	0	1	5

a) Completa la tabla con las marcas de clase.

b) Representa los datos en un diagrama de dispersión. Analiza la relación entre las variables.

c) Halla el coeficiente de correlación y contrasta su valor con el análisis del apartado anterior.

d) Obtén la recta de regresión que explica la dependencia anterior.

e) Si un grupo musical ha vendido 18.000 discos, ¿qué número de conciertos es previsible que dé?

Ejercicio 27.- Una planta envasadora de frutos secos necesita adquirir una máquina empaquetadora de bolsas de 50 gr, lo más precisa posible, para lo cual efectúa una prueba de diez pesadas con dos aparatos X e Y con los siguientes resultados:

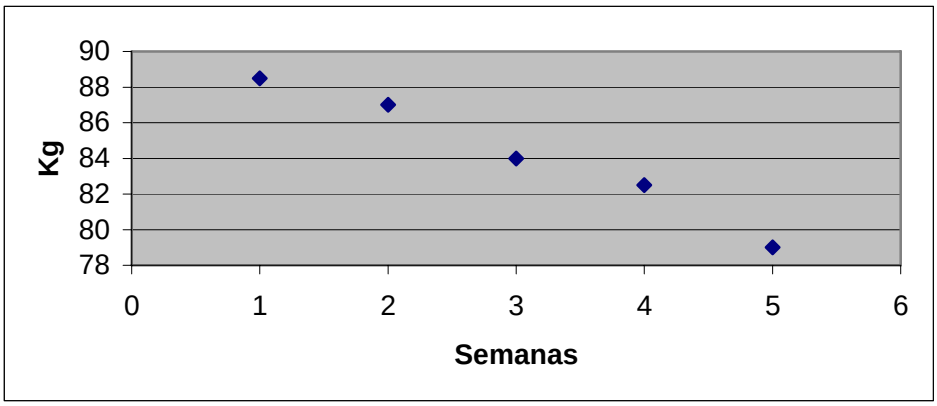
X	52	54	53	47	48	49	46	48	51	52
Y	51	54	51	46	49	49	48	49	51	52

a) Calcular la media y la desviación típica de cada una de las distribuciones marginales X e Y. ¿Qué maquina debe elegir y por qué?

b) Calcular el coeficiente de correlación entre X e Y

c) ¿Qué pesada se espera obtener de la máquina Y en una prueba si se sabe que X ha dado 54 gr?

Ejercicio 28.- Una persona se somete a una dieta de adelgazamiento durante cinco semanas con los resultados que aparecen en la gráfica:



a) ¿Dirías que existe una fuerte correlación lineal entre el peso y la duración de la dieta?
b) A la vista de la gráfica, ¿cuál te parece que es el coeficiente de correlación aproximado? ¿Por qué?
c) Calcula, a partir de la recta de regresión correspondiente, el peso de una persona si continuase con la dieta dos semanas más.

Ejercicio 29.- En una población de 60 individuos se han observado dos variables estadísticas X e Y, obteniéndose los datos que aparecen en la tabla:

		X				
		1	**2**	**3**	**4**	**5**
Y	**3**	1	2	---	---	---
	4	---	4	6	---	---
	5	---	---	10	12	---
	6	---	---	---	15	5
	7	---	---	---	1	4

a) Determina los parámetros de la distribución marginal de X.

b) Representa la nube de puntos de la distribución bidimensional.

c) Explica cuál de los siguientes valores: -0'9; 0'2; 0'85 puede ser el del coeficiente de correlación lineal de las variables X e Y.

Ejercicio 30.- Estudiando dos características, X=”número de miembros de una familia” e Y=”número de habitaciones de la vivienda”, se obtienen los siguientes valores de una variable bidimensional:

(5,4) (3,2) (1,3) (7,5) (6,4) (5,3) (3,2) (6,4) (7,5) (3,1)

a) Hacer las tablas de frecuencia simple y de doble entrada.

b) Representa el diagrama de dispersión y la recta de regresión de Y sobre X. ¿Cómo es la correlación?