ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

VARIABLES BIDIMENSIONALES

Hasta ahora las series estadísticas estudiadas estaban asociadas a variables estadísticas unidimensionales, es decir se estudiaba un solo carácter de la población. Sin embargo, en ocasiones es útil considerar a la vez varios caracteres de una misma población: la estatura, la edad y el peso de un grupo de 50 niños; el sueldo y la cualificación de un conjunto de asalariados; la extensión y el número de habitantes de los países europeos, la inversión en publicidad y la facturación de ciertas empresas, etc. Nosotros reduciremos el estudio de una población a dos caracteres, teniendo así variables estadísticas bidimensionales.

Las variables estadísticas bidimensionales las representamos por un par (X,Y) donde X es una variable estadística unidimensional que toma los valores x_1, x_2, x_3, ... e Y es otra variable unidimensional que toma los valores y_1, y_2, y_3, ...

PARÁMETROS ESTADÍSTICOS BIDIMENSIONALES

Al igual que las variables estadísticas unidimensionales, las bidimensionales también se pueden expresar mediante una tabla. En este caso, se pueden hacer dos tipos de tablas: simple y de doble entrada.

Ejemplo C: *Las calificaciones de 40 alumnos en Matemáticas y Física han sido las siguientes:*

X=Calificación Matemáticas	3	3	4	5	5	6	6	6	7	7	8
Y=Calificación Física	2	5	5	4	5	4	6	7	6	7	9
Número de alumnos	4	3	3	2	10	2	4	5	4	2	1

Esta información puede disponerse en una tabla de doble entrada:

Y \ X	**3**	**4**	**5**	**6**	**7**	**8**	
2	4						4
4			2	2			4
5	3	3	10				16
6				4	4		8
7				5	2		7
9						1	1
	7	3	12	11	6	1	40

Nosotros usaremos una tabla simple similar a las usadas hasta ahora:

x_i	y_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
3	2	4	12	36	8	16	24
3	5	3	9	27	15	75	45
4	5	3	12	48	15	75	60
5	4	2	10	50	8	32	40
5	5	10	50	250	50	250	250
6	4	2	12	72	8	32	48
6	6	4	24	144	24	144	144
6	7	5	30	180	35	245	210
7	6	4	28	196	24	144	168
7	7	2	14	98	14	98	98
8	9	1	8	64	9	81	72
		40	209	1165	210	1192	1159

Obsérvese que hemos añadido una columna nueva ($x_i \cdot y_i \cdot f_i$) con el fin de calcular un nuevo parámetro estadístico llamado:

Covarianza

Se llama covarianza de una variable bidimensional (X;Y) a la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas. Se representa por s_{xy} y su expresión es:

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i}{N} = \frac{\sum x_i \cdot y_i \cdot f_i}{N} - \bar{x} \cdot \bar{y}$$

Los cálculos que normalmente haremos son los siguientes (para la tabla anterior):

$$\bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{209}{40} = 5,225 \qquad \bar{y} = \frac{\sum y_i \cdot f_i}{N} = \frac{210}{40} = 5,25$$

$$s_x^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{1165}{40} - (5,225)^2 = 1,824 \qquad s_y^2 = \frac{\sum y_i^2 \cdot f_i}{N} - \bar{y}^2 = \frac{1192}{40} - (5,25)^2 = 2,238$$

$$s_x = \sqrt{s_x^2} = \sqrt{1,824} = 1,351 \qquad s_y = \sqrt{s_y^2} = \sqrt{2,238} = 1,496$$

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i}{N} = \frac{\sum x_i \cdot y_i \cdot f_i}{N} - \bar{x} \cdot \bar{y} = \frac{1159}{40} - (5,225) \cdot (5,25) = 1,544$$

Correlación y Regresión

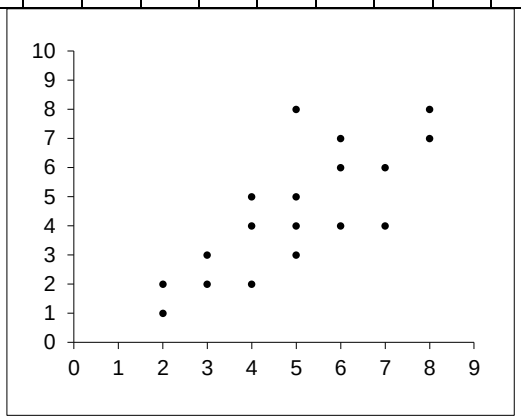
Al realizar un estudio de dos o más variables sobre un mismo colectivo, normalmente se hace para averiguar si existe alguna relación o dependencia entre ellas. A este hecho se le conoce con el nombre de **correlación**.

Una vez conocida la correlación que existe entre las dos variables, trataremos de encontrar una función matemática que relacione las dos variables, de tal manera que conocido un valor de una de ellas, sea posible calcular, con mayor o menor aproximación (dependiendo del grado de correlación), el correspondiente valor de la otra. Esto se denomina **regresión**.

Para encontrar la posible dependencia que existe entre las dos variables nos vamos a fijar en el **diagrama de dispersión** o **nube de puntos**, que no es más que la representación gráfica en el plano cartesiano de los pares (x_i, y_j). Veamos varios ejemplos:

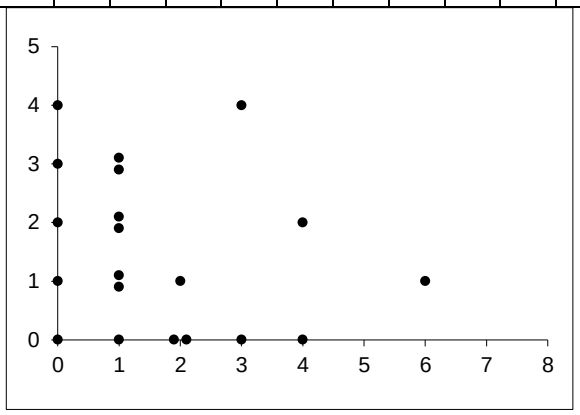
Ejemplo D: *Las notas de 18 alumnos en Matemáticas y Física son:*

Matemáticas:	2	3	4	4	5	6	6	2	7	8	5	4	6	5	8	7	3	5
Física:	1	3	2	4	4	4	6	2	6	7	8	5	7	5	8	4	2	3



Ejemplo E: *Se han clasificado 20 familias con arreglo al número de hijas e hijos obteniéndose los siguientes resultados:*

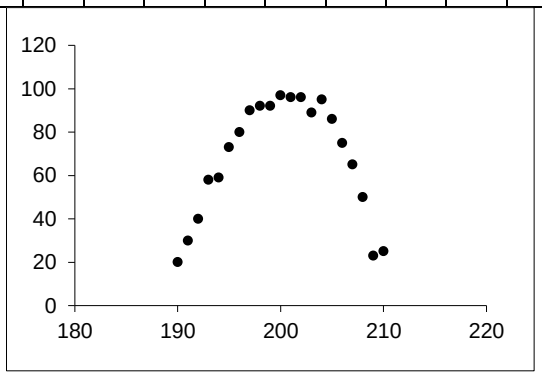
Hijos:	0	1	1	0	0	1	6	3	4	2	0	2	0	3	2	1	1	4	1	1
Hijas:	0	2	2	2	4	3	1	0	2	1	3	0	1	4	0	1	3	0	0	1



Cuando algún valor se repite se puede dibujar un punto más grueso o, como hemos hecho aquí dibujar varios puntos juntos.

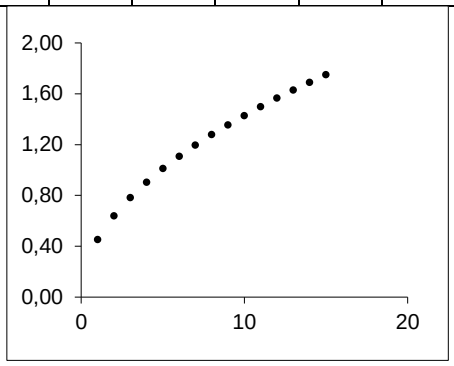
Ejemplo F: *Durante en un año, en un ensayo clínico, se han anotado la cantidad diaria de medicamento suministrado (en miligramos) a una serie de enfermos y el porcentaje de curación asociado, obteniéndose los siguientes datos:*

Dosis (en mg):	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210
% curación	20	30	40	58	59	73	80	90	92	92	97	96	96	89	95	86	75	65	50	23	25



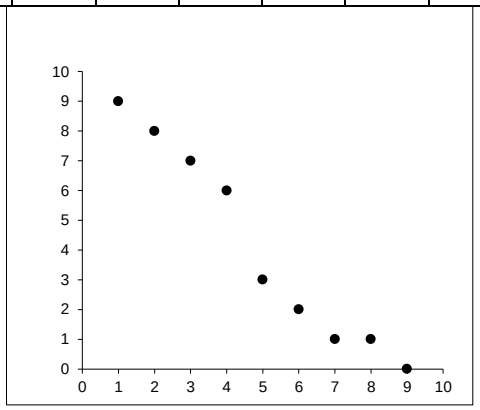
Ejemplo G: *Se ha dejado caer una canica desde diferentes alturas (en metros) y se ha medido el tiempo que ha tardado en llegar al suelo (en segundos):*

Altura:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tiempo:	0,45	0,64	0,78	0,90	1,01	1,11	1,20	1,28	1,36	1,43	1,50	1,56	1,63	1,69	1,75



Ejemplo H: *Un jugador de baloncesto lanza a canasta desde distintas distancias (en metros), 10 balones cada vez. Se anotan los encestes desde cada distancia:*

Distancia:	1	2	3	4	5	6	7	8	9
Encestes::	9	8	7	6	3	2	1	1	0



Dependiendo de cómo se distribuyen los puntos diremos que:

a) La correlación es lineal o curvilínea si la nube de puntos se condensa en torno a una línea recta (ejemplos D, G y H) o a una curva (ejemplo F).

b) La correlación es débil o fuerte según el grado de dicha condensación (es fuerte en los ejemplos F y H y débil en el ejemplo D). En caso de que la nube de puntos se ajuste perfectamente a una curva (ejemplo G) la dependencia se llama funcional.

c) La correlación es positiva o directa cuando a medida que crece una variable la otra también crece (ejemplos D y G). Será negativa o inversa cuando a medida que crece una variable la otra decrece (ejemplo H)

En caso de que la nube de puntos adopte una forma amorfa y no se ajuste a ninguna curva (ejemplo E) diremos que no existe correlación y se habla entonces de variables incorreladas o independientes.

Nosotros nos limitaremos a estudiar la correlación lineal, es decir, el grado en que una nube de puntos se ajusta a una línea recta. Trataremos de cuantificar esa correlación mediante un parámetro que llamaremos:

Coeficiente de correlación lineal (de Pearson)

El coeficiente de correlación lineal de Pearson es un número r comprendido entre –1 y 1 que mide hasta que punto puede representarse mediante una función del tipo $y = mx + n$ (una recta) la correlación existente entre dos variables. Se define como:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Si r es 1 o –1 existe dependencia funcional entre ambas variables expresable mediante una fórmula del tipo $y = mx + n$.

Si r está próximo a 1 (respectivamente a –1) existe correlación lineal fuerte y positiva (respectivamente negativa) entre las variables.

Si r está próximo a 0, apenas existe correlación lineal entre las variables

A partir de la fórmula, es obvio que el signo del coeficiente r viene dado por el signo de la covarianza ya que las desviaciones típicas son siempre positivas. Por tanto si:

$s_{xy} > 0 \Rightarrow$ correlación directa o positiva

$s_{xy} < 0 \Rightarrow$ correlación inversa o negativa.

$s_{xy} = 0 \Rightarrow$ no existe correlación (variables independientes)

A modo de ejemplo los coeficientes de correlación de Pearson de los ejemplos vistos anteriormente son:

Ejemplo D: 0,77 (correlación lineal moderadamente fuerte y positiva)

Ejemplo E: -0,23 (correlación lineal muy débil y negativa)

Ejemplo F: 0,10 (correlación lineal prácticamente inexistente)

Ejemplo G: 0,99 (correlación lineal muy fuerte y positiva)

Ejemplo H: -0,98 (correlación lineal muy fuerte y negativa)

NOTAS IMPORTANTES:

A) El coeficiente de correlación lineal de Pearson sirve, como indica su nombre, para detectar únicamente la correlación lineal. Si nos fijamos en la nube de puntos del ejemplo F, es evidente que las variables están fuertemente correlacionadas y sin embargo el coeficiente de correlación es tan solo de 0,10. Esto es así ya que la nube de puntos no se ajusta a una recta sino a lo que parece ser una parábola. Es decir: que el coeficiente de correlación este próximo a 0 significa, tan solo, que la nube de puntos no se ajusta a una recta.

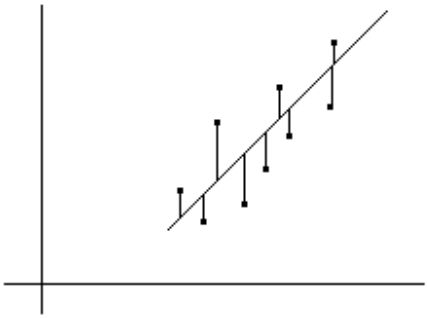
B) El concepto de correlación es delicado. Un error muy común es identificar correlación con causalidad, es decir, pensar que si dos variables están correlacionadas una es causa de la otra. Aunque en ocasiones se dé esta causalidad, no es difícil encontrar ejemplos en los que no. Supongamos que estudiamos la evolución de las ventas de automóviles y del consumo de leche en un determinado país en vías de desarrollo. Es posible que encontremos que ambas variables están correlacionadas. ¿Quiere esto decir que la compra de un automóvil implica un aumento en el consumo de leche? o, ¿el consumo de leche produce un aumento en las ventas de automóviles? Claramente la respuesta a estas preguntas debe ser negativa. Posiblemente lo que ocurre es que la variación de ambas variables se deriva de un tercer factor (por ejemplo la prosperidad relativa de dicho país, que hace aumentar tanto la venta de automóviles como el consumo de leche).

Supongamos ahora que estudiamos la evolución de las ventas de automóviles y de los casos de SIDA en España en la década 1990-2000. Es muy posible que exista correlación entre ambas variables. Y sin embargo es evidente que los automóviles no producen SIDA y más evidente, si cabe, que los enfermos de SIDA no se compran más automóviles que los que no padecen esa enfermedad. Ante un caso como este se suele hablar de "*correlación falsa*".

Una vez que se ha detectado que existe un alto grado de correlación lineal entre dos variables, surge el problema de encontrar la recta que mejor describa esa correlación, es decir, la recta $y = mx + n$ que mejor se ajuste a la nube de puntos. Dicha recta recibe el nombre de **recta de regresión.**

El procedimiento que se sigue es el llamado "Método de los mínimos cuadrados":

De entre todas las rectas nos quedamos con aquella para la cual la suma de los cuadrados de las distancias de cada uno de los puntos de la nube a la recta sea la mínima posible



De este modo se llega, usando métodos matemáticos superiores a este curso, a que:

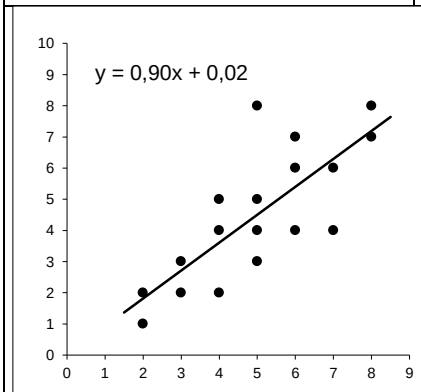
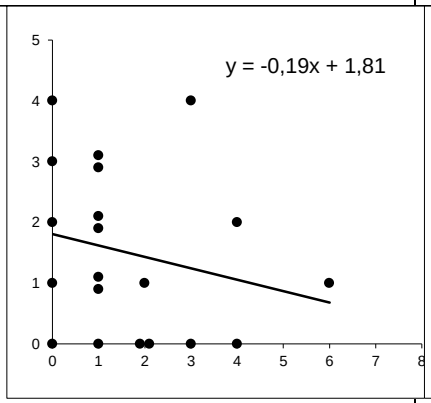
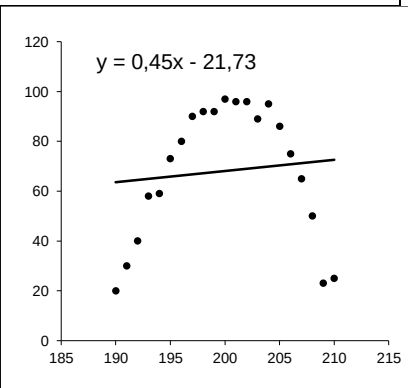
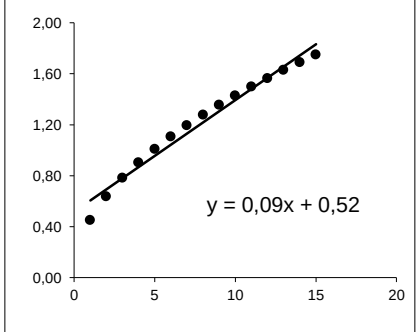
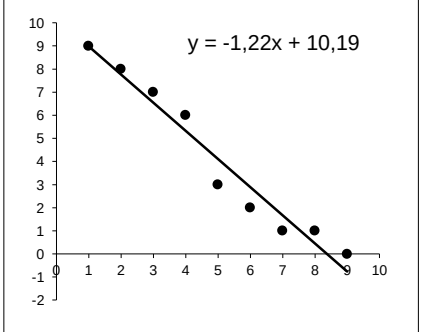
a) La recta pasa por el punto $(\overline{x}, \overline{y})$

b) Su pendiente es $\frac{s_{xy}}{s_x^2}$

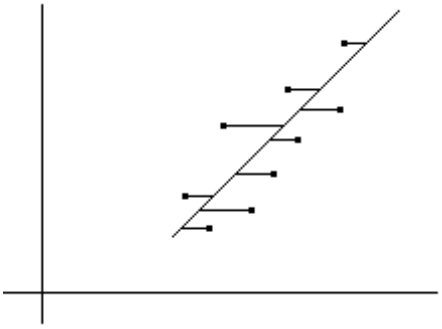
Por lo tanto la recta de regresión de Y sobre X tiene por ecuación:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

En los ejemplos anteriores las rectas quedarían de la siguiente forma:

Ejemplo D (r=0,77):	Ejemplo E (r=-0,23):	Ejemplo F (r=0,10):
		
Ejemplo G (r=0,99):	**Ejemplo H (r=-0,98):**	
		

Si en vez de tomar distancias verticales las hubiésemos tomado horizontales y siguiendo el mismo procedimiento obtendríamos



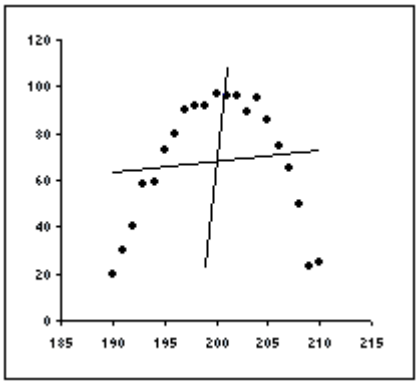
la recta de regresión de X sobre Y:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

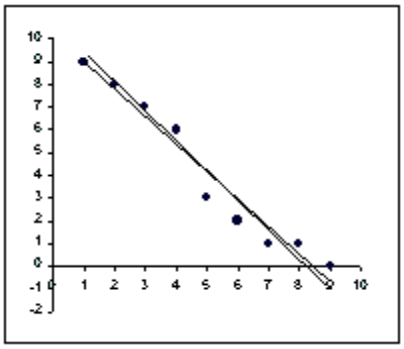
Ambas rectas pasan por el punto $(\bar{x}, \bar{y})$ y formaran un ángulo tanto menor cuanto mayor sea la correlación (es decir, cuanto más próximo esté el coeficiente de correlación a 1 o a –1)

Así pues, las dos rectas serán muy parecidas si la correlación es muy fuerte y formarán un ángulo cercano a los 90° si la correlación es muy débil.

En nuestro ejemplo F (correlación muy baja) las rectas quedarían de la siguiente manera:



En cambio, en el ejemplo H (correlación muy fuerte):



La recta de regresión para hacer estimaciones

Una aplicación de la recta de regresión es la obtención de valores esperados de una variable para ciertos valores de la otra; es decir, si $y = mx + n$ es la recta de regresión de Y sobre X y $x = x_0$ es un valor particular de X, el número $y_0 = mx_o + n$ constituye lo que podríamos llamar "el valor esperado de Y para el valor x_0 de la variable X". Análogamente, la recta de regresión de X sobre Y puede utilizarse para hallar "valores esperados de X para valores determinados de Y".

No obstante lo anterior, dado un valor de una variable, para estimar el valor correspondiente de la otra variable se suelen tener en cuenta las dos rectas de regresión y se da como estimación la media de los dos valores obtenidos.

Debemos tener en cuenta las siguientes consideraciones:

A) Evidentemente la estimación será tanto más fiable cuanto mayor sea la correlación. Si la correlación es débil los valores obtenidos deben ser tomados con muchas reservas.

B) Las estimaciones deben hacerse dentro del intervalo de valores utilizados o muy cerca de ellos. Así, si estudiamos datos sobre pesos y tallas de niños menores de 6 años encontraremos que ambas variables están fuertemente correlacionadas; por lo que la recta de regresión obtenida nos será muy útil. Sin embargo no tiene mucho sentido, a partir de dicha recta, estimar tallas a partir de pesos (o viceversa) de niños de (por ejemplo) 14 años. Si a pesar de todo lo hacemos, las estimaciones no serán fiables.

Ejemplo C: *Las calificaciones de 40 alumnos en Matemáticas y Física han sido:*

X=Calificación Matemáticas	3	3	4	5	5	6	6	6	7	7	8
Y=Calificación Física	2	5	5	4	5	4	6	7	6	7	9
Número de alumnos	4	3	3	2	10	2	4	5	4	2	1

Si un alumno ha obtenido un 6,5 en matemáticas, ¿qué nota se estima que obtendrá en física?

Si ha obtenido un 8 en física, ¿qué nota se espera que obtenga en matemáticas?.

¿son fiables estas previsiones?

Esta información puede disponerse en una tabla de doble entrada:

x_i	y_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
3	2	4	12	36	8	16	24
3	5	3	9	27	15	75	45
4	5	3	12	48	15	75	60
5	4	2	10	50	8	32	40
5	5	10	50	250	50	250	250
6	4	2	12	72	8	32	48
6	6	4	24	144	24	144	144
6	7	5	30	180	35	245	210
7	6	4	28	196	24	144	168
7	7	2	14	98	14	98	98
8	9	1	8	64	9	81	72
		40	209	1165	210	1192	1159

$$\bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{209}{40} = 5,225 \qquad \bar{y} = \frac{\sum y_i \cdot f_i}{N} = \frac{210}{40} = 5,25$$

$$s_x^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{1165}{40} - (5,225)^2 = 1,824 \qquad s_y^2 = \frac{\sum y_i^2 \cdot f_i}{N} - \bar{y}^2 = \frac{1192}{40} - (5,25)^2 = 2,238$$

$$s_x = \sqrt{s_x^2} = \sqrt{1,824} = 1,351 \qquad s_y = \sqrt{s_y^2} = \sqrt{2,238} = 1,496$$

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i}{N} = \frac{\sum x_i \cdot y_i \cdot f_i}{N} - \bar{x} \cdot \bar{y} = \frac{1159}{40} - (5,225) \cdot (5,25) = 1,544$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{1,544}{1,351 \cdot 1,496} = 0,764$$ correlación moderadamente fuerte y positiva

Esto indica que las estimaciones que hagamos no son muy fiables, aún así vamos a realizarlas:

Recta de regresión de Y sobre X, para saber que nota va a sacar en Física conocida la nota de Matemáticas: $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \Leftrightarrow y - 5,25 = \frac{1,544}{1,824}(x - 5,225) \Leftrightarrow y = 0,852x + 0,798$

Para x=6,5 la nota de física sería y=6,3

Recta de regresión de X sobre Y, para saber que nota va a sacar en Matemáticas conocida la nota de Física: $x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \Leftrightarrow x - 5,225 = \frac{1,544}{2,238}(y - 5,25) \Leftrightarrow x = 0,69y + 1,6$

Para y=8 en matemáticas sacaría un x=7,1