COEFICIENTE DE CORRELACIÓN DE KARL PEARSON

Autor: Mario Suárez
mgsmariosuarez@gmail.com

Llamando también coeficiente de correlación producto-momento.

a) Para datos no agrupados se calcula aplicando la siguiente ecuación:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

r = Coeficiente producto-momento de correlación lineal

$x = X - \bar{X}\,; y = Y - \bar{Y}$

Ejemplo ilustrativo:
Con los datos sobre las temperaturas en dos días diferentes en una ciudad, determinar el tipo de correlación que existe entre ellas mediante el coeficiente de PEARSON.

X	18	17	15	16	14	12	9	15	16	14	16	18	ΣX =180
Y	13	15	14	13	9	10	8	13	12	13	10	8	ΣY= 138

Solución:

Se calcula la media aritmética

$\bar{x} = \frac{\sum x_i}{n}$

Para X:

$\bar{X}_X = \frac{180}{12} = 15$

Para Y:

$\bar{Y}_Y = \frac{138}{12} = 11{,}5$

Se llena la siguiente tabla:

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
18	13	3	1,5	9	4,5	2,25
17	15	2	3,5	4	7	12,25
15	14	0	2,5	0	0	6,25
16	13	1	1,5	1	1,5	2,25
14	9	-1	-2,5	1	2,5	6,25
12	10	-3	-1,5	9	4,5	2,25
9	8	-6	-3,5	36	21	12,25
15	13	0	1,5	0	0	2,25
16	12	1	0,5	1	0,5	0,25
14	13	-1	1,5	1	-1,5	2,25
16	10	1	-1,5	1	-1,5	2,25
18	8	3	-3,5	9	-10,5	12,25
180	**138**			**72**	**28**	**63**

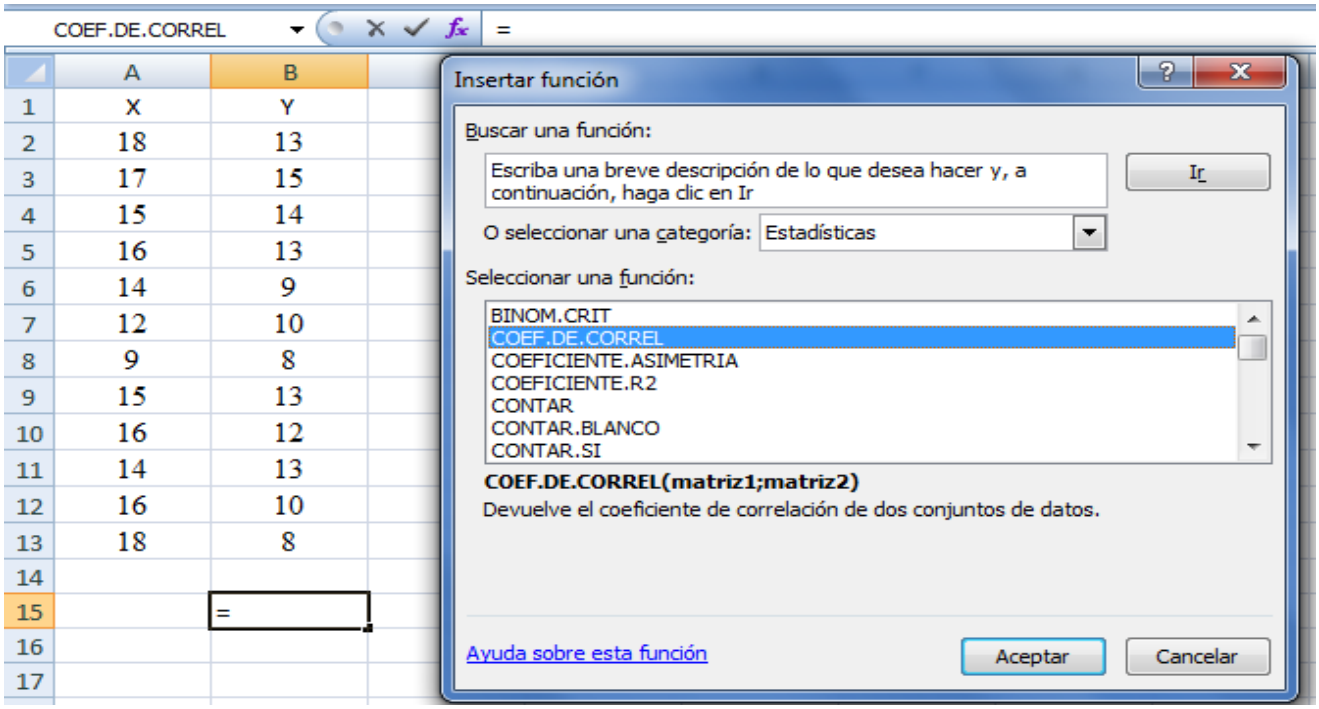
Se aplica la fórmula:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{28}{\sqrt{(72)(63)}} = 0{,}416$$

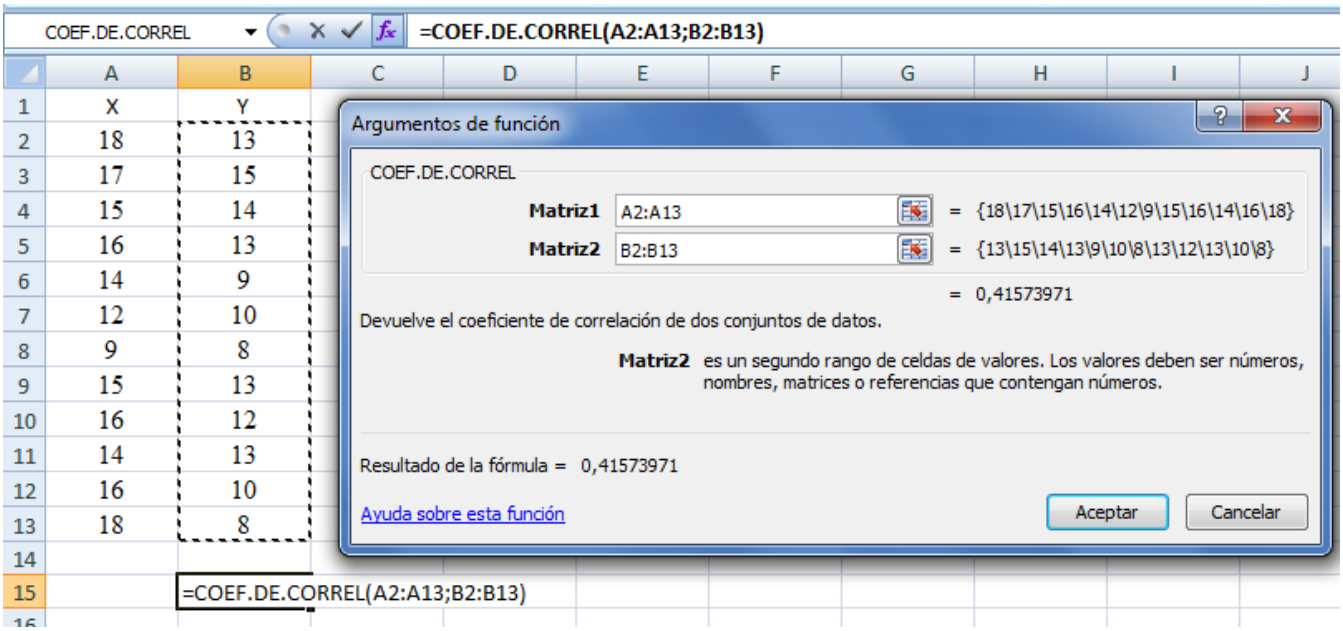
Existe una correlación moderada

En Excel se calcula de la siguiente manera:

a) Se inserta la función COEF.DE.CORREL y pulsar en Aceptar.



b) En el cuadro de argumentos de la función, en el recuadro de la Matriz 1 seleccionar las celdas de X, y en el recuadro de la Matriz 2 seleccionar las celdas de Y.

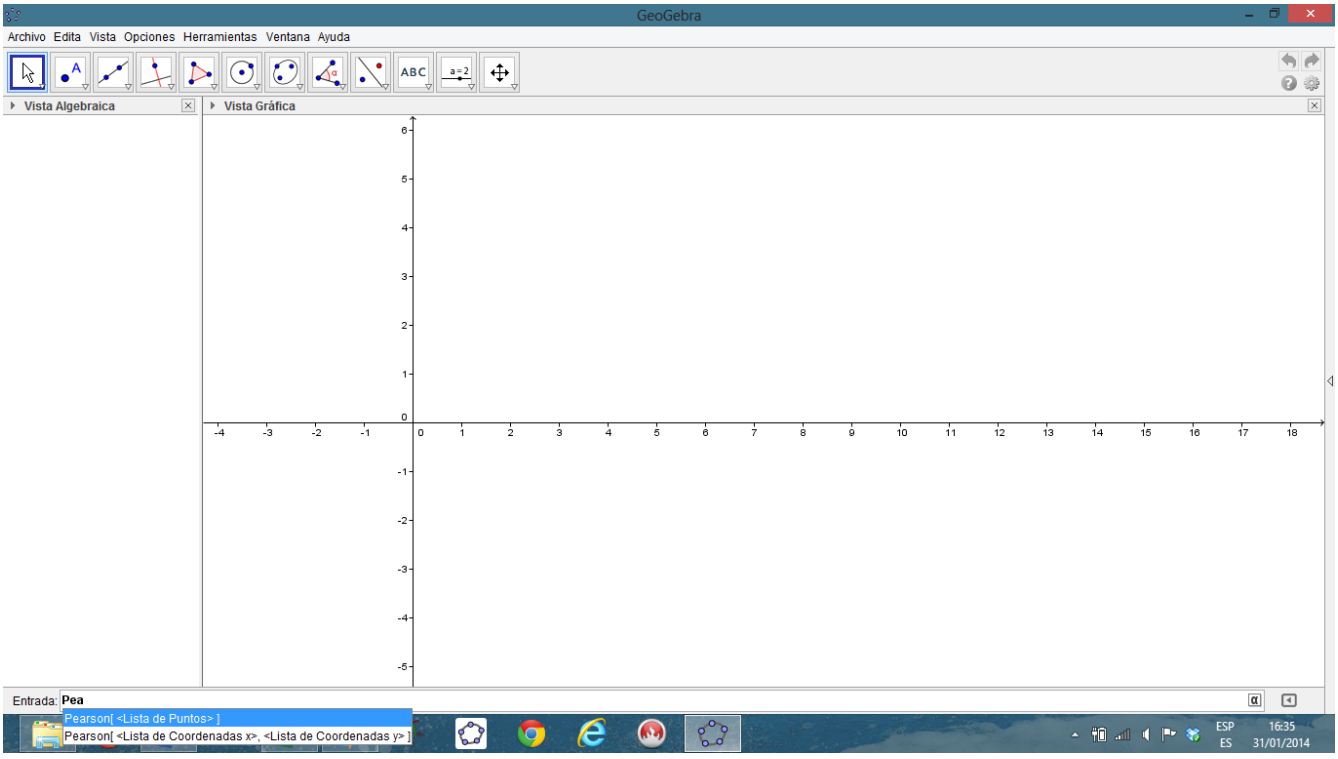


c) Pulsar en Aceptar.

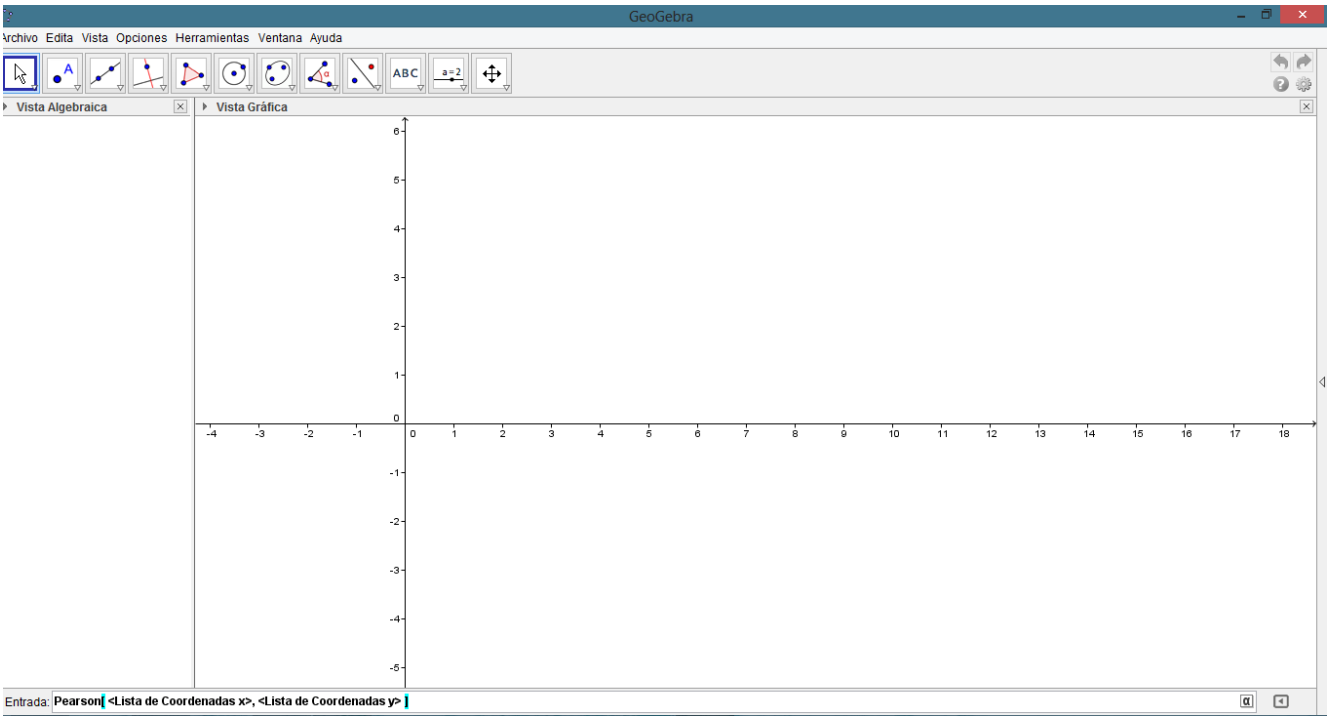
	A	B	C	D	E
1	X	Y			
2	18	13			
3	17	15			
4	15	14			
5	16	13			
6	14	9			
7	12	10			
8	9	8			
9	15	13			
10	16	12			
11	14	13			
12	16	10			
13	18	8			
14					
15		0,41573971	=COEF.DE.CORREL(A2:A13;B2:B13)		

En GeoGebra se calcula de la siguiente manera:

a) Escribir en Entrada Pearson.

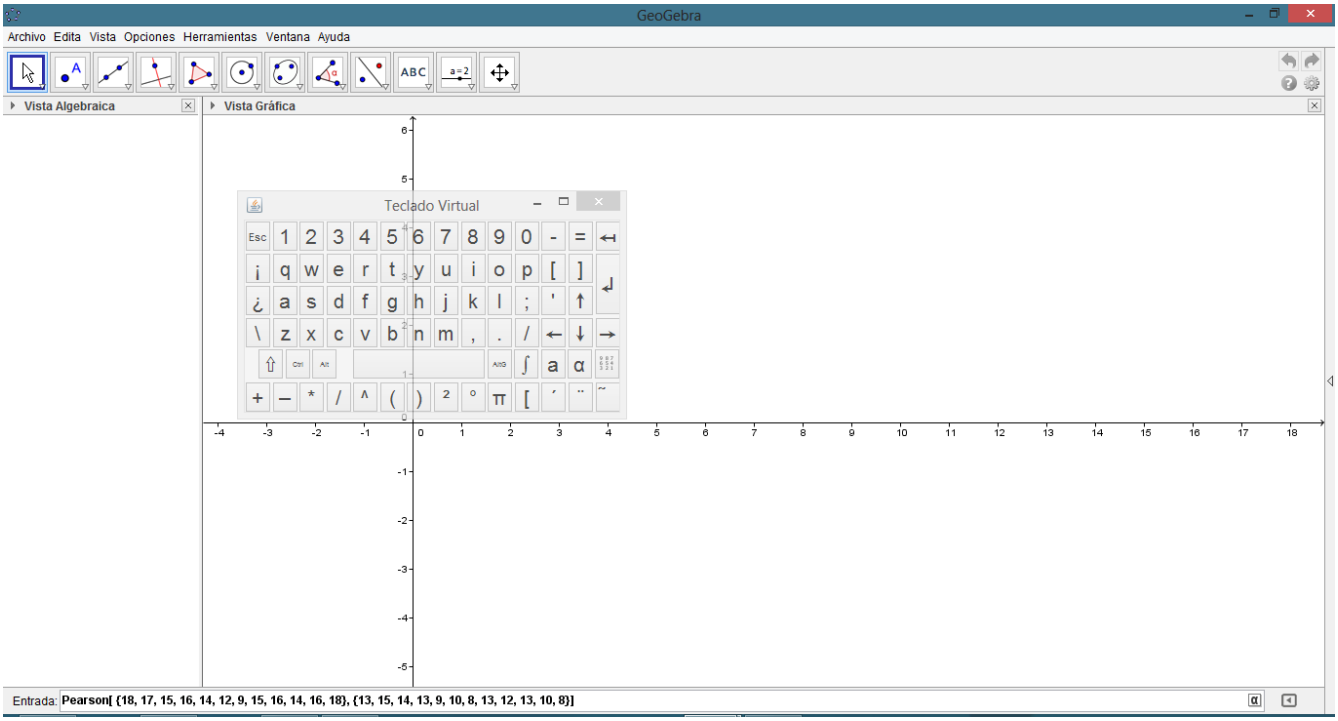


b) Seleccione la opción Pearson[<Lista de Coordenadas x>, <Lista de Coordenadas y>]

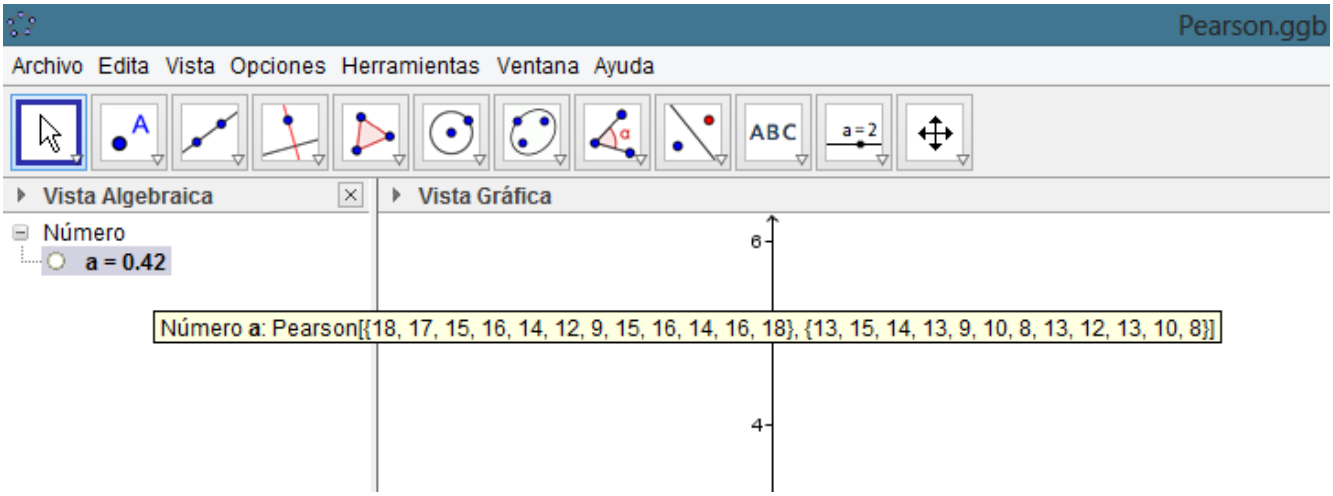


c) Escribir los datos de X y los datos de Y. Para escribir las llaves utilizar el teclado virtual:

Pearson[{18, 17, 15, 16, 14, 12, 9, 15, 16, 14, 16, 18}, {13, 15, 14, 13, 9, 10, 8, 13, 12, 13, 10, 8}]

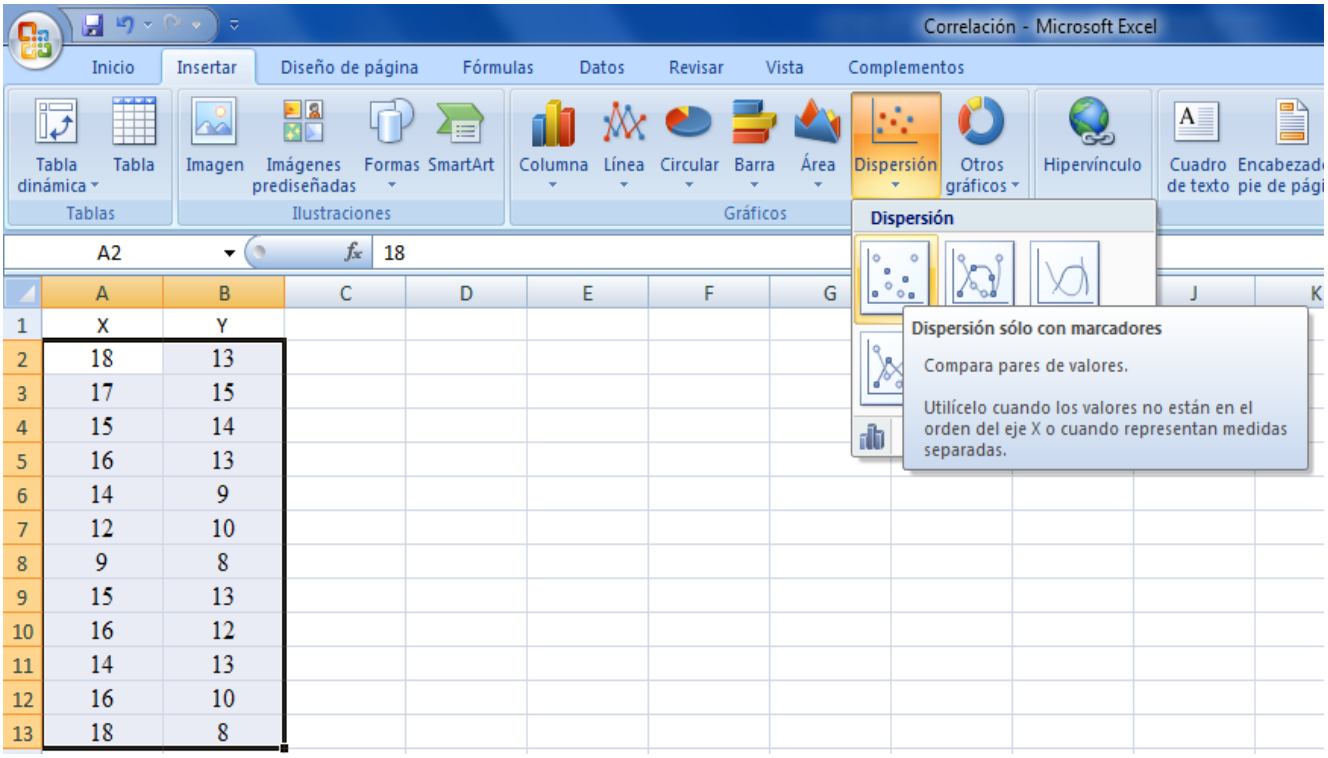


d) Enter

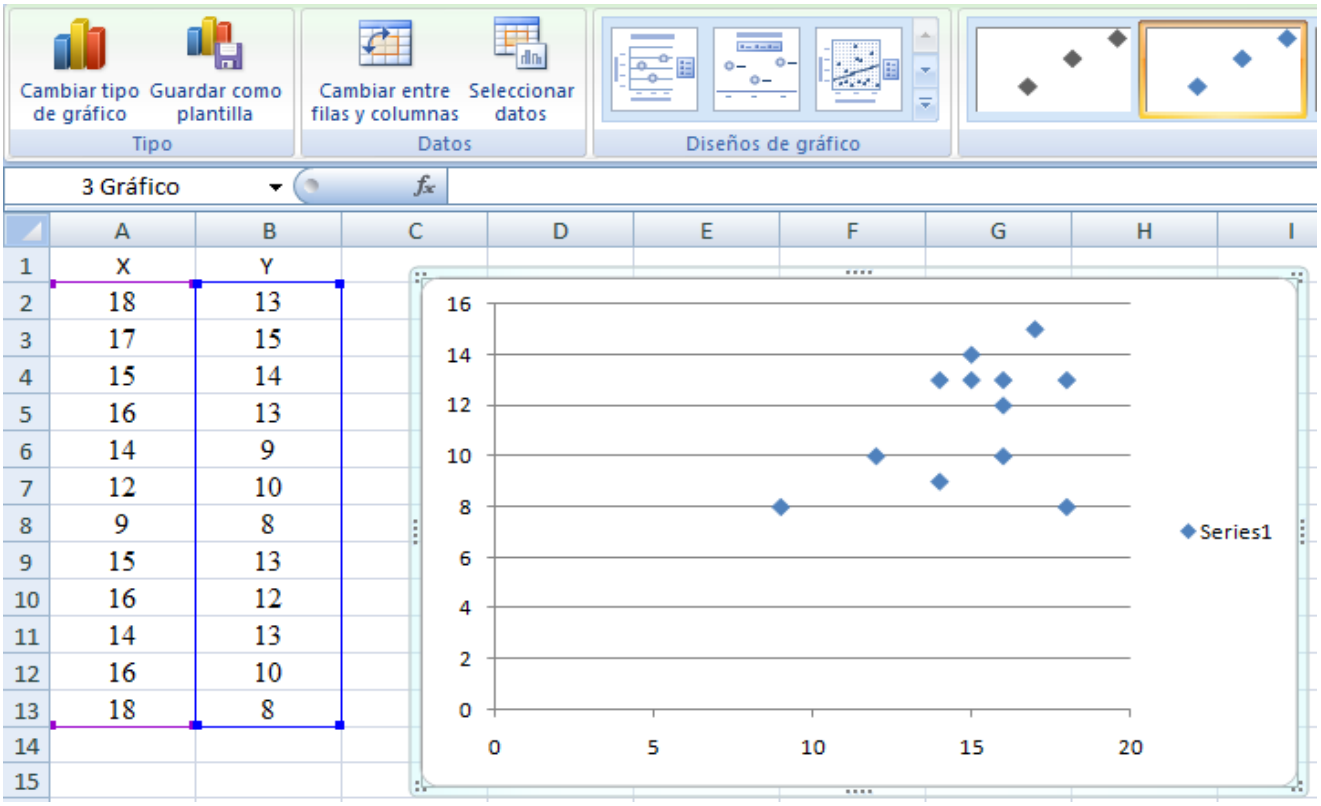


El diagrama de dispersión en Excel se realiza de la siguiente manera:

a) Seleccionar los datos e insertar diagrama de dispersión.

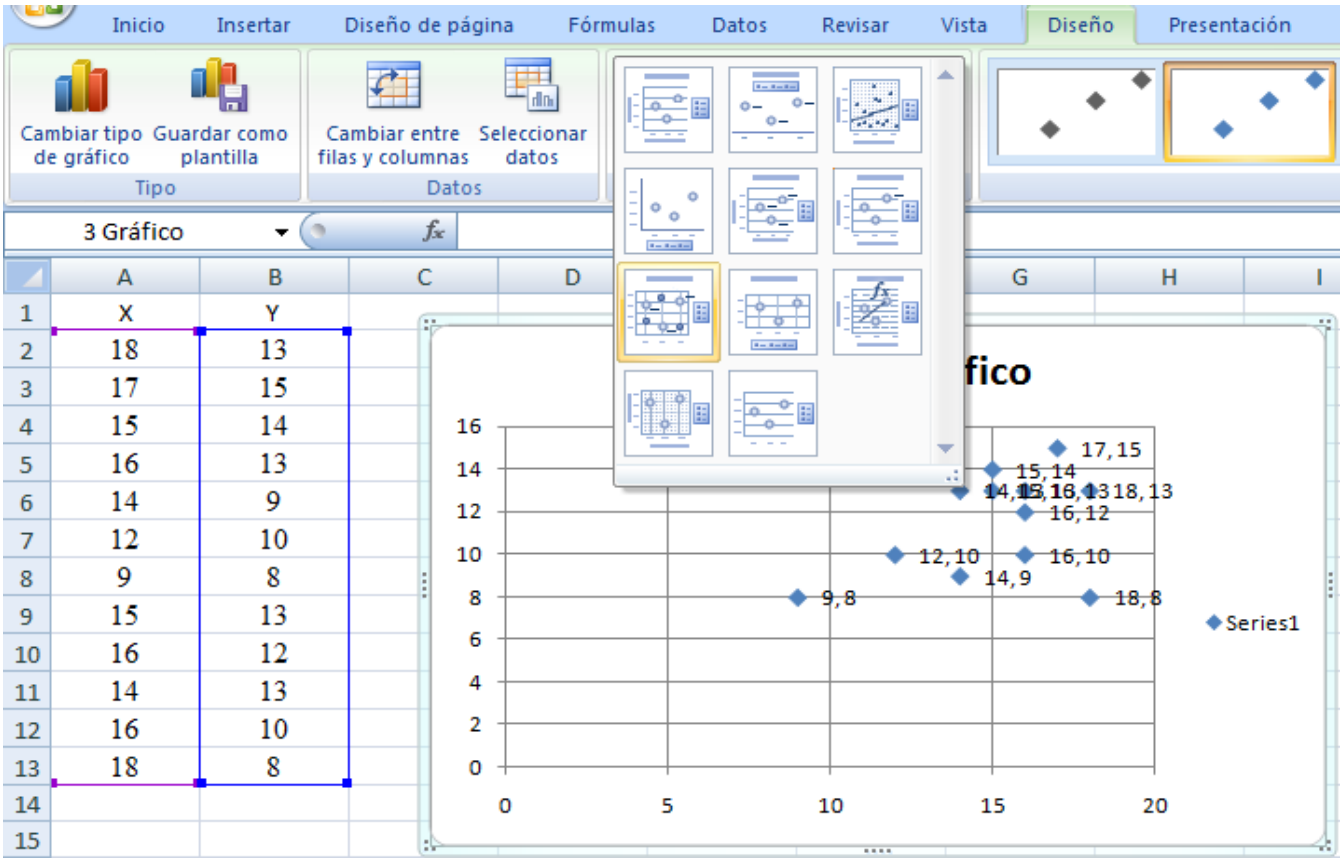


b) En diagrama dispersión, escoger el primero.



	A	B
1	X	Y
2	18	13
3	17	15
4	15	14
5	16	13
6	14	9
7	12	10
8	9	8
9	15	13
10	16	12
11	14	13
12	16	10
13	18	8
14		
15		

c) Para que ver las coordenadas escoger el diseño N° 7.

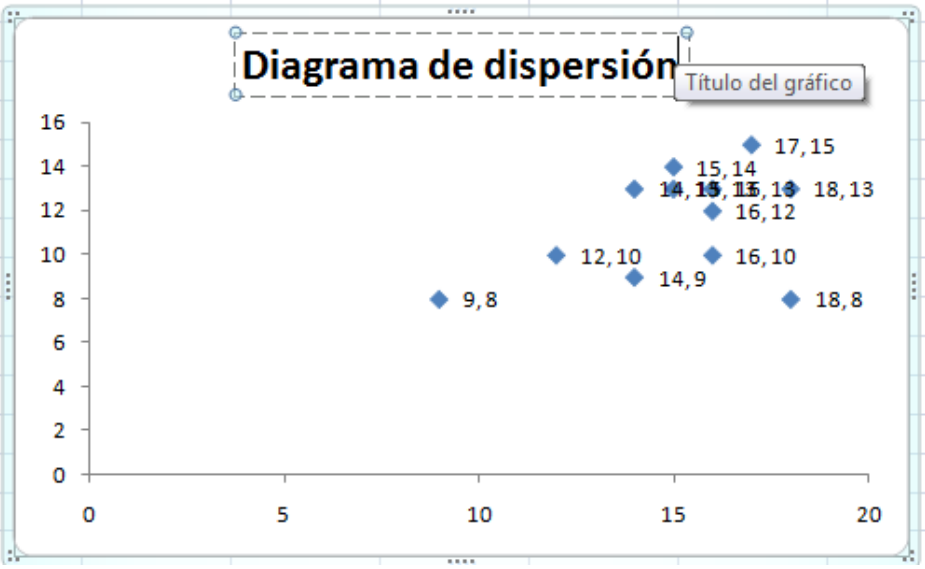


	A	B
1	X	Y
2	18	13
3	17	15
4	15	14
5	16	13
6	14	9
7	12	10
8	9	8
9	15	13
10	16	12
11	14	13
12	16	10
13	18	8
14		
15		

d) Borrar Serie 1, las líneas horizontales y verticales (haciendo clic y suprimir en cada objeto).

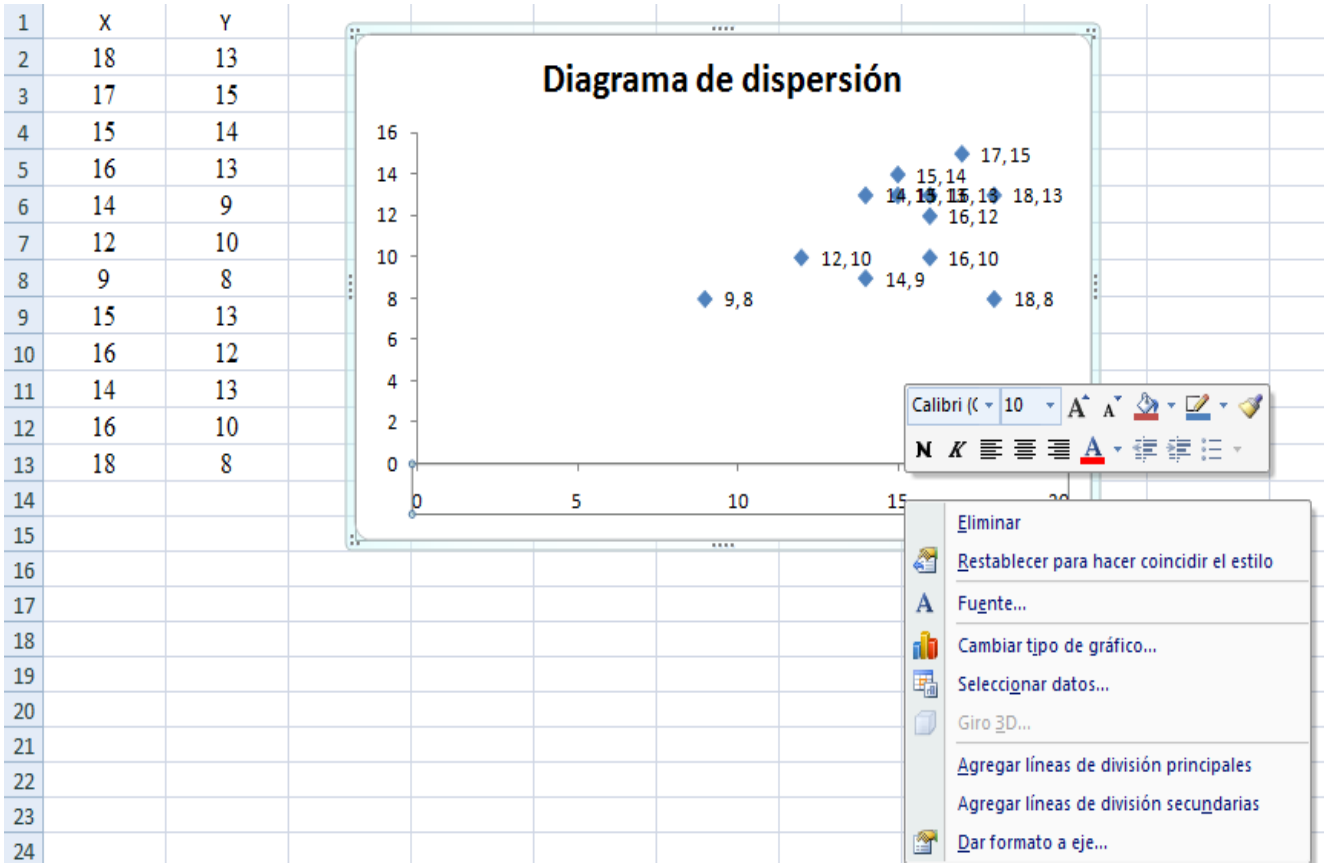
	A	B
1	X	Y
2	18	13
3	17	15
4	15	14
5	16	13
6	14	9
7	12	10
8	9	8
9	15	13
10	16	12
11	14	13
12	16	10
13	18	8
14		
15		

e) En título del gráfico escribir Diagrama de dispersión.

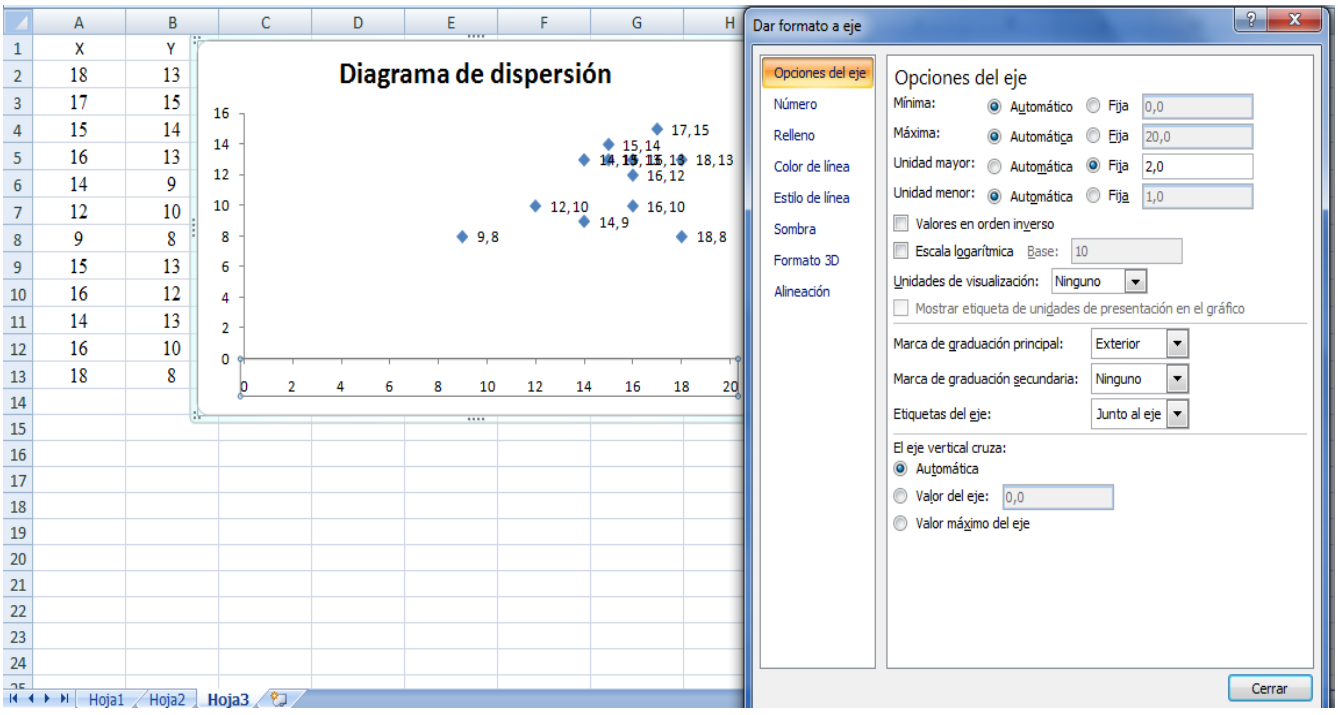


	X	Y
1		
2	18	13
3	17	15
4	15	14
5	16	13
6	14	9
7	12	10
8	9	8
9	15	13
10	16	12
11	14	13
12	16	10
13	18	8
14		
15		

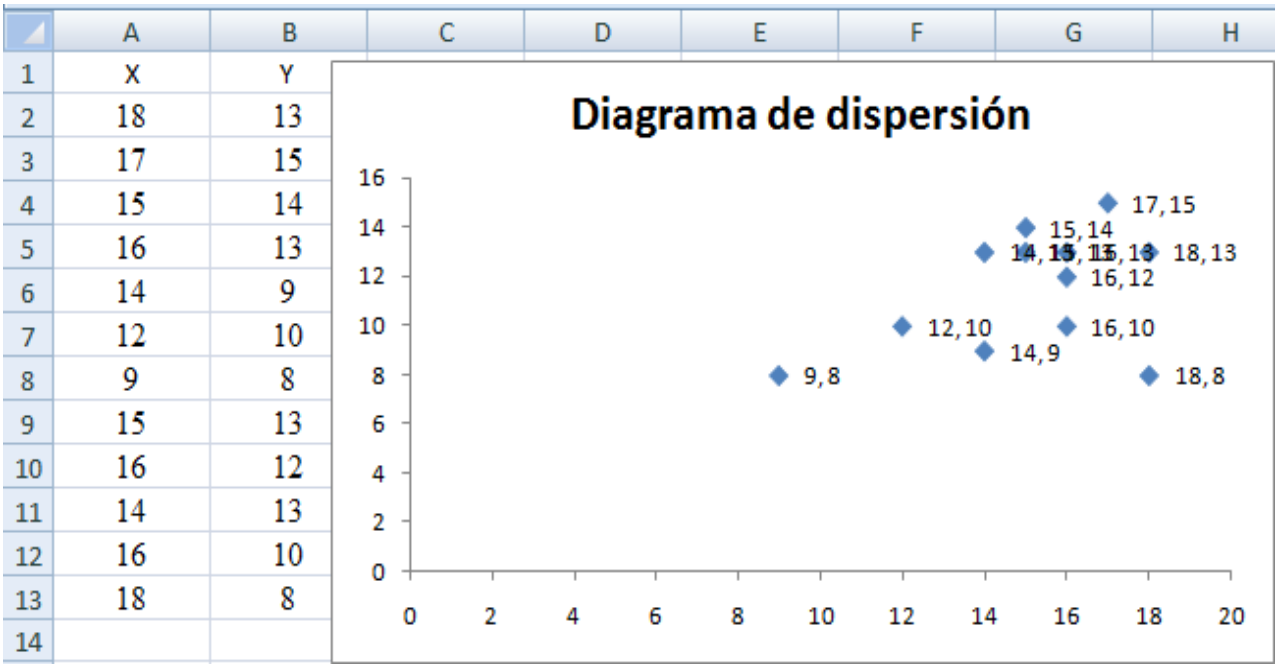
f) Clic en el eje x, y luego clic derecho para dar formato al eje.



g) Poner 2 en la casilla unidad mayor para ver los números de 2 en 2 en el eje x.



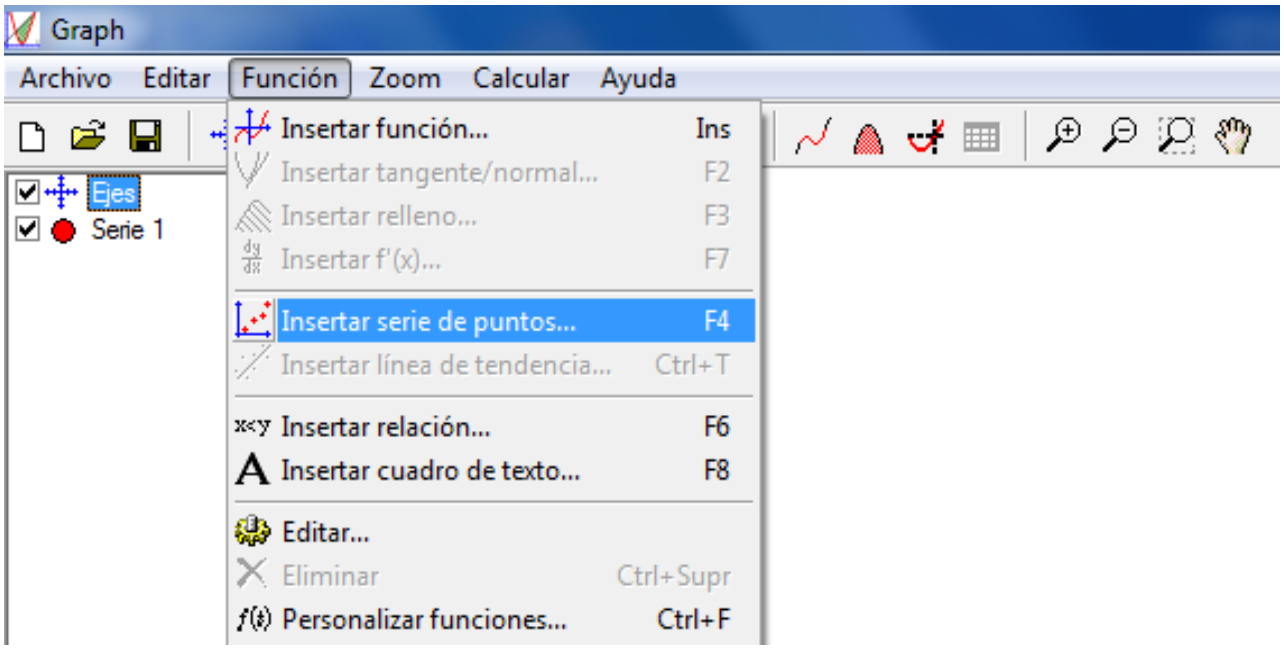
h) Clic en Cerrar para culminar la elaboración del diagrama de dispersión, aunque se le puede seguir haciendo más mejoras.



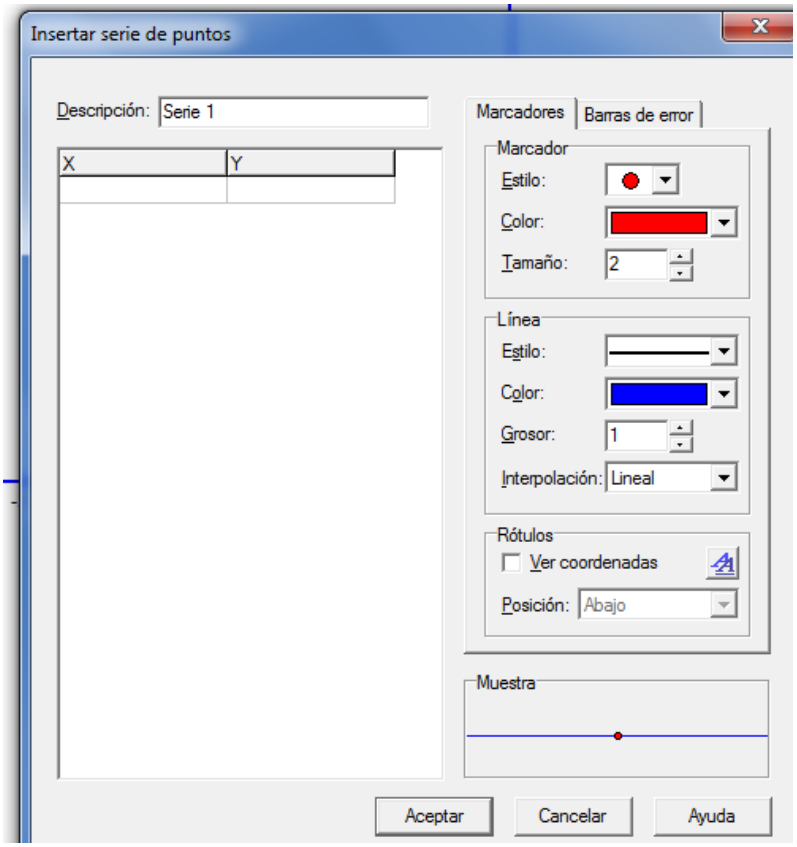
	A	B
1	X	Y
2	18	13
3	17	15
4	15	14
5	16	13
6	14	9
7	12	10
8	9	8
9	15	13
10	16	12
11	14	13
12	16	10
13	18	8
14		

Para realizar el diagrama de dispersión en el programa Graph se procede de la siguiente manera:

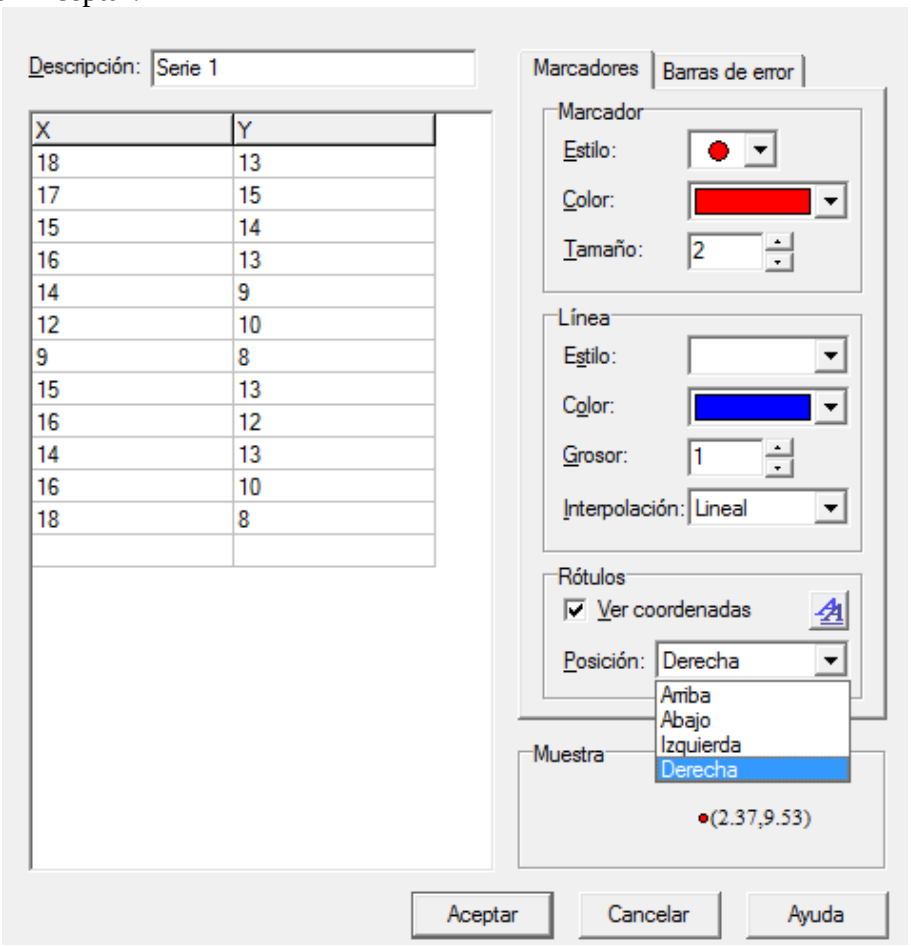
a) Clic en Función.



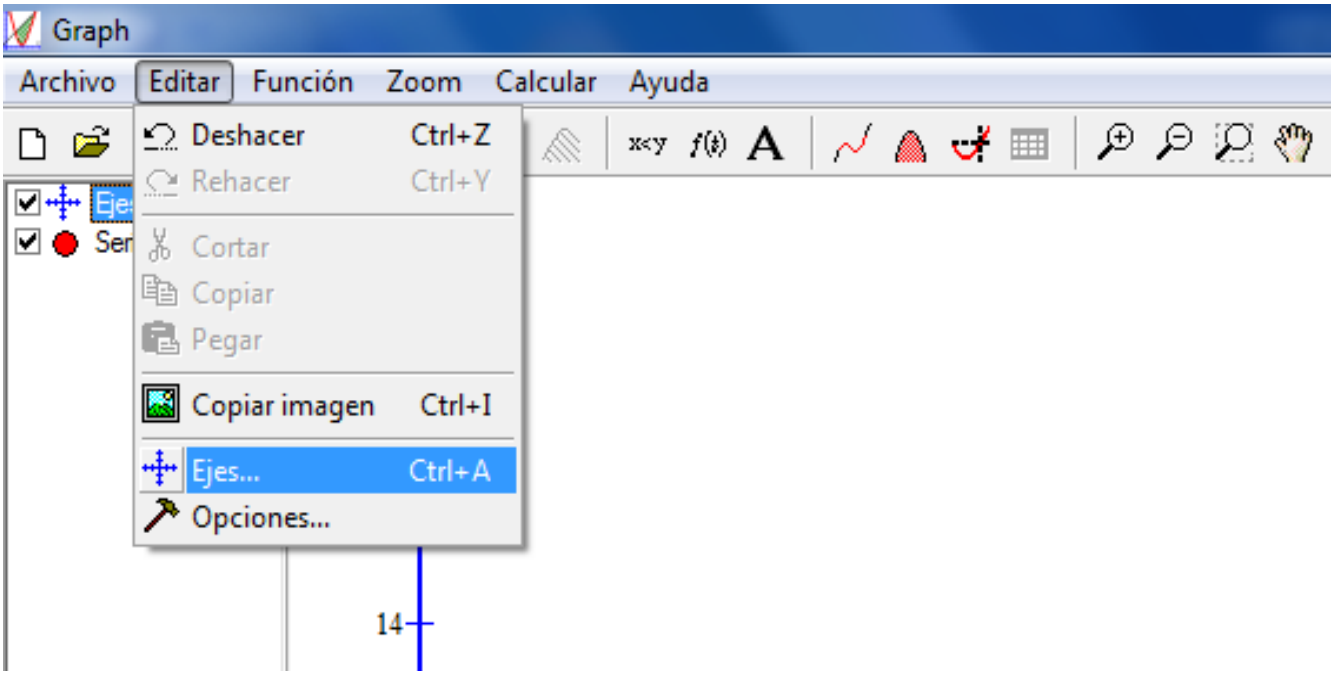
b) Clic en Insertar serie de puntos.



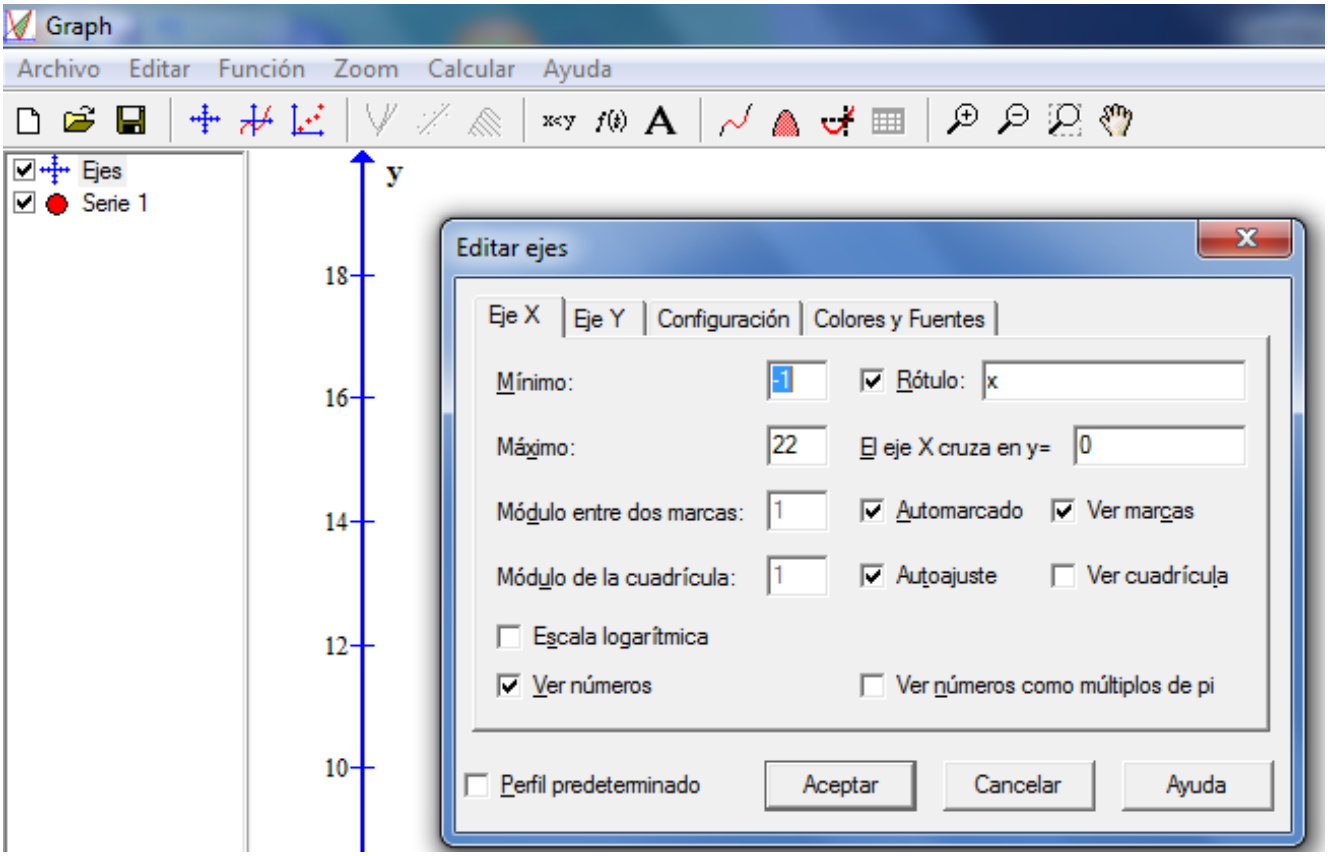
c) Escribir los puntos, y en estilo de línea, escoger sin línea. En rótulos poner en ver coordenadas a la derecha. Pulsar en Aceptar.



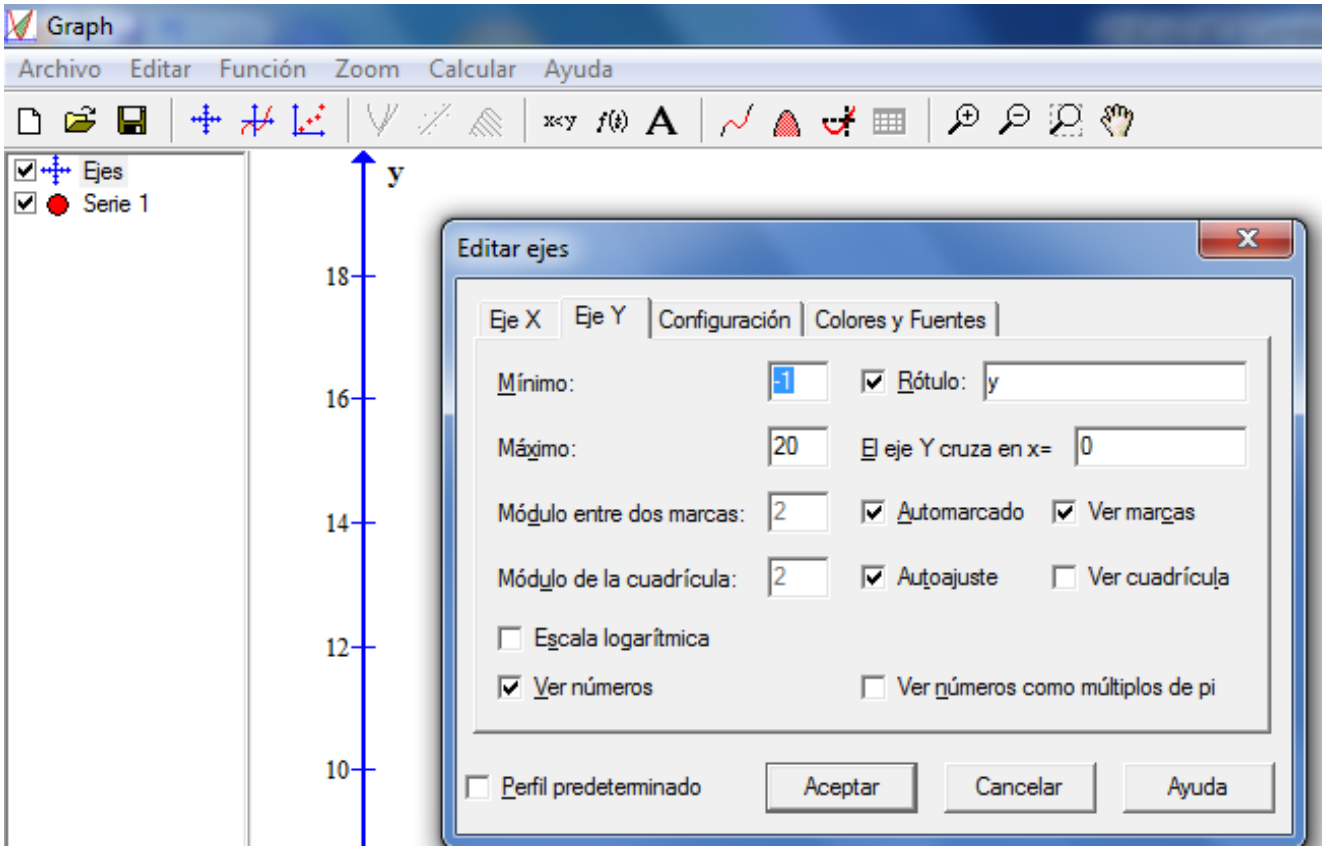
d) Para editar los ejes, hacer clic en Editar y luego en Ejes.



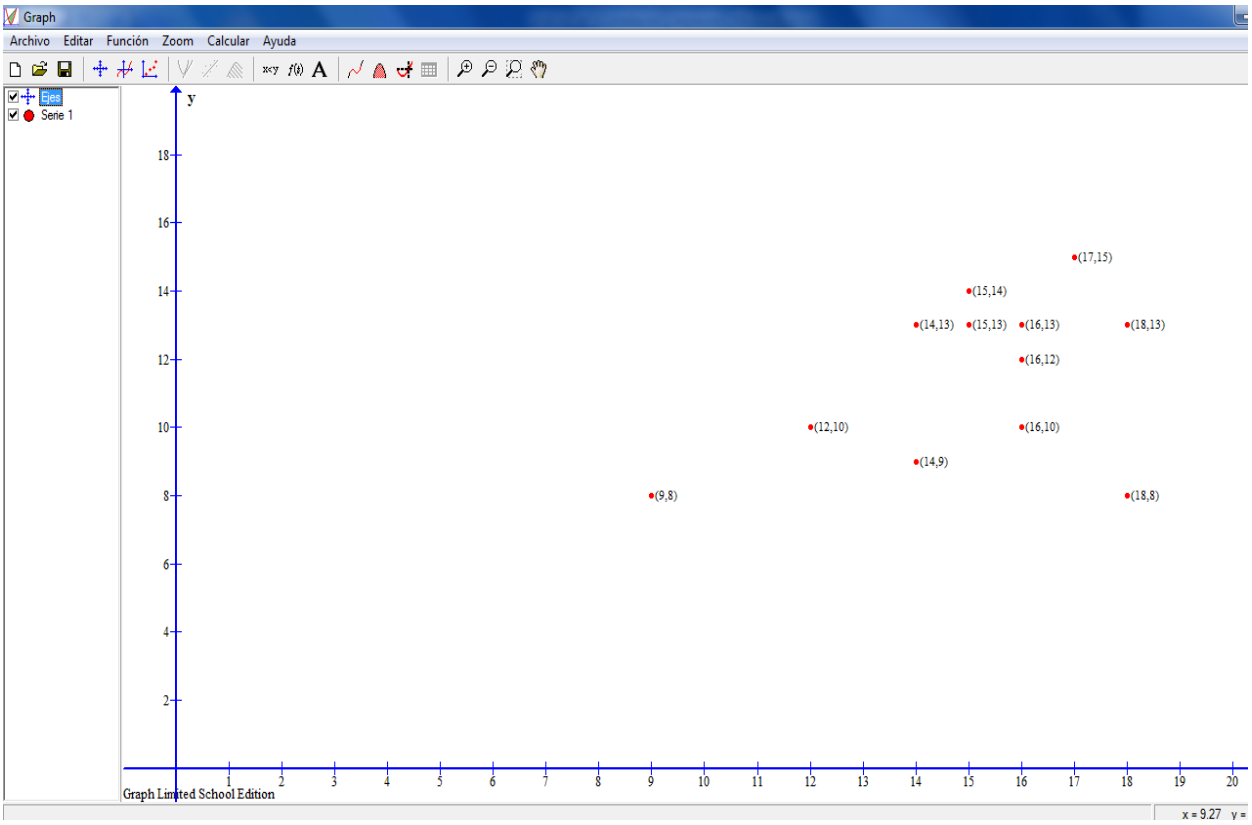
e) Llenar las casillas del Eje X de acuerdo a los datos del ejercicio.



f) Llenar las casillas del Eje Y de acuerdo a los datos del ejercicio.



g) Pulsar en Aceptar para dar por culminado la elaboración del diagrama de dispersión, el cual se presenta en la siguiente figura:



b) Para datos agrupados, el coeficiente de Correlación de Pearson se calcula aplicando la siguiente fórmula:

$$r = \frac{n \cdot \sum f \cdot dx \cdot dy - (\sum fx \cdot dx)\,(\sum fy \cdot dy)}{\sqrt{[n \cdot \sum fx \cdot dx^2 - (\sum fx \cdot dx)^2][n \cdot \sum fy \cdot dy^2 - (\sum fy \cdot dy)^2]}}$$

Donde:
n = número de datos.
f = frecuencia de celda.
fx = frecuencia de la variable X.
fy = frecuencia de la variable Y.
dx = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda $dx = 0$, para que se hagan más fáciles los cálculos.
dy = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda $dy = 0$, para que se hagan más fáciles los cálculos.

Ejemplo ilustrativo:
Con los siguientes datos sobre los Coeficientes Intelectuales (X) y de las calificaciones en una prueba de conocimiento (Y) de 50 estudiantes:

N° de estudiante	X	Y	N° de estudiante	X	Y
1	76	28	26	88	40
2	77	24	27	88	31
3	78	18	28	88	35
4	79	41	29	88	26
5	79	43	30	89	30
6	80	45	31	89	24
7	80	34	32	90	18
8	80	18	33	90	11
9	82	40	34	90	15
10	82	35	35	91	38
11	83	30	36	92	34
12	83	21	37	92	31
13	83	22	38	93	33
14	83	23	39	93	35
15	84	25	40	93	24
16	84	11	41	94	40
17	84	15	42	96	35
18	85	31	43	97	36
19	85	35	44	98	40
20	86	26	45	99	33
21	86	30	46	100	51
22	86	24	47	101	54
23	86	16	48	101	55
24	87	20	49	102	41
25	88	36	50	102	45

1) Elaborar una tabla de dos variables
2) Calcular el coeficiente de correlación

Solución:

1) En la *tabla de frecuencias de dos variables*, cada recuadro de esta tabla se llama una *celda* y corresponde a un par de intervalos, y el número indicado en cada celda se llama *frecuencia de celda*. Todos los totales indicados en la última fila y en la última columna se llaman *totales marginales o frecuencias marginales*, y corresponden, respectivamente, a las frecuencias de intervalo de las distribuciones de frecuencia separadas de la variable X y Y.

Para elaborar la tabla se recomienda:

- Agrupar las variables X y Y en un igual número de intervalos.

- Los intervalos de la variable X se ubican en la parte superior de manera horizontal (fila) y en orden ascendente.

- Los intervalos de la variable Y se ubican en la parte izquierda de manera vertical (columna) y en orden descendente.

Para elaborar los intervalos se procede a realizar los cálculos respectivos:

En la variable X:

Calculando el Rango se obtiene:
$R = x_{máx} - x_{mín} = 102 - 76 = 26$

Calculando el número de intervalos se obtiene:
$n_i = 1 + 3{,}32 \cdot log(n) = 1 + 3{,}32 \cdot log 50 = 6{,}6 = 7$

Calculando el ancho se obtiene:
$i = \frac{R}{ni} = \frac{26}{6{,}6} = 3{,}93 = 4$

En la variable Y:

Calculando el Rango se obtiene:
$R = y_{máx} - y_{mín} = 55 - 11 = 44$

Calculando el número de intervalos se obtiene:
$n_i = 1 + 3{,}32 \cdot log(n) = 1 + 3{,}32 \cdot log 50 = 6{,}64 = 7$

Calculando el ancho se obtiene:
$i = \frac{R}{ni} = \frac{44}{6{,}64} = 6{,}62 = 7$

Nota: Para la variable X se tomará un ancho de intervalo igual a 4 y para la variable Y un ancho de intervalo igual a 7. Debe quedar igual número de intervalos para cada variable, que en este ejemplo es igual a 7.

Contando las frecuencias de celda para cada par de intervalos de las variables X y Y se obtiene la siguiente tabla de frecuencias de dos variables:

		Coeficientes Intelectuales (X)							
		76-79	80-83	84-87	88-91	92-95	96-99	100-103	fy
Calificaciones (Y)	53-59							2	2
	46-52							1	1
	39-45	2	2		1	1	1	2	9
	32-38		2	1	3	3	3		12
	25-31	1	1	4	3	1			10
	18-24	2	4	2	2	1			11
	11-17			3	2				5
	fx	5	9	10	11	6	4	5	50

Interpretación:

- El número 2 es la frecuencia de la celda correspondiente al par de intervalos 76-79 en Coeficiente Intelectual y 39-45 en Calificación obtenida en la prueba de conocimiento.
- El número 5 en la fila de fx es el total marginal o frecuencia marginal del intervalo 76-79 en Coeficiente Intelectual.
- El número 2 en la columna de fy es el total marginal o frecuencia marginal del intervalo 53-59 en Calificación obtenida en la prueba de conocimiento.
- El número 50 es total de frecuencias marginales y representa al número total de estudiantes.

2) Realizando los cálculos respectivos se obtiene la siguiente tabla:

			Coeficientes Intelectuales (X)										
			76-79	80-83	84-87	88-91	92-95	96-99	100-103				
		dx / dy	-3	-2	-1	0	1	2	3	fy	$fy \cdot dy$	$fy \cdot dy^2$	$f \cdot dx \cdot dy$
Calificaciones (Y)	53-59	3							2 (18)	2	6	18	18
	46-52	2							1 (6)	1	2	4	6
	39-45	1	2 (-6)	2 (-4)		1 (0)	1 (1)	1 (2)	2 (6)	9	9	9	-1
	32-38	0		2 (0)	1 (0)	3 (0)	3 (0)	3 (0)		12	0	0	0
	25-31	-1	1 (3)	1 (2)	4 (4)	3 (0)	1 (-1)			10	-10	10	8
	18-24	-2	2 (12)	4 (16)	2 (4)	2 (0)	1 (-2)			11	-22	44	30
	11-17	-3			3 (9)	2 (0)				5	-15	45	9
	fx		5	9	10	11	6	4	5	50	-30	130	70
	$fx \cdot dx$		-15	-18	-10	0	6	8	15	-14			
	$fx \cdot dx^2$		45	36	10	0	6	16	45	158			
	$f \cdot dx \cdot dy$		9	14	17	0	-2	2	30	70			

Nota:

Los números de las esquinas de cada celda en la anterior tabla representan el producto $f \cdot dx \cdot dy$, así por ejemplo, para obtener el número el número -6 de los intervalos 76-79 en X y 39-45 en

Y se obtiene multiplicando 2·(-3)·1 = -6. Para obtener el número 18 de los intervalos 100-103 en X y 53-59 en Y se obtiene multiplicando 2·3·3 = 18

Los números de la última columna (18, 6, -1, 0, 8, 30 y 9) se obtienen sumando los números de las esquinas en cada fila, así por ejemplo, para obtener el número -1 se suma (-6) + (-4) + 0 + 1 +2 + 6 = -1

Los números de la última fila (9, 14, 17, 0, -2, 2 y 30) se obtienen sumando los números de las esquinas en cada columna, así por ejemplo, para obtener el número 9 se suma (-6) + 3 + 12 = 9.

Para obtener el número -30 de la antepenúlmina columna se obtiene sumando los resultados de $fy \cdot dy$, es decir, representa la $\sum fy \cdot dy$

Para obtener el número -14 de la antepenúlmina fila se obtiene sumando los resultados de $fx \cdot dx$, es decir, representa la $\sum fx \cdot dx$

Para obtener el número 130 de la penúltima columna se obtiene sumando los resultados de $fy \cdot dy^2$, es decir, representa $\sum fy \cdot dy^2$

Para obtener el número 158 de la penúltima fila se obtiene sumando los resultados de $fx \cdot dx^2$, es decir, representa $\sum fx \cdot dx^2$

Para obtener último número 70 de la última columna se obtiene sumando los resultados de la última columna 18 +6 +(-1) +0 + 8 + 30 +9 = 70, es decir, representa $\sum fx \cdot dx \cdot dy$

Para obtener último número 70 de la última fila se obtiene sumando los resultados de la última fila 9 + 14 + 17 + 0 +(-2) +2 + 30 = 70 , es decir, representa $\sum fx \cdot dx \cdot dy$. Por lo tanto tiene que ser igual al último número de la última columna como comprobación que los cálculos de la tabla han sido correctos.

Observando los datos en la tabla anterior se reemplaza los valores en la ecuación del Coeficiente de Correlación de Pearson para datos agrupados, obteniéndose:

$$r = \frac{n \cdot \sum f \cdot dx \cdot dy - (\sum fx \cdot dx)\,(\sum fy \cdot dy)}{\sqrt{[n \cdot \sum fx \cdot dx^2 - (\sum fx \cdot dx)^2][n \cdot \sum fy \cdot dy^2 - (\sum fy \cdot dy)^2]}}$$

$$r = \frac{50 \cdot 70 - (-14)(-30)}{\sqrt{[50 \cdot 158 - (-14)^2][50 \cdot 130 - (-30)^2]}} = \frac{3500 - 420}{\sqrt{[7900 - 196][6500 - 900]}} = \frac{3080}{\sqrt{[7704][5600]}}$$

$$r = \frac{3080}{\sqrt{43142400}} = \frac{3080}{6568{,}287448} = 0{,}469$$

Existe una correlación positiva moderada

TAREA DE INTERAPRENDIZAJE

1) Elabore un organizador gráfico de los tipos de correlación.

2) Con los datos de la siguiente tabla sobre las temperaturas del día X y del día Y en determinadas horas en una ciudad

X	9	10	12	14	16	18	20	22	24	26	28	30
Y	12	14	15	16	17	20	22	23	26	28	31	32

2.1) Calcule el coeficiente de correlación de Pearson empleando la fórmula y mediante Excel.

0,99

2.2) Elabore el diagrama de dispersión de manera manual.
2.3) Elabore el diagrama de dispersión empleando Excel.
2.4) Elabore el diagrama de dispersión empleando el programa Graph.

3) Cree y resuelva un ejercicio similar al anterior.

4) Dada la siguiente tabla de frecuencias de dos variables, con los datos sobre las calificaciones obtenidos en un curso de 50 estudiantes en la asignatura de Matemática (X) y en la asignatura de Estadística (Y), determinar el tipo de correlación que existe entre ellas mediante el coeficiente de Pearson.

		X					
		1-2	3-4	5-6	7-8	9-10	fy
Y	9-10				7	8	15
	7-8				6		6
	5-6			3	4		7
	3-4	5	5	1			11
	1-2	7	4				11
	fx	12	9	4	17	8	50

Correlación positiva muy alta de 0,91

5) Dada la siguiente tabla de frecuencias de dos variables, con los datos sobre los pesos en kilogramos en dos barrios diferentes en una ciudad, determinar el tipo de correlación que existe entre ellas mediante el coeficiente de Pearson.

		X						
		40-49	50-59	60-69	70-79	80-89	90-99	fy
Y	90-99				3	3	4	10
	80-89			8	2	2	4	16
	70-79			5	10	8	1	24
	60-69	8	1	2	5	2		18
	50-59	3	10	6	2			21
	40-49	4	6	1				11
	fx	15	17	22	22	15	9	100

Correlación positiva moderada de 0,688

6) Dada la siguiente tabla de frecuencias de dos variables, con los datos sobre las calificaciones obtenidos en un curso de 100 estudiantes en la asignatura de Matemática (X) y en la asignatura de Estadística (Y),

determinar el tipo de correlación que existe entre ellas mediante el coeficiente de Pearson para datos agrupados.

N° de estudiante	X	Y	N° de estudiante	X	Y	N° de estudiante	X	Y	N° de estudiante	X	Y
1	40	60	26	57	73	51	71	86	76	84	83
2	41	50	27	58	78	52	72	88	77	84	84
3	42	55	28	60	79	53	72	89	78	85	86
4	43	59	29	61	60	54	72	70	79	86	88
5	44	40	30	62	61	55	73	71	80	86	89
6	45	42	31	63	62	56	74	72	81	86	70
7	45	49	32	64	63	57	74	73	82	87	78
8	45	60	33	64	64	58	74	74	83	87	79
9	45	62	34	65	65	59	75	75	84	88	78
10	48	66	35	65	66	60	76	76	85	88	77
11	49	69	36	66	67	61	76	77	86	88	79
12	50	50	37	66	69	62	77	78	87	88	78
13	50	52	38	66	50	63	77	79	88	89	78
14	56	54	39	66	52	64	78	60	89	89	60
15	56	56	40	67	55	65	78	67	90	89	69
16	56	59	41	68	56	66	78	65	91	90	90
17	56	59	42	68	57	67	78	68	92	91	96
18	56	40	43	68	59	68	79	69	93	92	97
19	57	45	44	69	40	69	79	50	94	93	99
20	57	47	45	69	45	70	79	59	95	94	80
21	57	48	46	69	47	71	80	90	96	95	81
22	57	49	47	69	49	72	81	94	97	96	82
23	57	80	48	70	90	73	82	96	98	97	83
24	57	70	49	70	99	74	82	99	99	98	89
25	57	72	50	70	80	75	83	80	100	99	70

	X	40-48						94-102	
Y	dx / dy	-3	-2	-1	0	1	2	3	fy
94-102	3								7
	2								
	1								
	0								
	-1	5							
	-2	3							
40-48	-3	2							9
	fx	10							100

Correlación positiva moderada de 0,62

7) Cree y resuelva un ejercicio similar al anterior.

8) Consulte en la biblioteca o en el internet un ejercicio resuelto sobre el coeficiente de correlación de Pearson para datos agrupados en intervalos.