Correlación entre variables

Apuntes de clase del curso Seminario Investigativo VI

Por:

Gustavo Ramón S.*

* Doctor en *Nuevas Perspectivas en la Investigación en Ciencias de la Actividad Física y el Deporte (Universidad de Granada).*
Docente – Investigador del Instituto Universitario de Educación Física, Universidad de Antioquia (Colombia).
Correo: gusramon2000@yahoo.es

Correlación entre variables

La Correlación es una técnica estadística usada para determinar la relación entre dos o más variables.

La relación entre la duración de una carrera de distancia y el test del escalón, o la relación entre las características de la personalidad y la participación en deportes de alto riesgo.

La correlación puede ser de al menos dos variables o de una variable dependiente y dos o más variables independientes, denominada correlación múltiple.

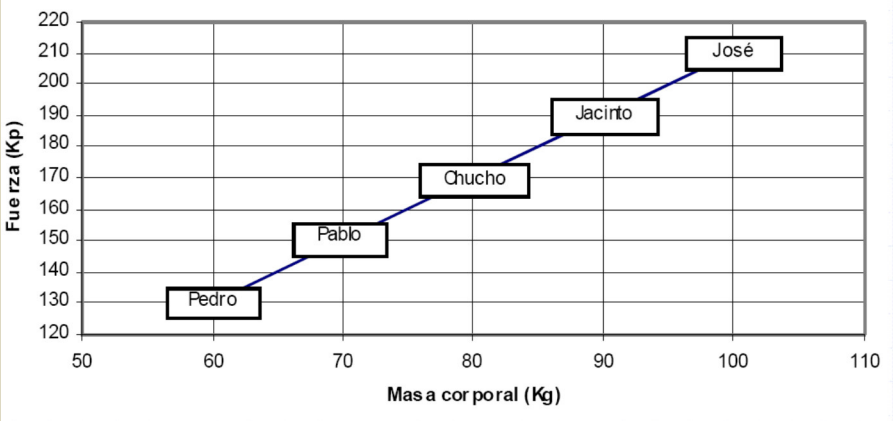
Coeficiente de correlación

El Coeficiente de Correlación es un valor cuantitativo de la relación entre dos o más variables.

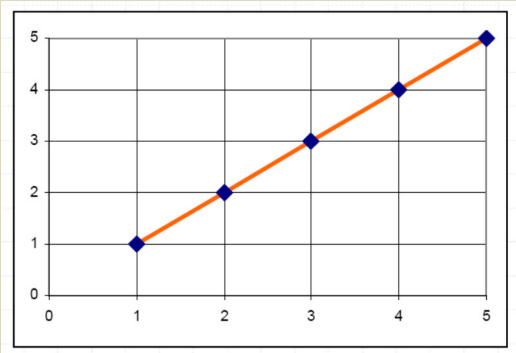
La coeficiente de correlación puede variar desde -1.00 hasta 1.00.

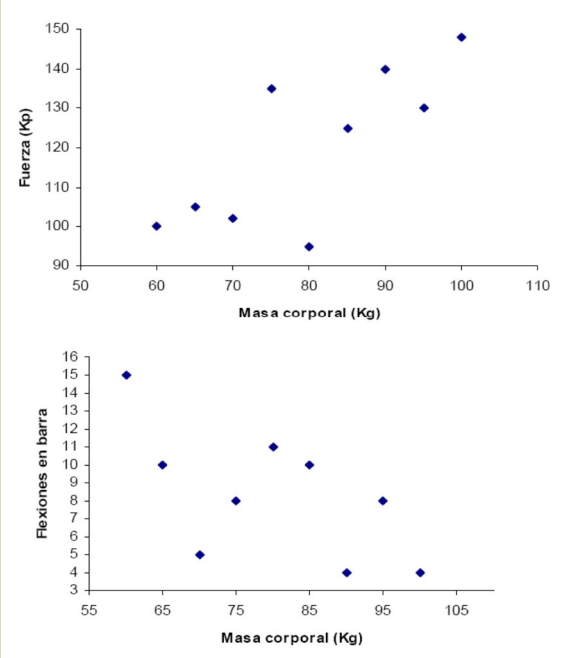
La correlación de proporcionalidad directa o positiva se establece con los valores +1.00 y de proporcionalidad inversa o negativa, con -1.00. No existe relación entre las variables cuando el coeficiente es de 0.00.

Nombre	Masa corporal	Fuerza
Pedro	60	130
Pablo	70	150
Chucho	80	170
Jacinto	90	190
José	100	210



Nombre	Masa corporal	Fuerza
Pedro	1	1
Pablo	2	2
Chucho	3	3
Jacinto	4	4
José	5	5

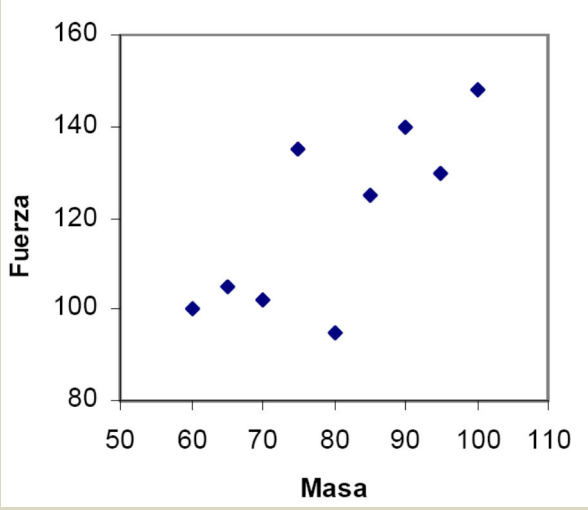




Coeficiente de correlación = r

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2} * \sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

N	Masa		Fuerza		
	X	X2	Y	Y2	XY
1	60	3600	100	10000	6000
2	65	4225	105	11025	6825
3	70	4900	102	10404	7140
4	75	5625	135	18225	10125
5	80	6400	95	9025	7600
6	85	7225	125	15625	10625
7	90	8100	140	19600	12900
8	95	9025	130	16900	12350
9	100	10000	148	21904	14800
Σ	720	59100	1080	13270	88065



N = 9; Σ(XY)= 88.065; Σ(X) 720; = Σ(Y) =1080;

X2 = 59.100; Y2= 132.708;

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2} * \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{9(88.065) - (720*1080)}{\sqrt{9*59.100 - 720^2} * \sqrt{9*132.708 - 1080^2}}$$

$$r = \frac{14.985}{116,19*167,25}$$ r = 0.77

Ecuaciones de Regresión

La fórmula general para una ecuación de regresión lineal es:

Y'= a+bX

donde Y' es el valor predicho

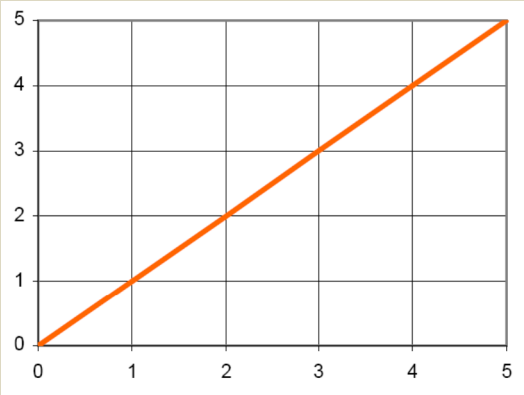
- **a** es el intercepto
- **b** es la pendiente de la línea
- y **X** es el predictor

- **a** puede ser calculada a partir de la siguiente fórmula:

a = My - bMx , donde My es la media de Y, y Mx es la media de X

- **b** puede ser calculada a partir de la siguiente fórmula:

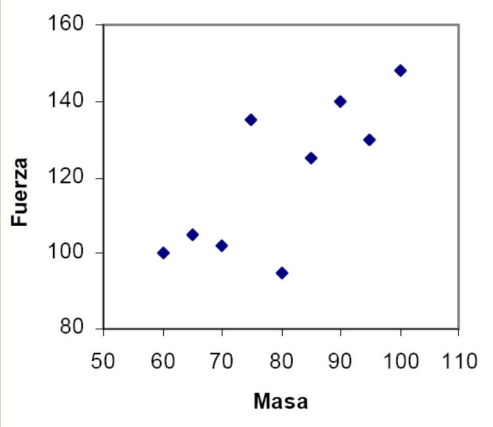
4 b = r (S_y/S_x) , donde S_y es la desviación estándar de Y, y S_x la de X



Intercepto = a = O

Pendiente = b = ΔY / ΔX = (5-0) / (5 – 0) = 1

Si X = 2 → Y = O + 1*2 = 2



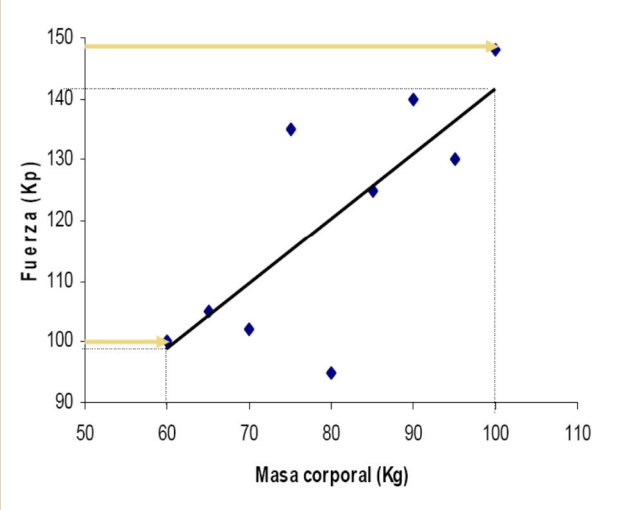
b= r (Sy/Sx)= 0.771 (19.71/13.69)= 1.110

a= My – bMx= 120 - 1.110*80= 31.2

Con esta ecuación de regresión podemos calcular los valores de los extremos para la masa corporal (60 y 100 kg):

Y_{60} = 31.2 + 1.110*60 = 97.8

Y_{100} = 31.2 + 1.110*100 = 142.2



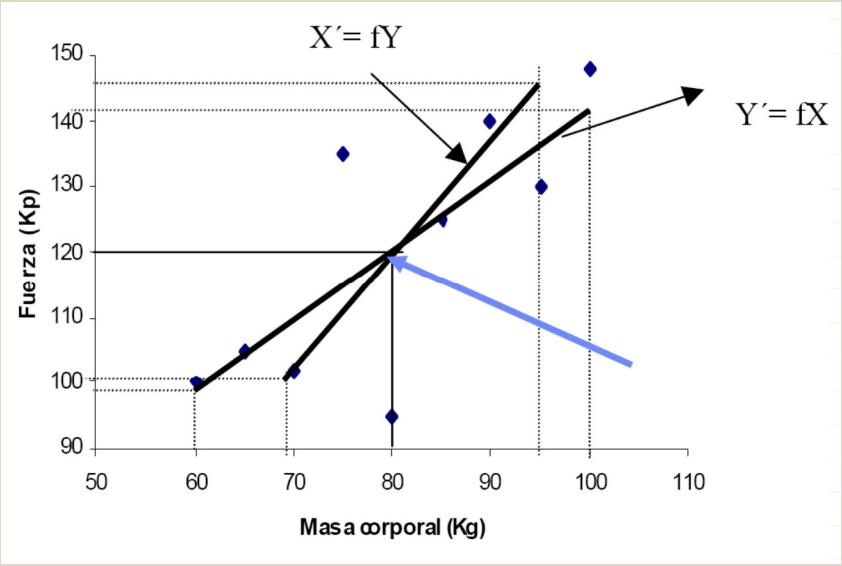
Valores reales para una masa corporal de 60kg era de 100 Kp y en el caso estimado fue de 97.8 (una diferencia de -2.2 kp)

Para el 100 kg, era de 150 y su estimación fue de 142.2 (una diferencia de -7.8kp).

Esto sucede porque la correlación no es de 1.00.

error estándar de la estimación.

En el anterior ejemplo, hicimos la recta de regresión de Y sobre los valores de X. Pero igualmente podríamos calcular y dibujar la línea de regresión de los valores de X sobre Y (X'= 15.71 + O.536Y). El resultado final sería el gráfico siguiente.



Se puede observar que ambas rectas se cortan en el punto correspondiente a la media de X y la media de Y.

Se podría decir que la relación entre las rectas de regresión gira en este punto común. De manera que, cuando r es igual a 1, las rectas se superponen y cuando r es cero, las rectas son perpendiculares.

Se pueden realizar diagramas de dispersión en los que aparece una sola recta de regresión: la que sirve para predecir Y a partir de los valores de X.

Aunque este estudio se refiera a una sola recta, todas las conclusiones serán también aplicables a la recta que sirve para predecir X a partir de Y.

La recta de regresión representa el mejor fundamento para predecir valores de Y a partir de valores conocidos de X.

No todos los puntos que representan las calificaciones caen sobre la recta de regresión. Las desviaciones de los valores reales menos los valores predichos representan los errores de la predicción.

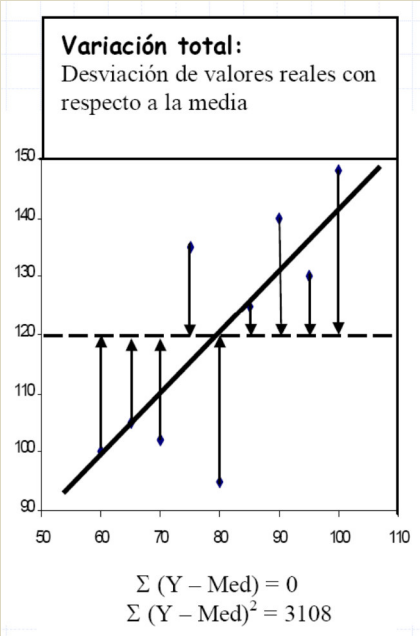
N	Masa X	Fuerza Y	Y´	(Y-Med)2	(Y´-Med)2	(Y - Y´)2
1	60	100	97.8	400	492.8	4.8
2	65	105	103.35	225	277.2	2.7
3	70	102	108.9	324	123.2	47.6
4	75	135	114.45	225	30.8	422.3
5	80	95	120	625	0.0	625.0
6	85	125	125.55	25	30.8	0.3
7	90	140	131.1	400	123.2	79.2
8	95	130	136.65	100	277.2	44.2
9	100	148	142.2	784	492.8	33.6
Media		120	Suma	3108.0	1848.15	1259.85

$$\sum (Y-\overline{Y})^2 = \sum (Y-Y')^2 + \sum (Y'-\overline{Y})^2$$

Variación total — Variación no explicada — Variación explicada

Las tres sumas de cuadrados son:

1. Variaciones de los valores con respecto a la media de la muestra. Esta variación está dada por (Y - Media)² y es básica para la determinación de la varianza y de la desviación estándar de la muestra. Es la variación total.

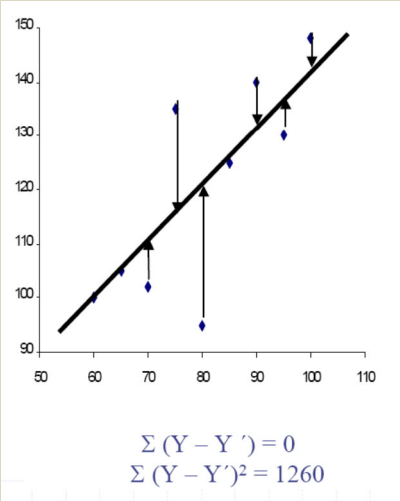


2. Variación de los valores reales con respecto a la recta de regresión (o valores predichos) Esta variación está dada por (Y - Y′) ² y se llama variación no explicada.

Si la correlación fuese de ±1.00, todos los valores caen en la recta de regresión y en consecuencia, se explicarían toda la variación de los valores de Y en función de la variación en X. Cuando existe una correlación perfecta, no existe variación no explicada.

• Cuando la correlación no es perfecta, muchos de los puntos no caen en la recta de regresión. Las desviaciones de estos valores con respecto a la recta de regresión representan las variaciones que no pueden ser explicadas mediante la correlación entre ambas variables, de ahí el uso del término.

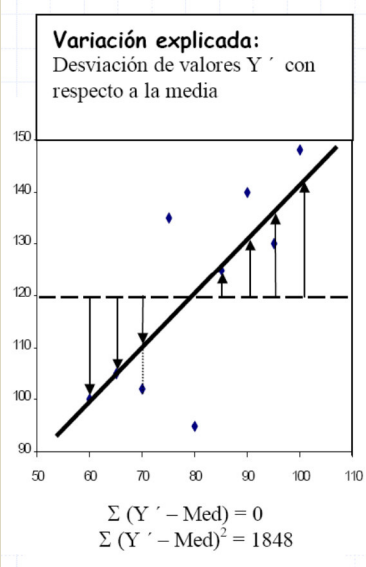
Variación no explicada: Desviación de valores estimados menos los reales



3. Variación de los valores estimados respecto a la media de la distribución. Esta variación está dada por (Y'- Media)² y se la conoce como variación explicada.

Este nombre deriva, de manera análoga, a la dada para la variación anterior.

Variación explicada: Desviación de valores Y' con respecto a la prima.



Coeficiente de determinación

$$r^2 = \frac{Variacion \quad \exp licada}{Variacion \quad total} = \frac{\sum (Y' - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

r = raíz (r^2)

Puesto que r^2 representa la proporción de la variación explicada, (1- r^2) representará la proporción de la variación que no es explicada, conocido como coeficiente de no determinación y se representa por k^2.

La relación entre r^2 y k^2 es $k^2+r^2 = 1$