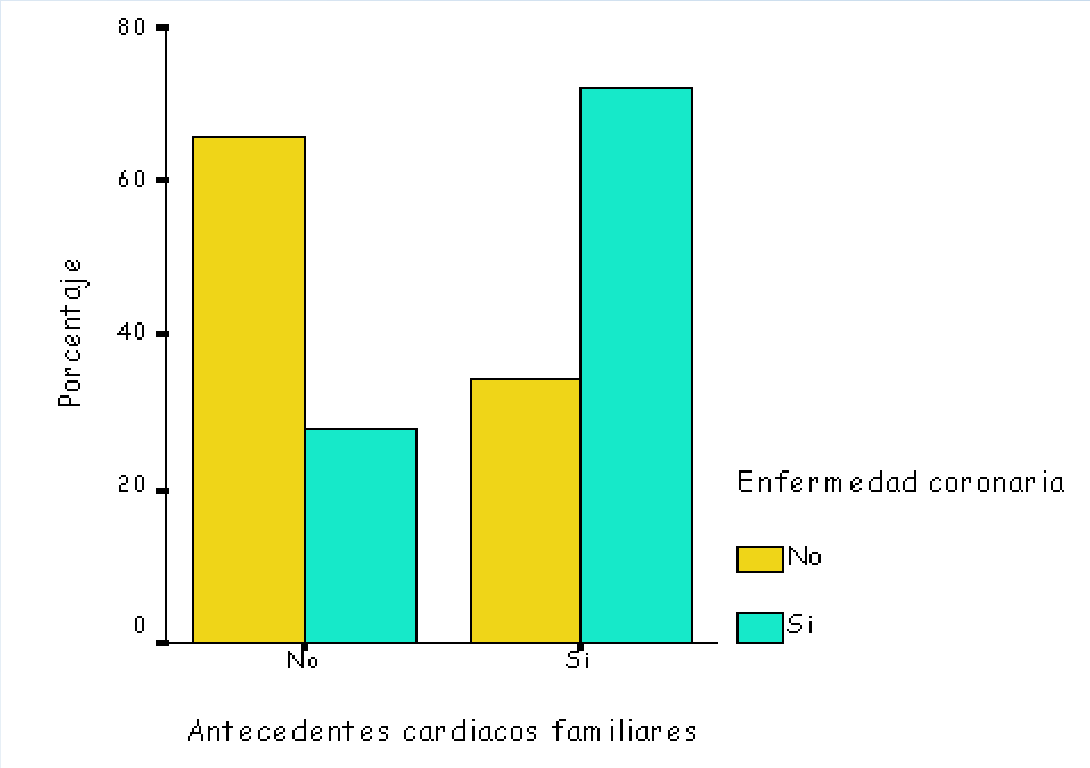
TEMA IV
ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

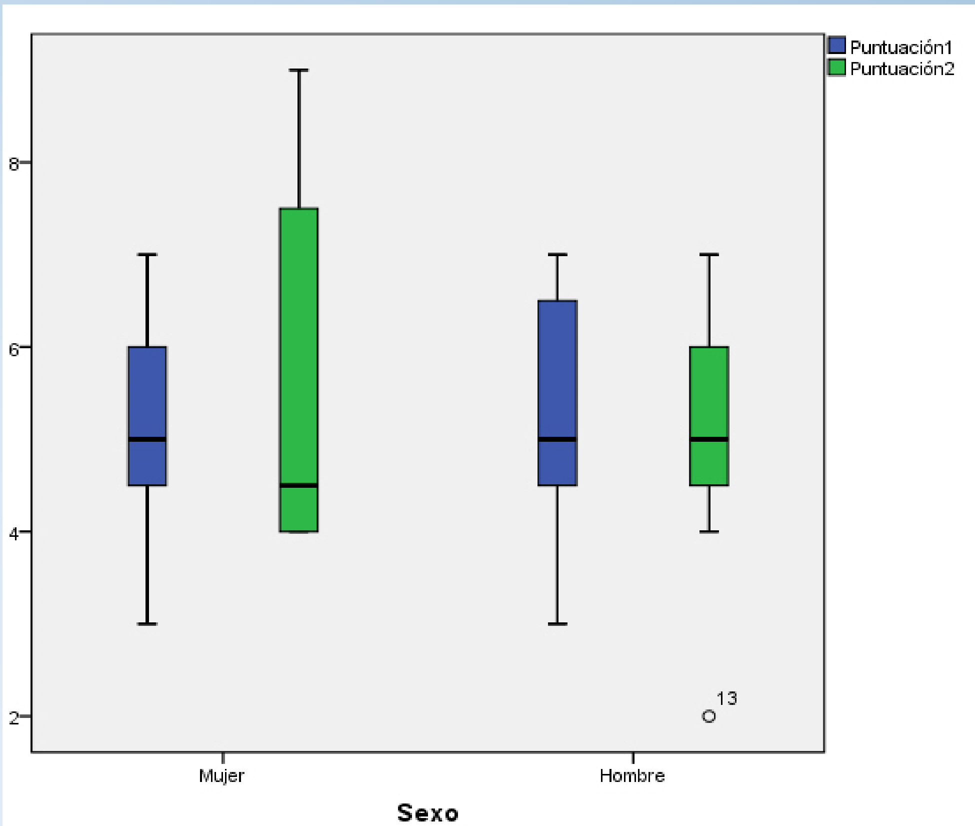
1. Introducción.
2. Tablas de frecuencias bivariadas y gráficos.
3. Variables cuantitativas.
    3.1. Covarianza.
    3.2. Coeficiente de correlación de Pearson.

1.Introducción

Hasta ahora nos hemos centrado en medidas de tendencia central, variabilidad, asimetría y curtosis de una única variable.

No obstante, en la práctica es común examinar dos o más variables conjuntamente (v.g., relación entre inteligencia y rendimiento, etc.)





2. Tablas de Frecuencias Bivariadas y Gráficos

$X \setminus Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1k}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_h	n_{h1}	n_{h2}	$\dots$	n_{hj}	$\dots$	n_{hk}	$n_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot k}$	N

Distribuciones marginales

2. Tablas de Frecuencias Bivariadas y Gráficos

Tabla de Correlación o Contingencia

Permite ayudarnos a determinar si existe relación de **interdependencia** entre 2 variables, es decir, si se influyen mutuamente.

	B_1	B_2	...	B_J
A_1	n_{11}	n_{12}	...	n_{1J}
A_2	n_{21}	n_{22}	...	n_{2J}
...	...	...	...	...
A_I	n_{I1}	n_{I2}	...	n_{IJ}

donde **nij** es el número de observaciones que presentan **simultáneamente** las características i, j de las variables A y B, respectivamente.

Así, una tabla de contingencia es una una tabla de doble entrada, donde en cada casilla figurará el número de casos o individuos que poseen un nivel de una de las características analizadas y otro nivel de la otra característica.

Distribuciones marginales

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de como se comporta la otra. Estaríamos así en el análisis de una **distribución marginal**.

Distribución marginal de A

A_i	$n_{i.}$
A_1	$n_{1.}$
A_2	$n_{2.}$
…	…
A_{n-1}	$n_{n-1.}$
A_n	$n_{n.}$

Distribución marginal de B

B_j	$n_{.j}$
B_1	$n_{.1}$
B_2	$n_{.2}$
…	…
B_{m-1}	$n_{.m-1}$
B_m	$n_{.m}$

En las tablas de contingencia:

a) **Distribuciones marginales**

	B_1	B_2	...	B_J	Distribución marginal de A
A_1	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...
A_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I.}$
Distribución marginal de B	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	n

b) **Distribuciones de frecuencias relativas**

	B_1	B_2	...	B_J	Distribución marginal de A
A_1	$\frac{n_{11}}{n}$	$\frac{n_{12}}{n}$	...	$\frac{n_{1J}}{n}$	$\frac{n_{1.}}{n}$
A_2	$\frac{n_{21}}{n}$	$\frac{n_{22}}{n}$	...	$\frac{n_{2J}}{n}$	$\frac{n_{2.}}{n}$
...	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
A_I	$\frac{n_{I1}}{n}$	$\frac{n_{I2}}{n}$	...	$\frac{n_{IJ}}{n}$	$\frac{n_{I.}}{n}$
Distribución marginal de B	$\frac{n_{.1}}{n}$	$\frac{n_{.2}}{n}$	...	$\frac{n_{.J}}{n}$	1

c) **Perfiles fila**

	B_1	B_2	...	B_J	
A_1	$\frac{n_{11}}{n_{1.}}$	$\frac{n_{12}}{n_{1.}}$		$\frac{n_{1J}}{n_{1.}}$	1
A_2	$\frac{n_{21}}{n_{2.}}$	$\frac{n_{22}}{n_{2.}}$	...	$\frac{n_{21}}{n_{2.}}$	1
...	⋮	⋮	...	⋮	⋮
A_I	$\frac{n_{I1}}{n_{I.}}$	$\frac{n_{I2}}{n_{I.}}$	...	$\frac{n_{IJ}}{n_{I.}}$	1
Distribución marginal de B	$\frac{n_{.1}}{n}$	$\frac{n_{.2}}{n}$	...	$\frac{n_{.J}}{n}$	1

Del total de individuos con la característica "A_1" que porcentaje comparte a su vez la "B_1"

d) **Perfiles columna**

	B_1	B_2	...	B_J	*Distribución marginal de A*
A_1	$\frac{n_{11}}{n_{.1}}$	$\frac{n_{12}}{n_{.2}}$	...	$\frac{n_{1J}}{n_{.J}}$	$\frac{n_{1.}}{n}$
A_2	$\frac{n_{21}}{n_{.1}}$	$\frac{n_{22}}{n_{.2}}$	...	$\frac{n_{2J}}{n_{.J}}$	$\frac{n_{2.}}{n}$
...	⋮	⋮	...	⋮	⋮
A_I	$\frac{n_{I1}}{n_{.1}}$	$\frac{n_{I2}}{n_{.2}}$	...	$\frac{n_{IJ}}{n_{.J}}$	$\frac{n_{I.}}{n}$
	1	1	...	1	1

Cómo es lógico, el porcentaje de individuos con "A_1" que, o bien comparten B_1 o B_2 y hasta B_j será el 100% = 1

Independencia Estadística

$$\frac{n_{ij}}{N}$$

		Graves Y				
	Averias	0	1	2	3	Marginal de leves
Leves X	0	0,2308	0,0385	0,0077	0,0000	0,2769
	[illegible]	0,1692	0,0615	0,0231	0,0077	0,26[illegible]
	2	0,07[illegible]	0,0385	0,0154	0,0154	0,1462
	3	0,0923	0,0615	0,0077	0,0154	0,1769
	4	0,0615	0,0308	0,0000	0,0077	0,1000
	5	0,0308	0,0077	0,0000	0,0000	0,0385
Marginal de Graves		0,6615	0,2385	0,0538	0,0462	1

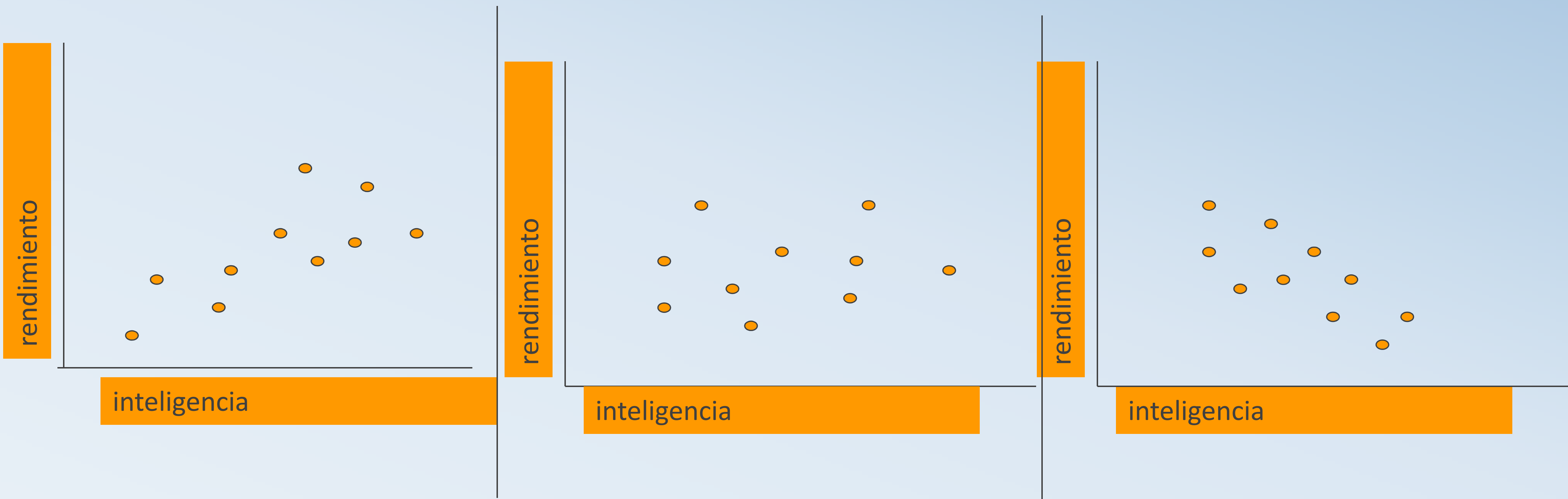
$$\frac{n_{i.}}{N}$$

$$\frac{n_{.j}}{N}$$

$$Si \quad \frac{n_{i.}}{N}\frac{n_{.j}}{N} = \frac{n_{ij}}{N} \quad \forall\, ij \Rightarrow Independencia$$

Representación gráfica: Nube de puntos o diagrama de dispersión

Representación gráfica de una relación



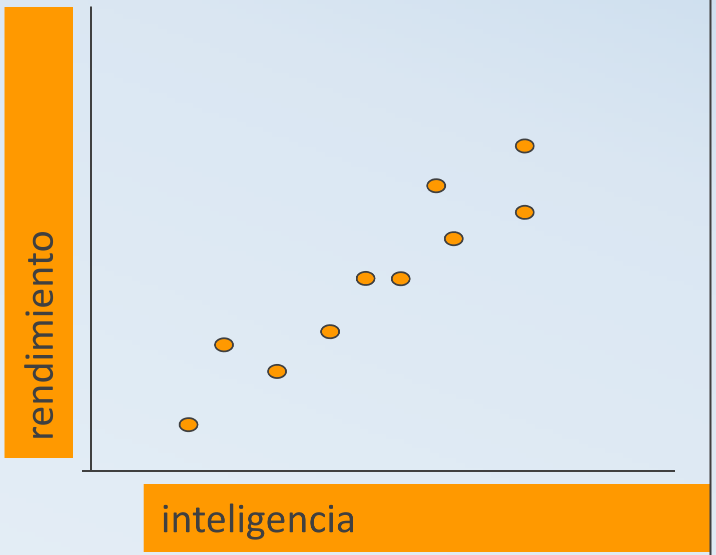
Relación lineal positiva

Sin relación

Relación lineal negativa

Nota: El coeficiente de correlación de Pearson mide relación LINEAL.

Representación gráfica de una relación (2)

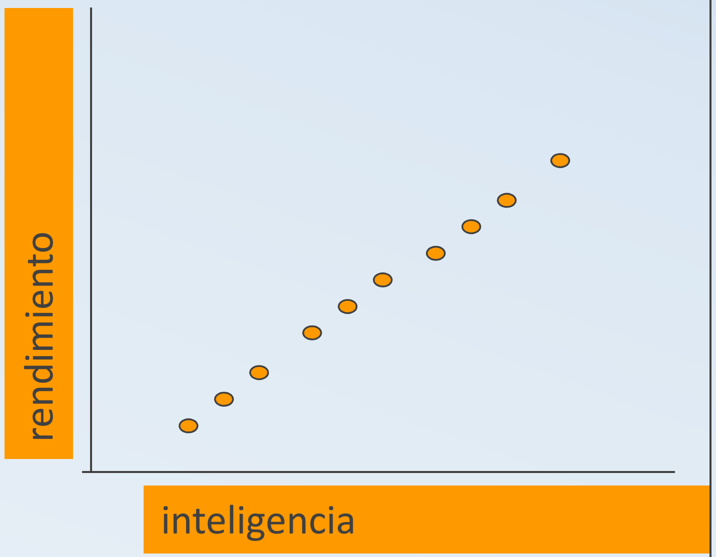


Relación lineal

Relación no lineal

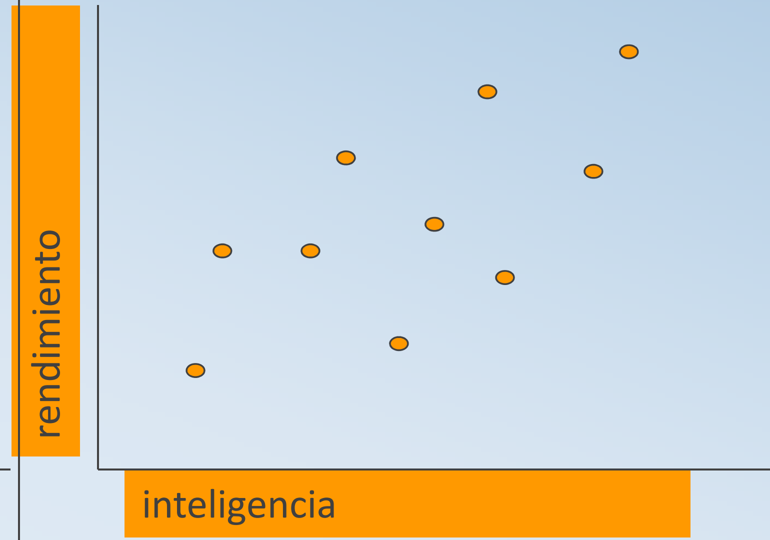
Nota: El coeficiente de correlación de Pearson mide relación LINEAL.

Representación gráfica de una relación (3)



Relación lineal perfecta (casi perfecta)

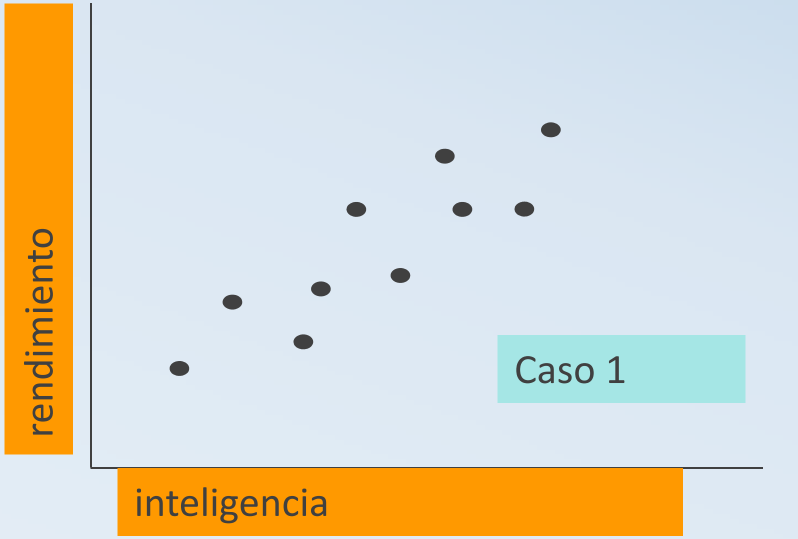
Relación lineal fuerte/moderada



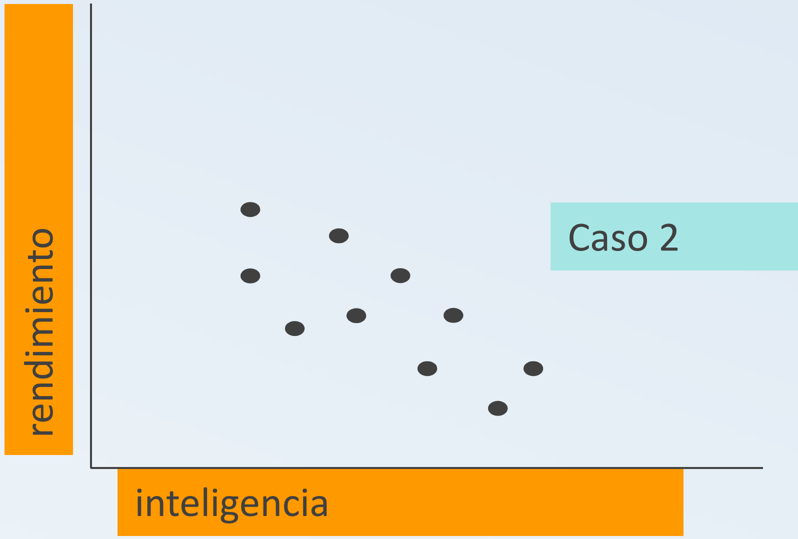
Relación lineal débil

Ahora necesitamos un índice que nos informe tanto del grado en que X e Y están relacionadas, y si la relación es positiva o negativa

3. Covarianza e índice de correlación de Pearson



Observad que cuando la relación lineal es positiva, cuando las puntuaciones diferenciales de X son positivas, las puntuaciones diferenciales de Y suelen ser positivas.



Observad que cuando la relación lineal es negativa, cuando las puntuaciones diferenciales de X son positivas, las puntuaciones diferenciales de Y suelen ser negativas.

3.1 Covarianza

La covarianza aprovecha esta característica señalada en la transparencia anterior (al emplear el producto de las puntuaciones diferencias de X e Y). He aquí la fórmula:

$$s_{xy} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{n}$$

En el caso 1, la covarianza será un valor positivo, y en el caso 2, la covarianza será un valor negativo. Por tanto la covarianza nos da una idea de si la relación entre X e Y es positiva o negativa.

Problema: la covarianza no en un índice acotado (v.g., cómo interpretar una covarianza de 6 en términos del grado de asociación), y no tiene en cuenta la variabilidad de las variables. Por eso se emplea el siguiente índice....

3.2 Coeficiente de correlación (lineal) de Pearson

El coeficiente de correlación de Pearson parte de la covarianza:

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{n \cdot s_x \cdot s_y}$$

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

Ahora veremos varias propiedades del índice...

3.2 Coeficiente de correlación (lineal) de Pearson

Propiedad 1. El índice de correlación de Pearson no puede valer menos de -1 ni más de +1.

Un índice de correlación de Pearson de -1 indica una relación lineal negativa perfecta

Un índice de correlación de Pearson de +1 indica una relación lineal positiva perfecta.

Un índice de correlación de Pearson de 0 indica ausencia de relación lineal. (Observad que un valor cercano a 0 del índice no implica que no haya algún tipo de relación no lineal: el índice de Pearson mide **relación lineal**.)

3.2 Coeficiente de correlación (lineal) de Pearson

Propiedad 2. El índice de correlación de Pearson (en valor absoluto) no varía cuando se transforman linealmente las variables.

Por ejemplo, la correlación de Pearson entre la temperatura (en grados celsius) y el nivel de depresión es la misma que la correlación entre la temperatura (medida en grados Fahrenheit) y el nivel de depresión.

Evidentemente, el índice de correlación de Pearson es el mismo entre las puntaciones directas de X e Y, o entre las puntuaciones diferenciales de X e Y, o entre las puntuaciones típicas de X e Y. (Recordad que las puntuaciones diferenciales y las puntuaciones típicas son transformaciones lineales de las puntuaciones directas.)

3.2 Coeficiente de correlación (lineal) de Pearson

Interpretación

Hemos de tener en cuenta qué es lo que estamos midiendo para poder interpretar cuán grande es la relación entre las variables bajo estudio. En muchos casos, depende del área bajo estudio.

En todo caso, es muy importante efectuar el diagrama de dispersión. Por ejemplo, en el caso de la izquierda, es claro que no hay relación entre inteligencia y rendimiento. Sin embargo, si calculamos el índice de correlación de Pearson nos dará un valor muy elevado, causado por la puntuación atípica en la esquina superior derecha.

3.2 Coeficiente de correlación (lineal) de Pearson

Interpretación (2)

Es importante indicar que “CORRELACIÓN NO IMPLICA CAUSALIDAD”. El que dos variables estén altamente correlaciones no implica que X causa Y ni que Y causa X.

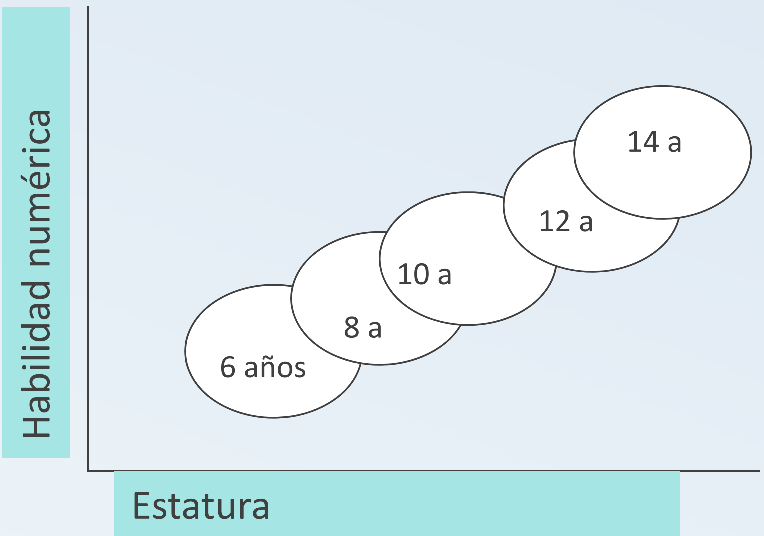
(Esa es una de las razones empleadas por las tabaqueras en el tema de la correlación entre cáncer de pulmón y el hecho de fumar.)

3.2 Coeficiente de correlación (lineal) de Pearson

Interpretación (3)

Es importante indicar que el coeficiente de correlación de Pearson puede verse afectado por la influencia de terceras variables.

Por ejemplo, si fuéramos a un colegio y medimos la estatura y pasamos una prueba de habilidad verbal, saldrá que los más altos también tienen más habilidad verbal...claro, que eso puede ser debido simplemente a que en el colegio los niños más altos serán mayores en edad que los más bajos.



Si se parcializa esta "tercera" variable (mediante "correlación parcial"), difícilmente habrá una relación de importancia entre estatura y habilidad numérica.

Hay muchos casos en que es la tercera variable la causante de una alta relación entre X e Y (y ello muchas veces es difícil de identificar)

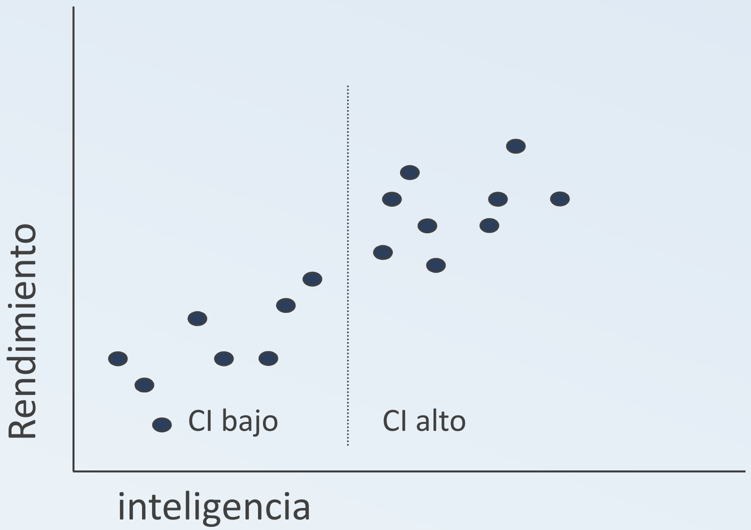
3.2 Coeficiente de correlación (lineal) de Pearson

Interpretación (4)

Por otra parte, el valor del coeficiente de Pearson depende en parte de la variabilidad del grupo.

Si efectuamos el coeficiente de Pearson entre inteligencia y rendimiento con todos los sujetos, el valor del coeficiente de Pearson será bastante elevado.

Sin embargo, si empleamos únicamente los individuos con CI bajo (o CI alto) y calculamos la correlación con Rendimiendo, el valor del coeficiente de Pearson será claramente menor.



Un grupo heterogéneo daría pues un mayor grado de relación entre variables que un grupo homogéneo.