Estadística Descriptiva II: Relación entre variables

Iniciación a la Investigación Ciencias de la Salud

Jesús Montanero Fernández

MUI Ciencias de la Salud, UEx

25 de octubre de 2010

¿De qué trata?

Descripción conjunto concreto de datos (sin generalizar)

- Clasificación $\rightsquigarrow$ Tablas de frecuencia
- Representación $\rightsquigarrow$ Gráficos
- Resumen $\rightsquigarrow$ Valores típicos

Tipos de variables ¿?

Según SPSS

- Cualitativas (factores)
  - Nominales: Grupo sanguíneo
  - Ordinales: Grado enfermedad
- Cuantitativas o de escala : Temperatura, estatura, glucemia, n^o hijos...

Problemas estadísticos: **relación**

Tipos de relaciones

- Cualitativa (factor) $\leftrightarrow$ Cuantitativa (Comparación de grupos o tratamientos)
- Cuantitativa $\leftrightarrow$ Cuantitativa (Regresión)
- Cualitativa (factor) $\leftrightarrow$ Cuanlitativa (Tablas de contingencia)

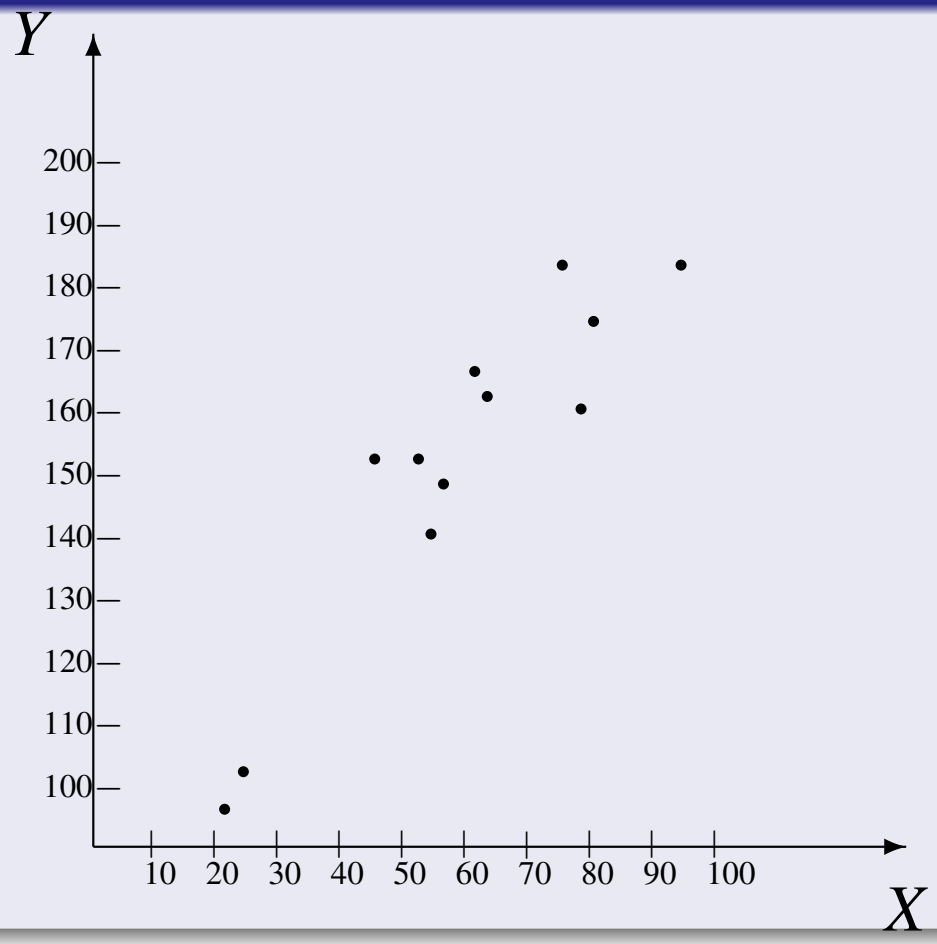
Relación entre dos variables cuantitativas

Peso-altura

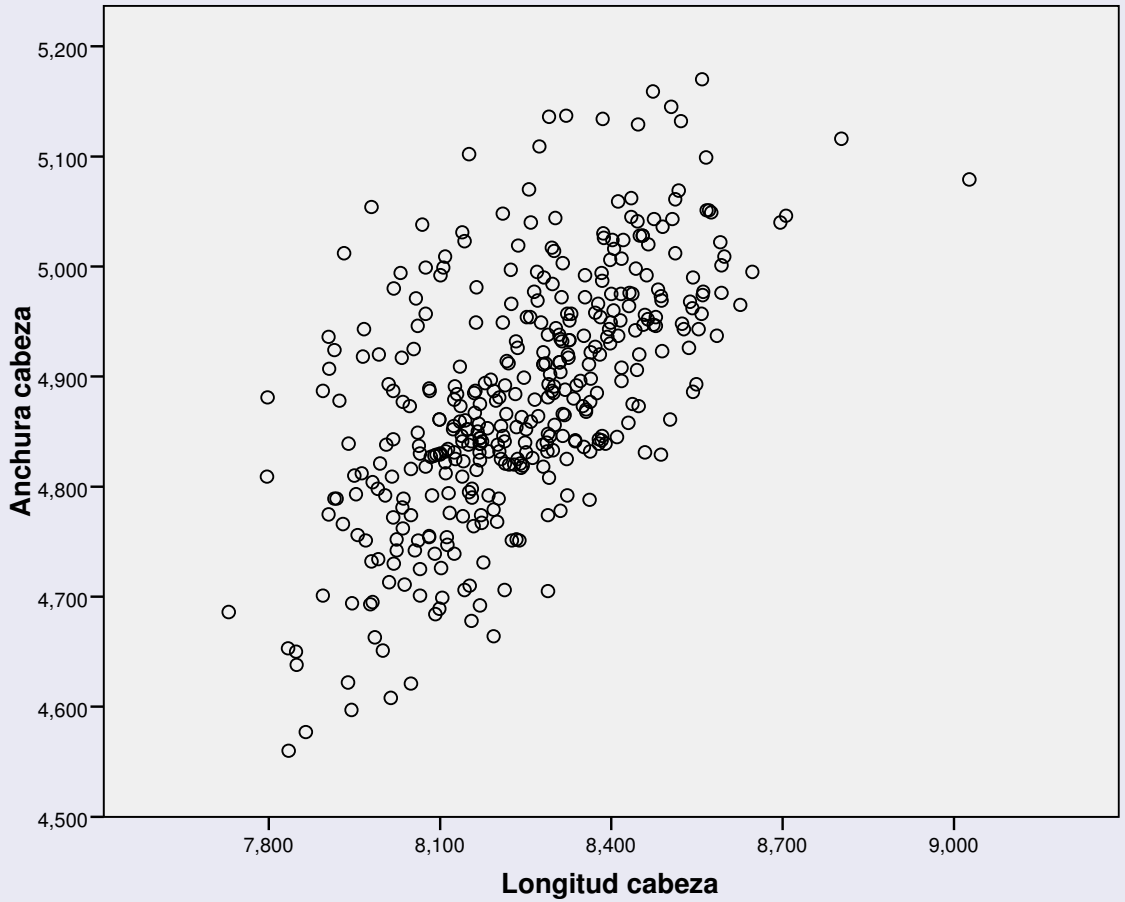
X =peso(kg)	80	45	63	94	24	75	56	...
Y =altura(cm)	174	152	160	183	102	183	148	...

Gráfico

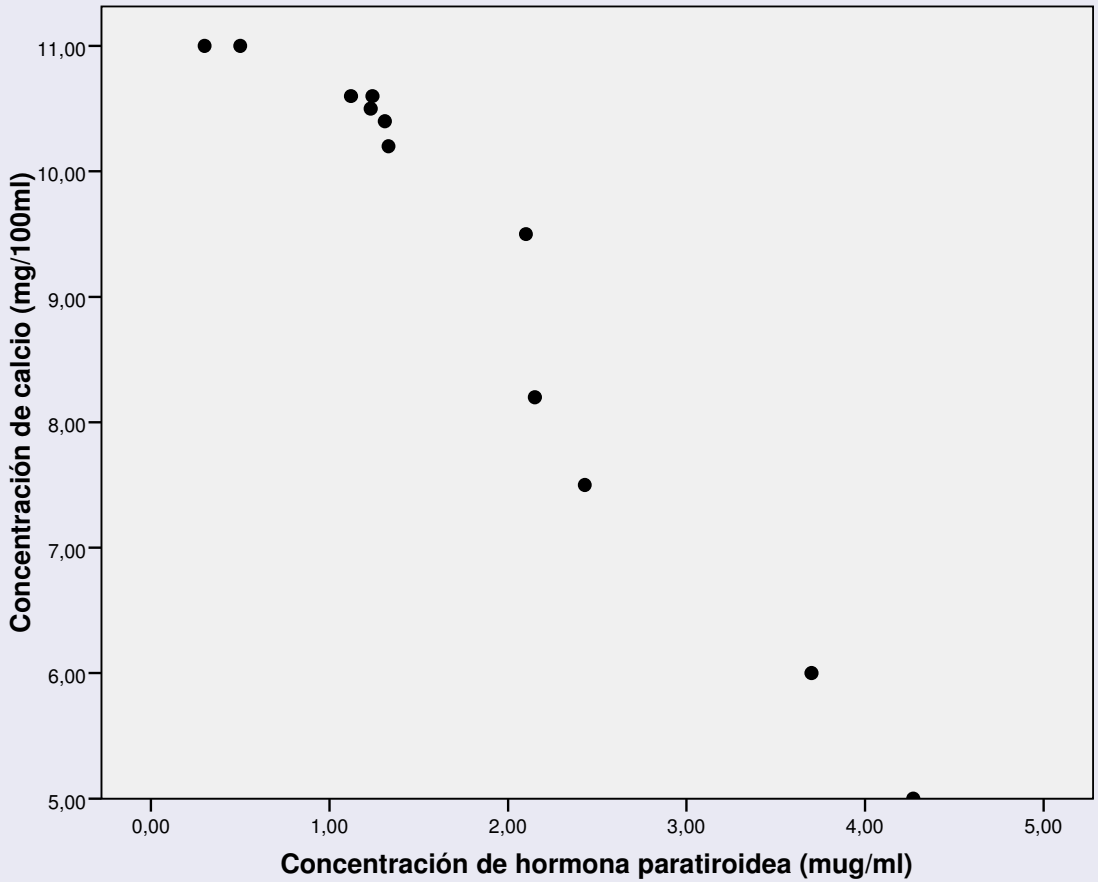
Diagrama de dispersión



Otro ejemplo



Estudiamos inicialmente relaciones lineales



Valores típicos

Dos tipos

- De las variables por separados.
- **Referentes a la relación entre las variables**

Variables por separado

$$\bar{x}, \quad s_x, \quad \bar{y}, \quad s_y, \quad \tilde{y}, \ldots$$

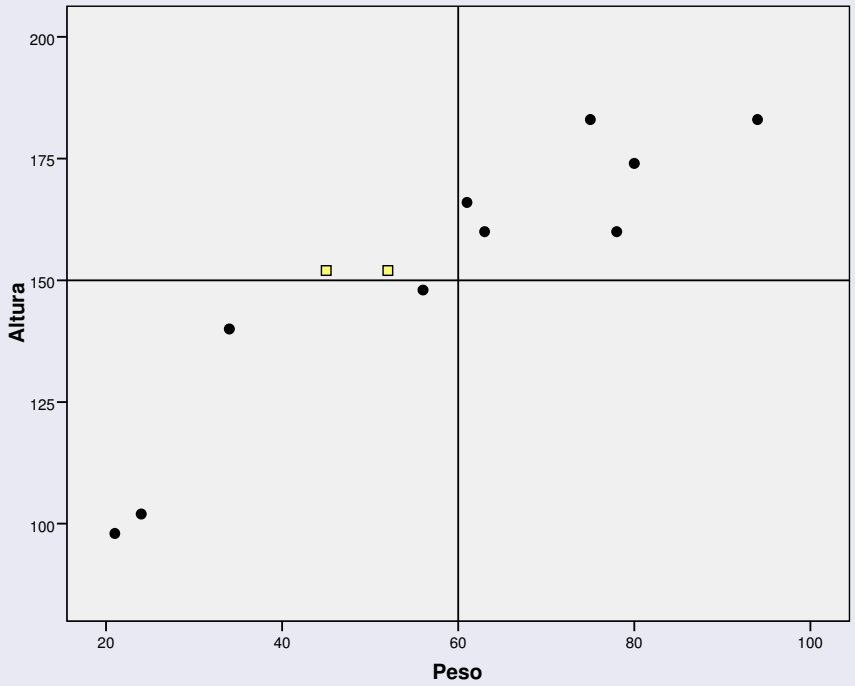
Referentes a la relación entre las variables: Covarianza $\rightarrow$ r

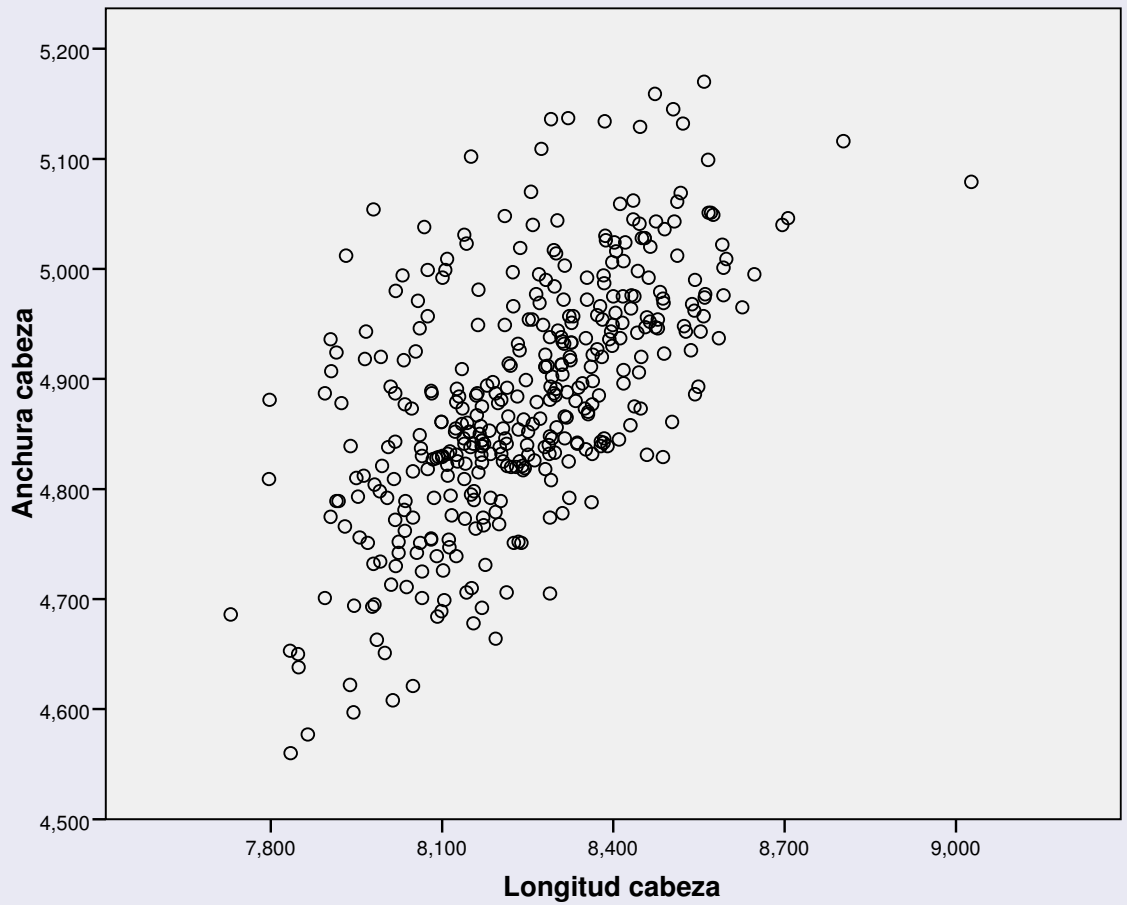
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$-s_x \cdot s_y \leq s_{xy} \leq +s_x \cdot s_y \, .$$

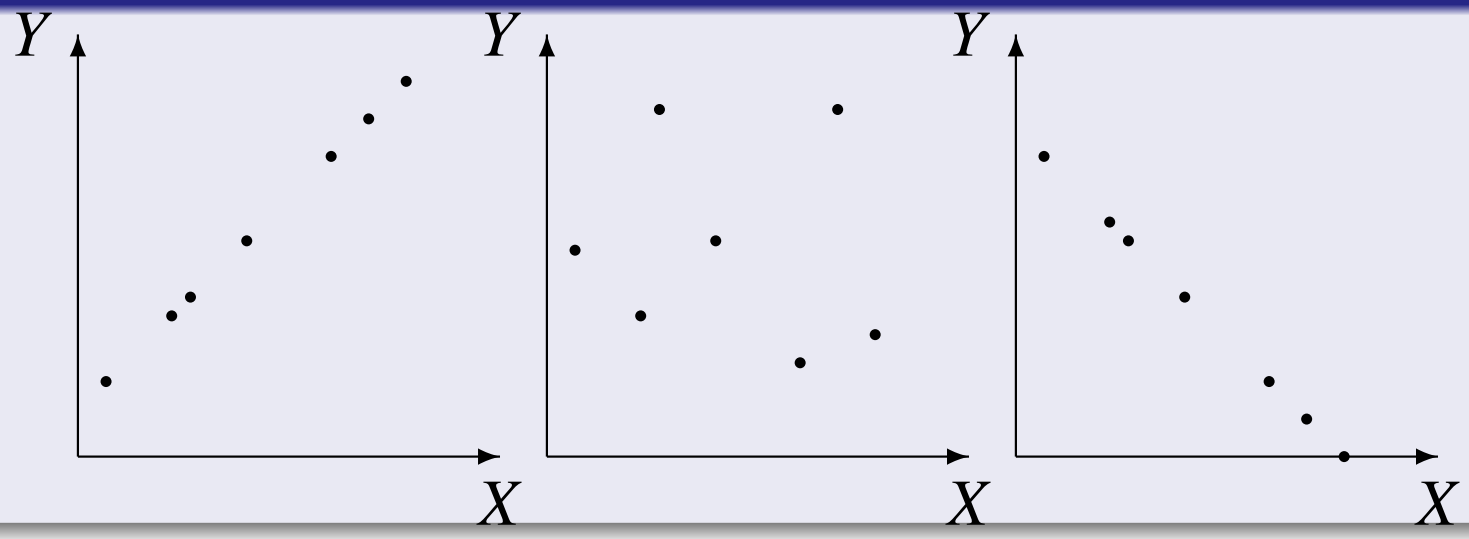
Interpretación gráfica

$$-630{,}71 \leq s_{xy} \leq +630{,}71 \qquad s_{xy} = 577{,}86$$

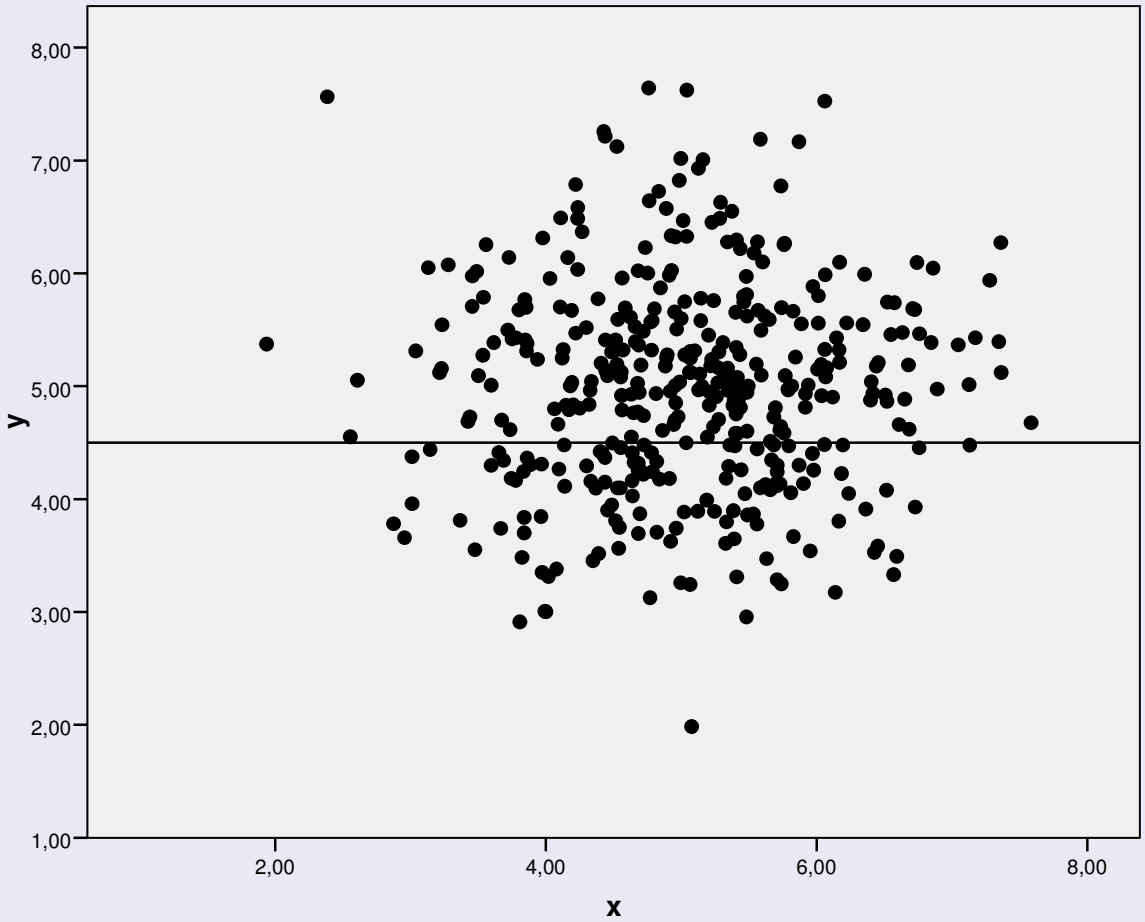



5,200
5,100
5,000
4,900
4,800
4,700
4,600
4,500
Anchura cabeza
7,800
8,100
8,400
8,700
9,000
Longitud cabeza

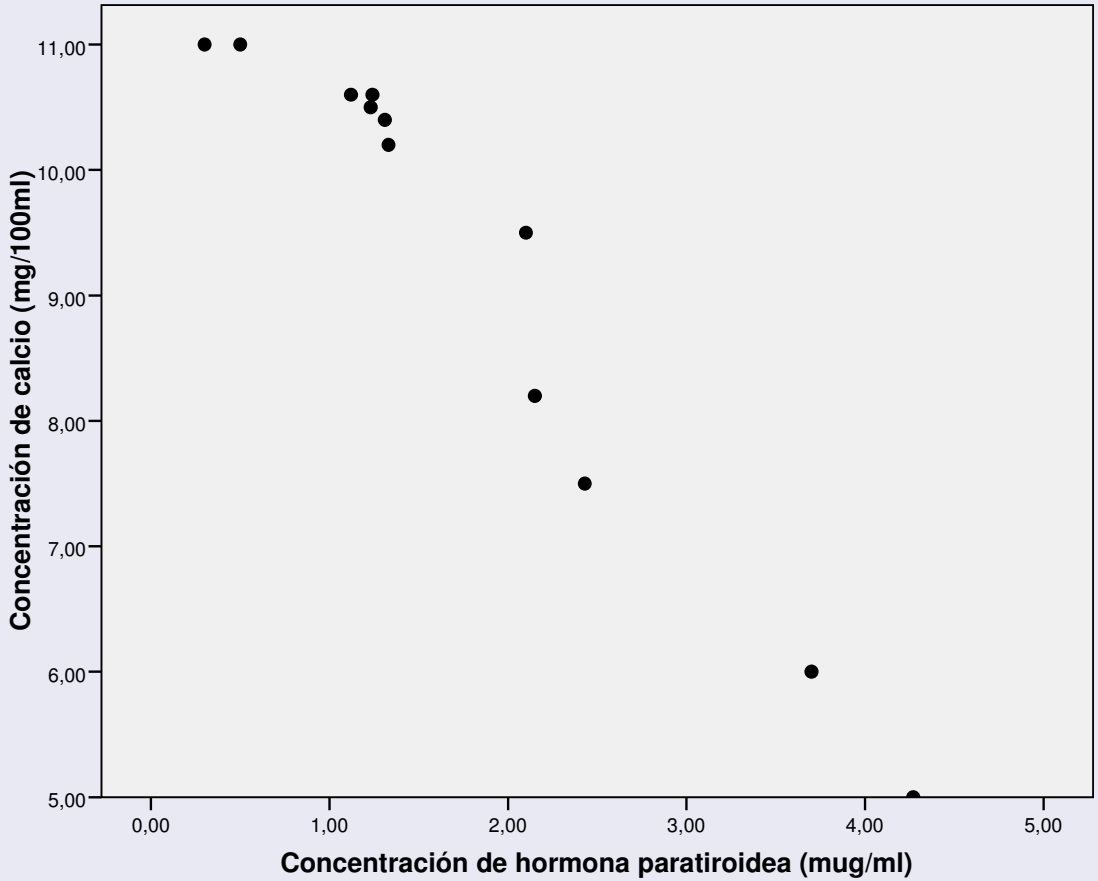
Interpretación gráfica covarianza



Covarianza próxima a cero



Covarianza negativa



Coeficiente de correlación lineal r

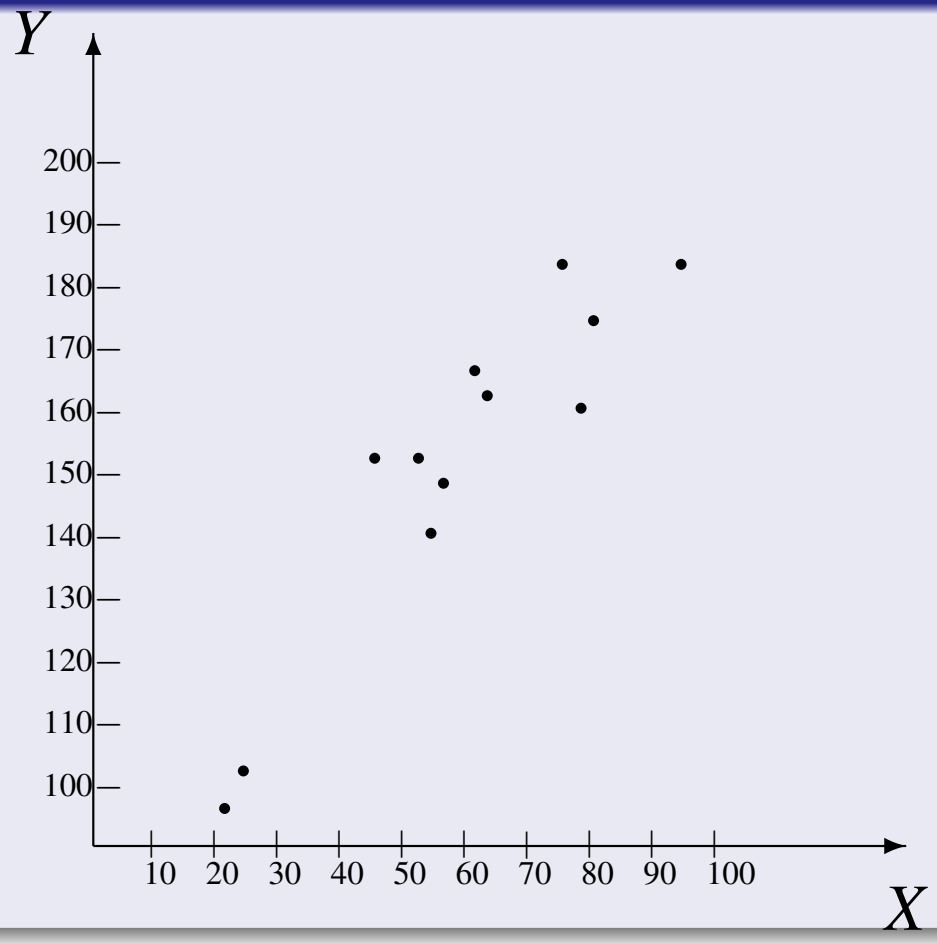
Medida adimensional del grado de correlación

$$-s_x \cdot s_y \leq s_{xy} \leq +s_x \cdot s_y \, .$$

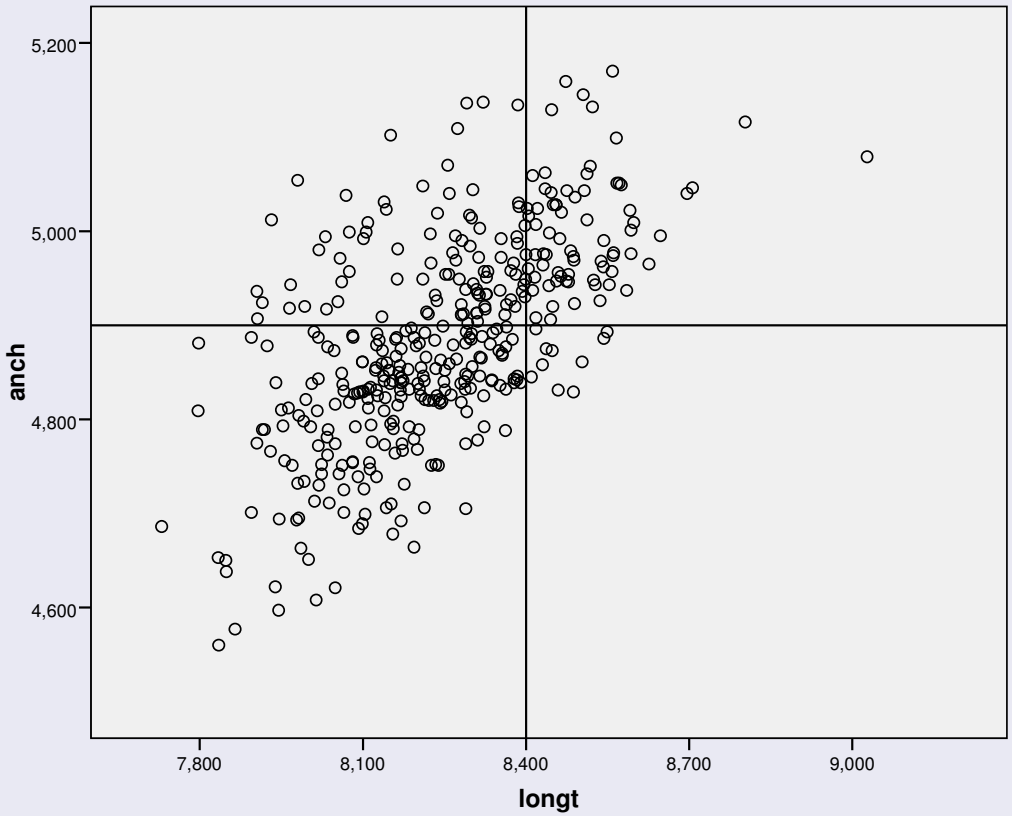
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$-1 \leq r \leq 1$$

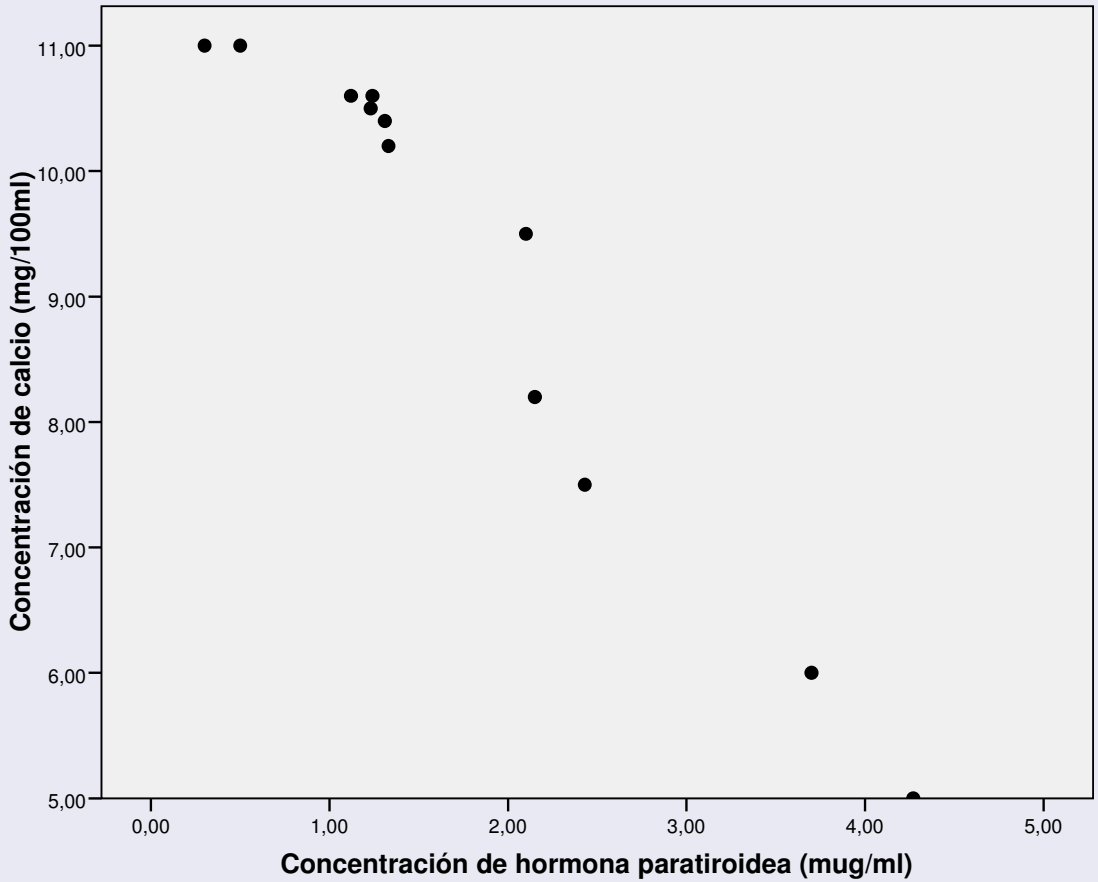
$r = 0{,}91$


Y
200
190
180
170
160
150
140
130
120
110
100
10
20
30
40
50
60
70
80
90
100
X

$r = 0{,}625$



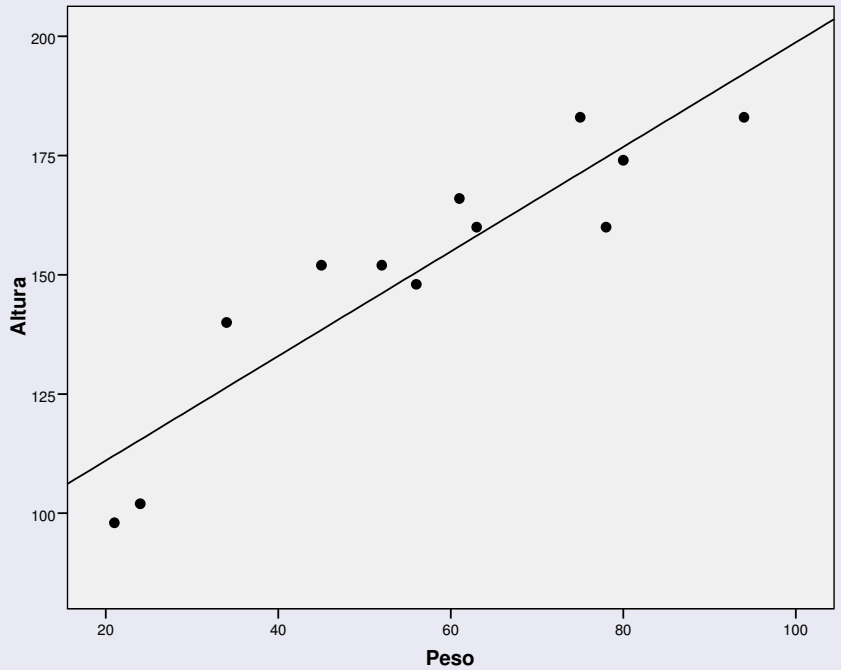
$r = -0{,}97$


Concentración de calcio (mg/100ml)
11,00
10,00
9,00
8,00
7,00
6,00
5,00
0,00
1,00
2,00
3,00
4,00
5,00
Concentración de hormona paratiroidea (mug/ml)

Recta de regresión lineal

$y = a + b \cdot x \quad y = 89{,}11 + 1{,}10x$

Predicciones: $x = 62kg \rightarrow \hat{y} = 89{,}11 + 1{,}10 \cdot 60 = 155{,}11cm$



Regresión múltiple

¿Pueden introducirse más variables explicativas en la ecuación?

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Varianza residual

Mide el error cometido por la recta de regresión

$$s^2_{y\leftarrow x} = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (a+bx_i)]^2 = 1335{,}32/10$$

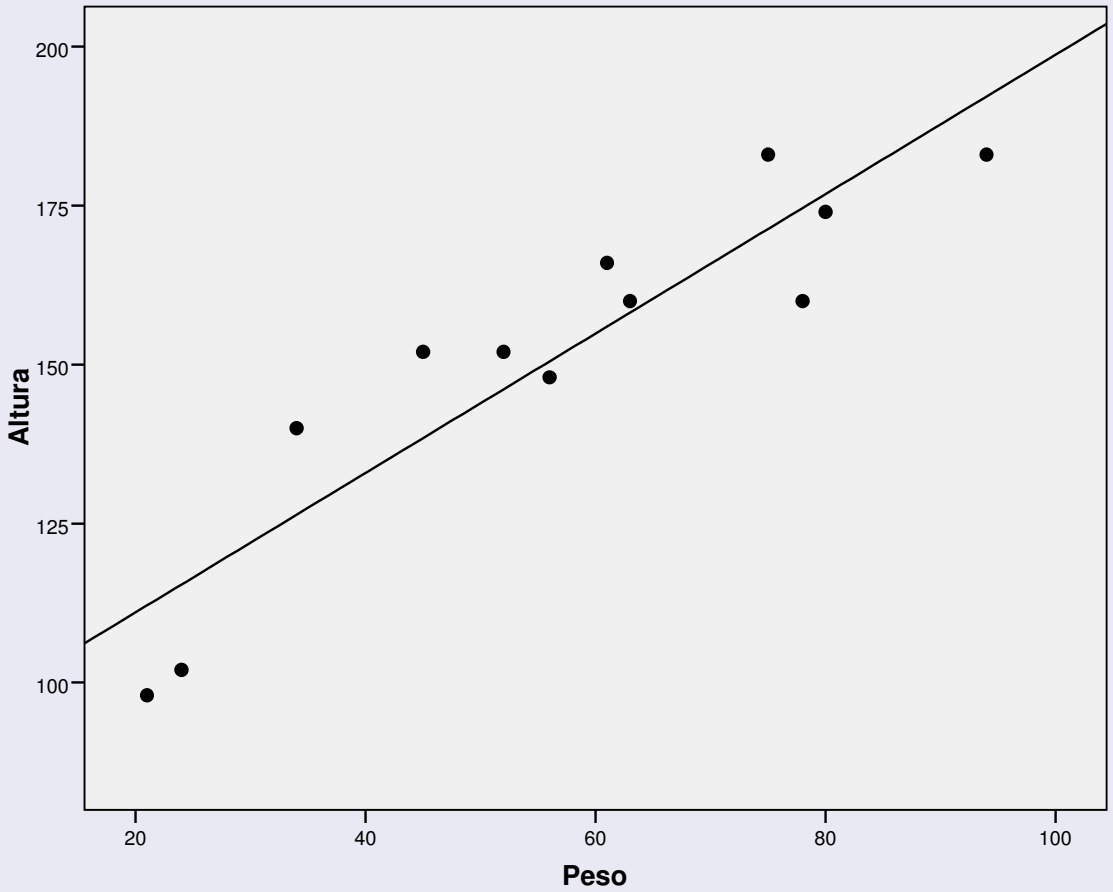
x_i	y_i	$(a+bx_i)$	$[y_i-(a+bx_i)]^2$
80	174	176.80	7.86
45	152	138.44	183.94
63	160	158.17	3.36
94	183	192.15	83.70
24	102	115.42	180.05
75	183	171.32	136.37
56	148	150.50	6.23
52	152	146.11	34.69
61	166	155.98	100.48
34	140	126.38	185.51
21	98	112.12	199.66
78	160	174.61	213.47
			1335.32

Coeficiente de determinación r^2

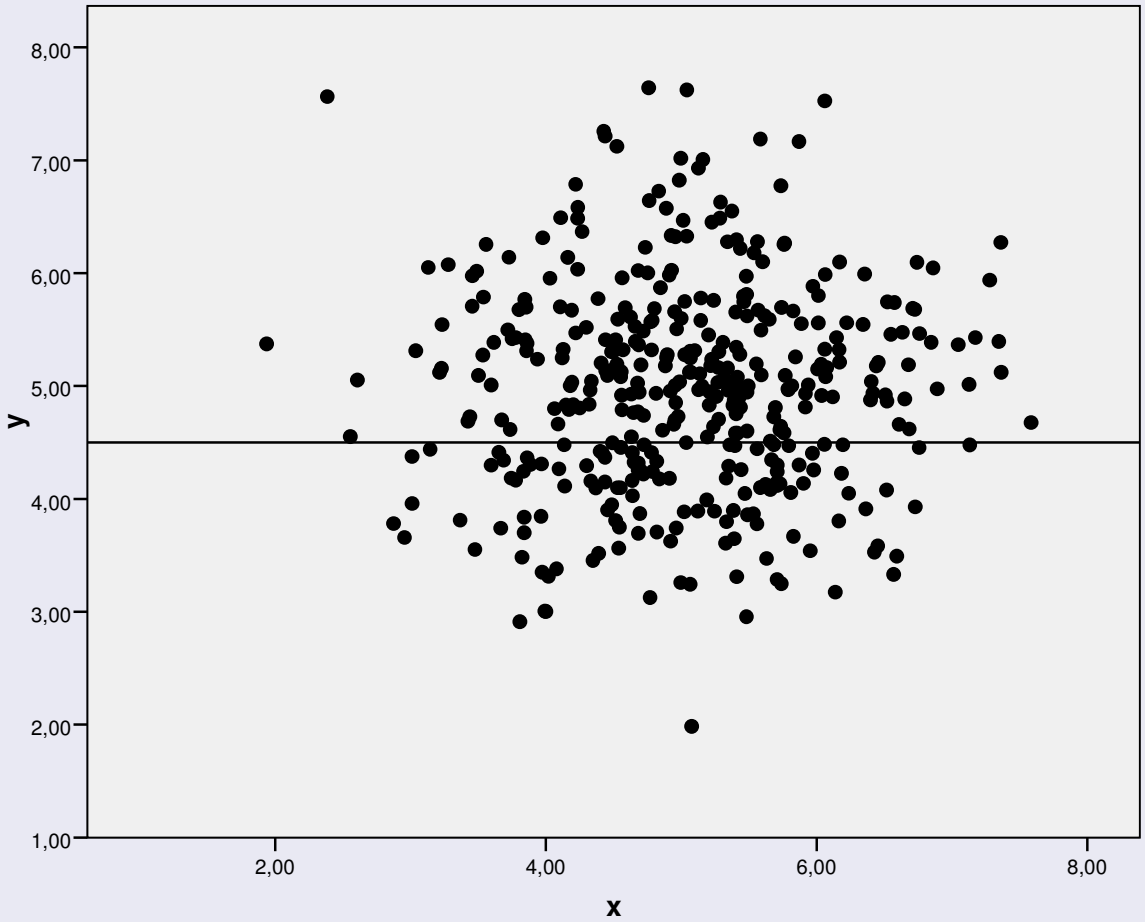
$$\frac{s^2_{y\leftarrow x}}{s^2_y} = 1 - r^2_{xy}$$

$1 - r^2_{xy}$ indica la proporción de la variabilidad total de Y no explicada por la regresión.
r^2_{xy} expresa lo contrario.

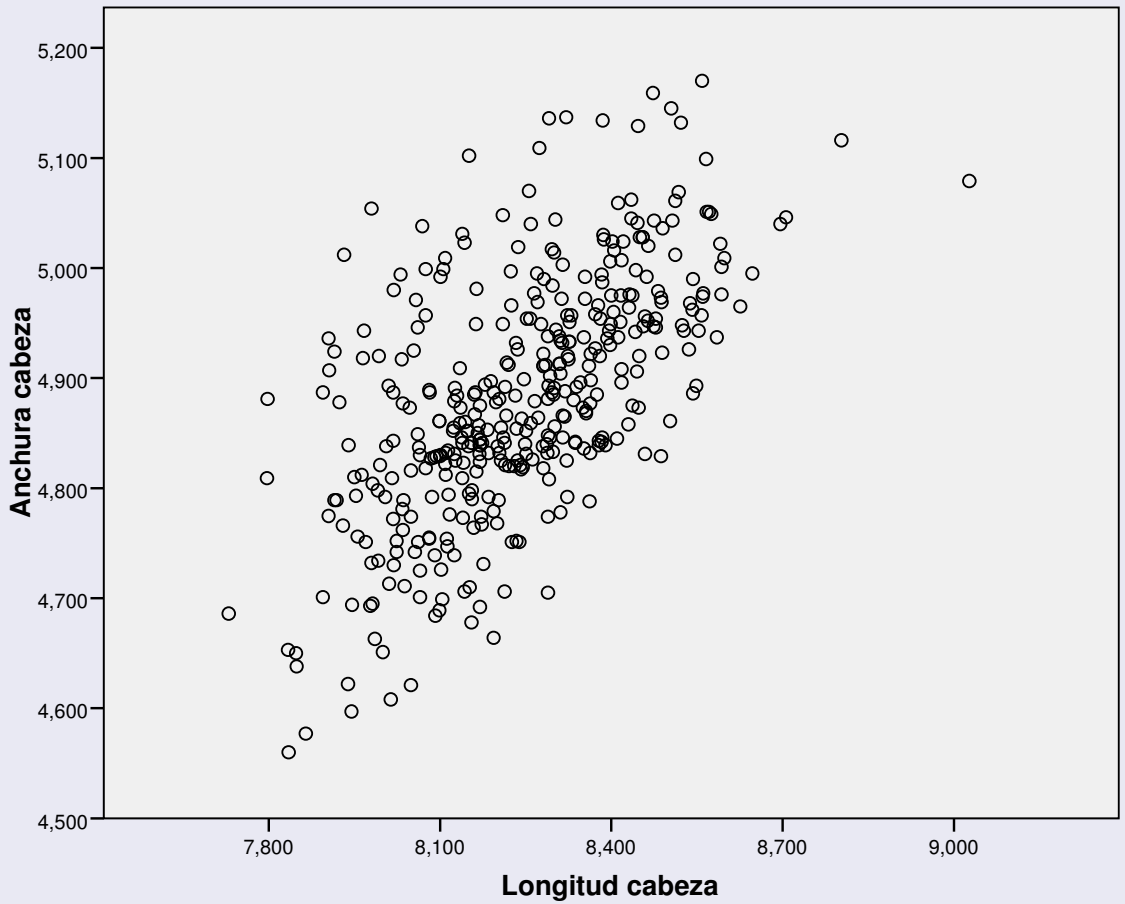
$r^2 = 0{,}82$



$r^2 \simeq 0$

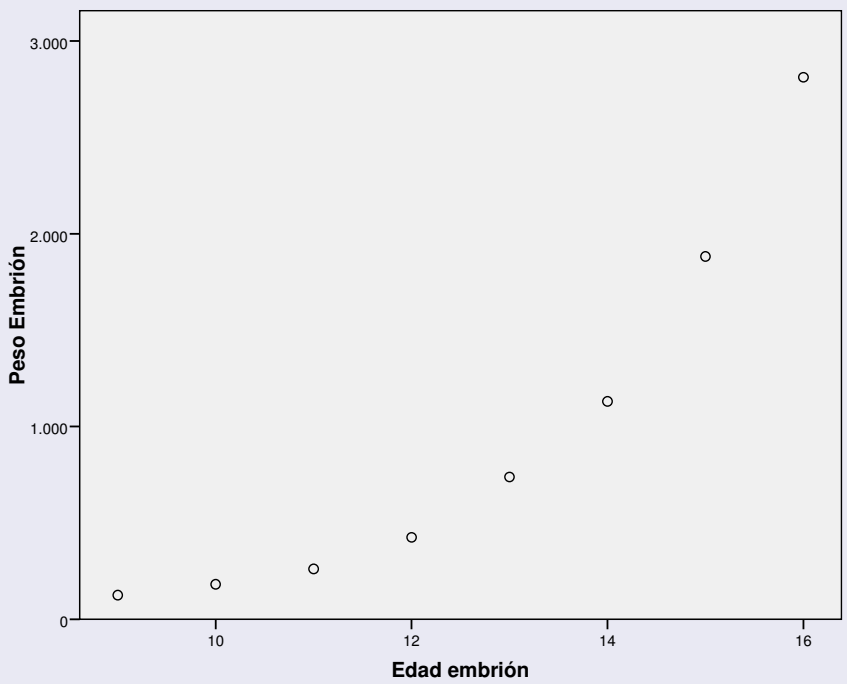


$r^2 = 0{,}39$


5,200
5,100
5,000
4,900
4,800
4,700
4,600
4,500
Anchura cabeza
7,800
8,100
8,400
8,700
9,000
Longitud cabeza

Regresión no lineal

Edad días-Peso embrión: Transformar variables



Relación entre variables cualitativas

nivel contaminación - salud árboles

		Cloroplastos			
	(3×3)	Alto	Medio	Bajo	Total
SO_2	Alto	3	4	13	20
	Medio	5	10	5	20
	Bajo	7	11	2	20
	Total	15	25	20	60

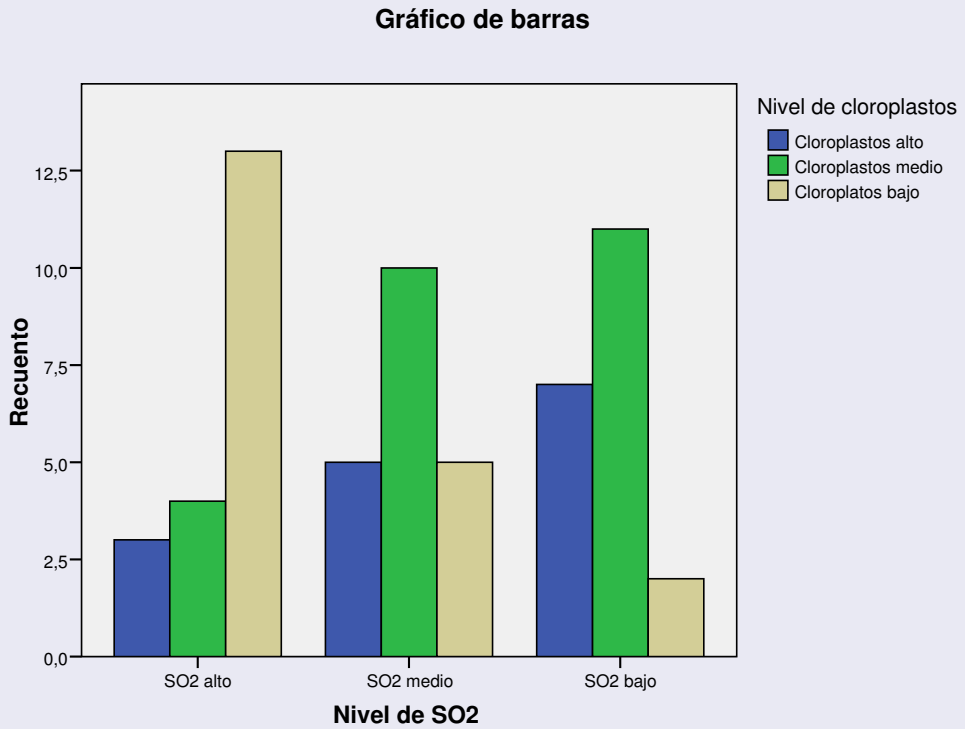
Vacunación-hepatitis

Vacunación

Hepatitis (2×2)	Sí	No	Total
Sí	11	70	81
No	538	464	1002
Total	549	534	1083

Gráfico

Barras agrupadas



Medidas del grado de dependencia

Observados vs Esperados independencia: distancia χ^2

$$\chi^2_{exp} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$0 \leq \chi^2_{exp} \leq +\infty$$

Coeficiente de contingencia de Pearson C

$$C = \sqrt{\frac{\chi^2_{exp}}{\chi^2_{exp} + n}}$$

$$0 \leq C \leq \sqrt{\frac{q-1}{q}}, \quad q = \texttt{min}\{\text{nº filas, nº colunas}\}$$

Ejemplo: cloropastos

Tabla 3×3. Por lo tanto,

$$0 \leq C \leq \sqrt{\frac{2}{3}} = 0{,}816$$

En este caso concreto,

$$C = 0{,}444$$

Grado de asociación medio

Independencia $C = 0$

Cloroplastos

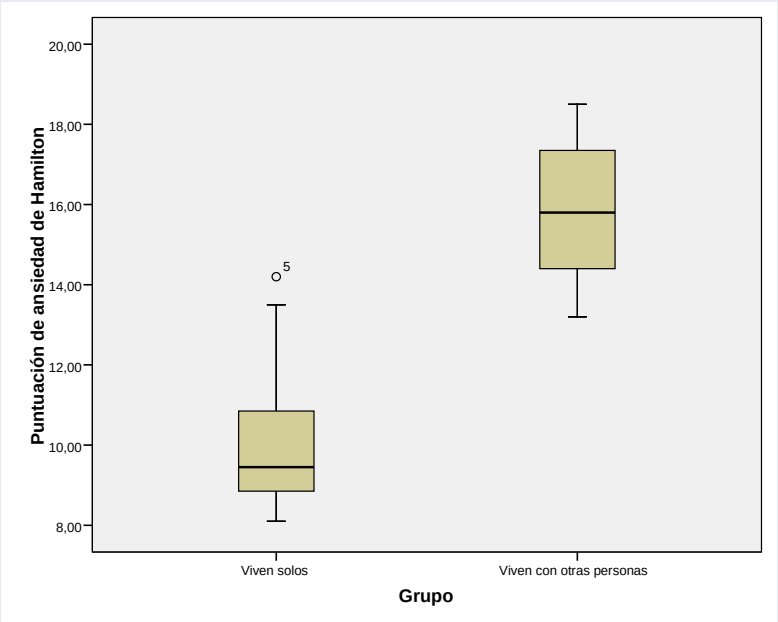
SO_2	(3×3)	Alto	Medio	Bajo	Total
	Alto	5	8.3	6.7	20
	Medio	5	8.3	6.7	20
	Bajo	5	8.3	6.7	20
	Total	15	25	20	60

Máxima dependencia $C = 0,816$

Los valores observados deberían ser éstos:

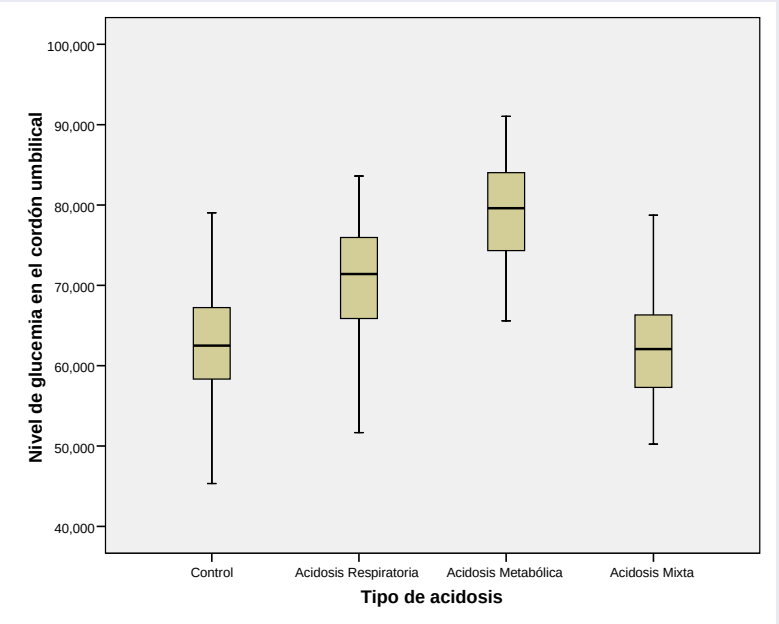
		Cloroplastos			
	(3×3)	Alto	Medio	Bajo	Total
SO_2	Alto	0	0	20	20
	Medio	0	20	0	20
	Bajo	20	0	0	20
	Total	20	20	20	60

Cualitativa → cuantitativa



¿Influye el estilo de vida en la ansiedad?

¿Influye la acidosis en la glucemia?



Contrastes de hipótesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Parámetros poblacionales

μ denota la media poblacional de una variable cuantitativa

Parámetros muestrales

Nosotros sólo contamos con los valores típicos ($\overline{x}$,s, etc) de una muestra de cada población.

Inferencia Estadística

En general, ¿cómo generalizar conclusiones a partir de una muestra?