Estudio descriptivo de dos variables

Metodología de la Investigación en Enfermería

Cátedra de Bioestadística
Universidad de Extremadura

1 de febrero de 2012

Índice

Índice

Índice

Índice

¿De qué trata?

Relación a nivel descriptivo entre dos variables

- Cuantitativa-Cuantitativa: 1. Regresión-Correlación (Bioestadística 2.1)
- Cualitativa-Cualitativa:
  - 2. Tablas de contingencia (Bioestadística 2.2)
  - 3. Factores de riesgo (Bioestadística: 5.4.3)
  - 4. Diagnóstico clínico (Bioestadística 5.4.4)
- Cualitativa-Cuantitativa: 5. Comparaciones de grupos o tratamientos (Bioestadística 5.5-Introducción)

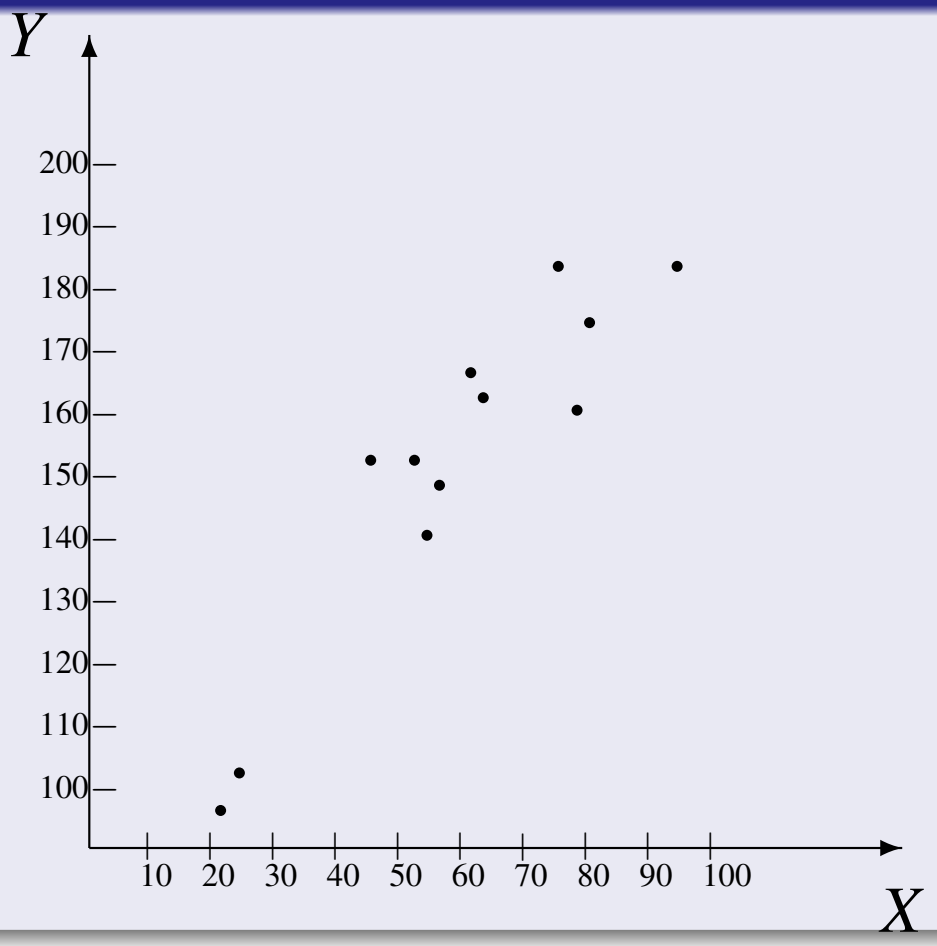
1. Tabla de frecuencias

No suele mostrarse pues los pares de datos raramente se repiten

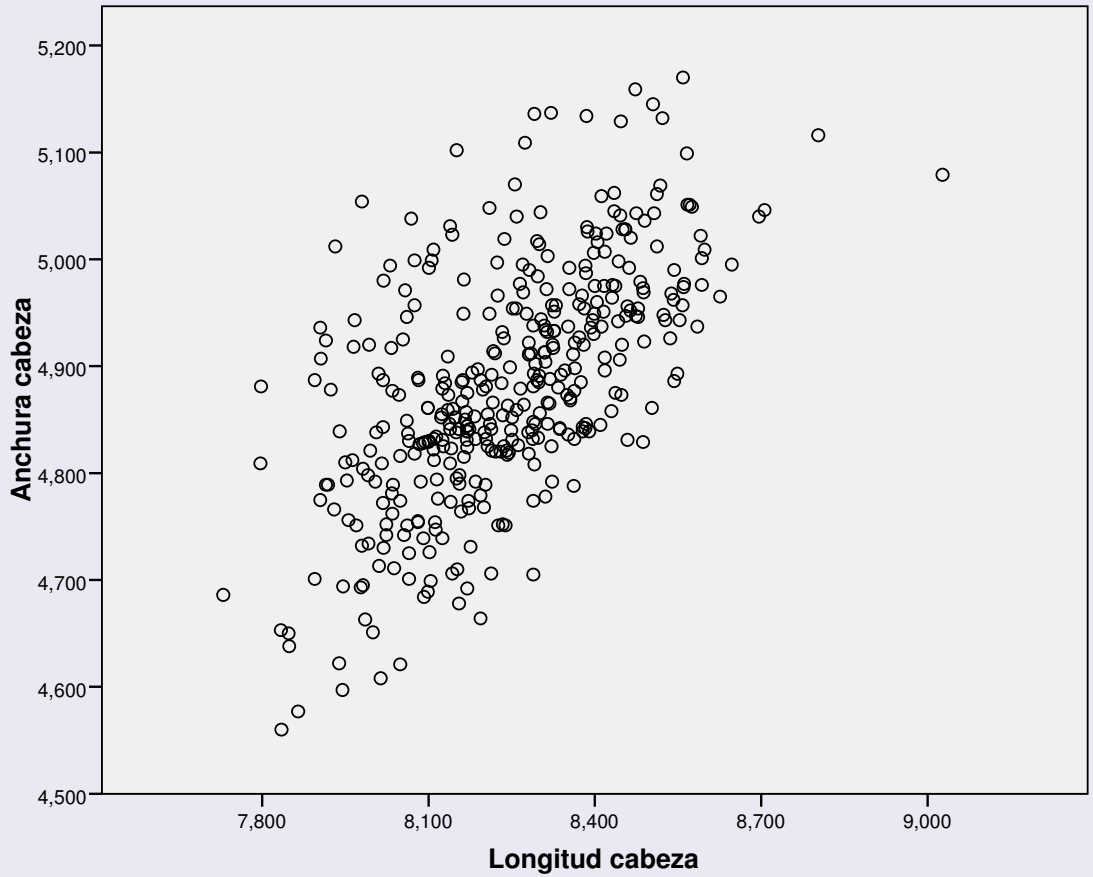
X =peso(kg)	80	45	63	94	24	75	56	...
Y =altura(cm)	174	152	160	183	102	183	148	...

2. Gráfico

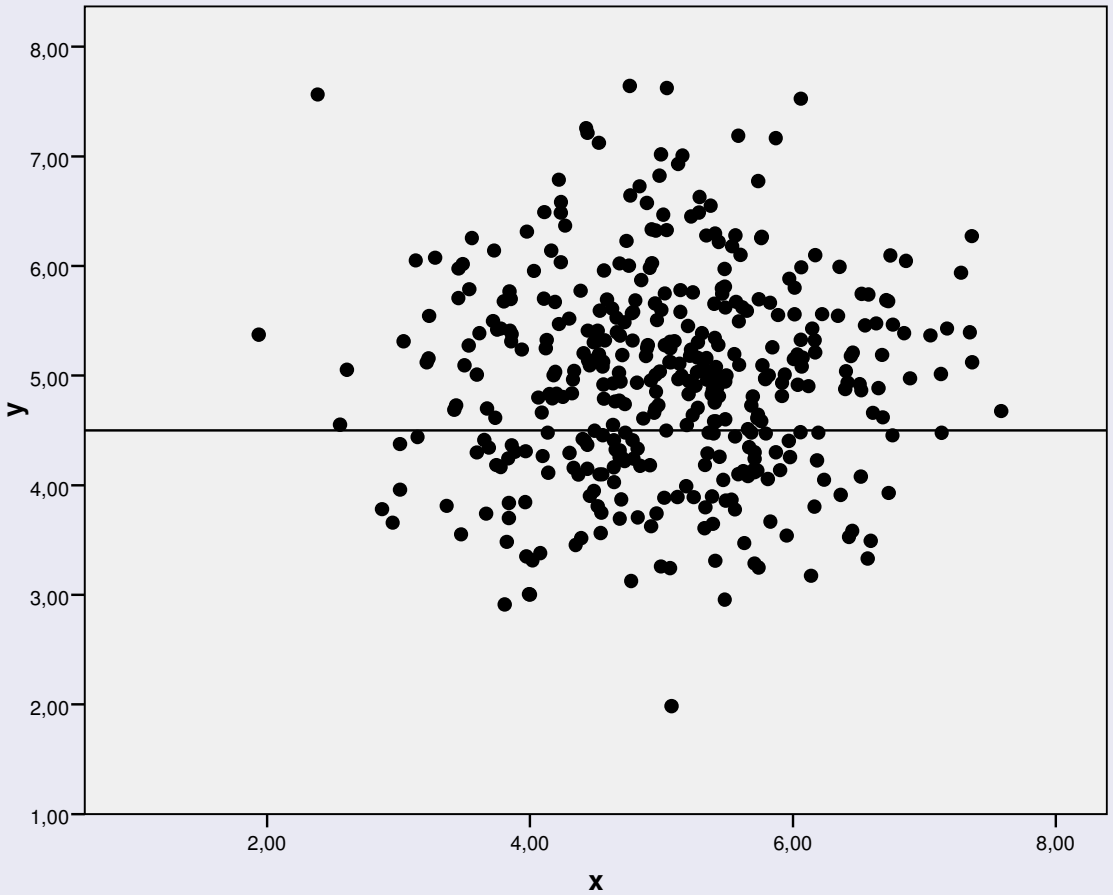
Diagrama de dispersión



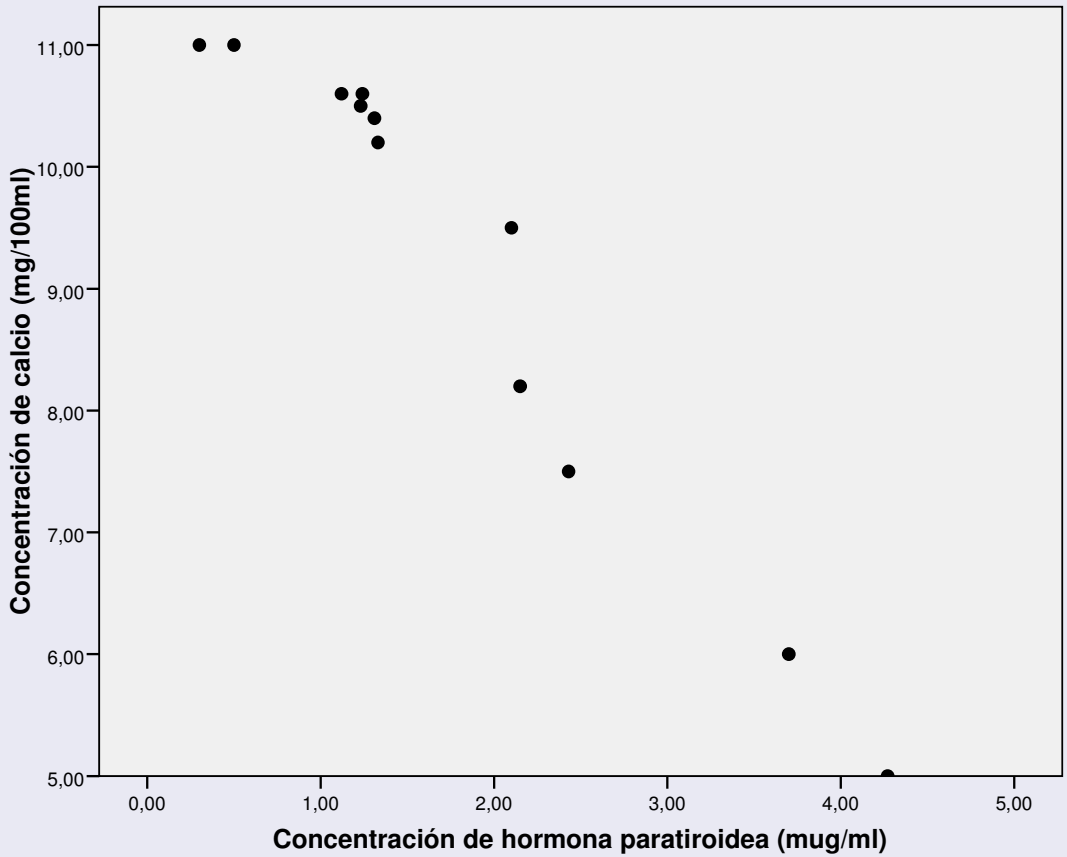
Otro ejemplo



Ausencia de de relación (independencia)



Estudiamos relaciones lineales



3. Valores típicos

Dos tipos

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

Variables por separado

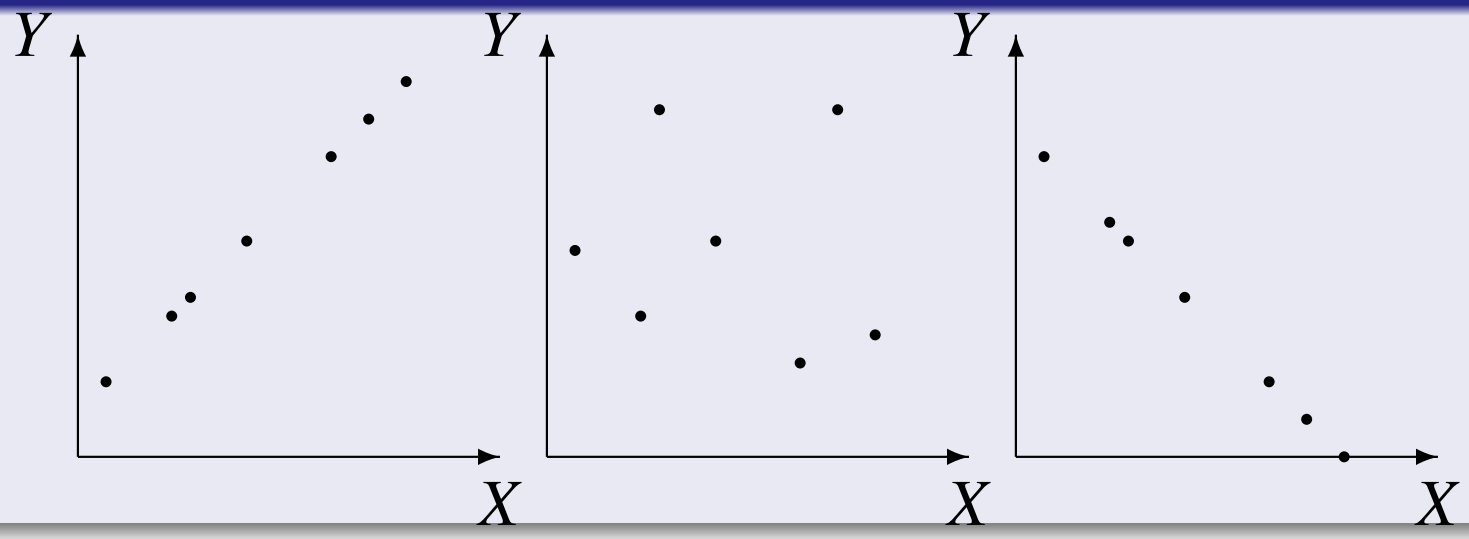
$$\bar{x}, \quad s_x, \quad \bar{y}, \quad s_y, \quad \tilde{y}, \ldots$$

Referentes a la relación entre las variables: Covarianza

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

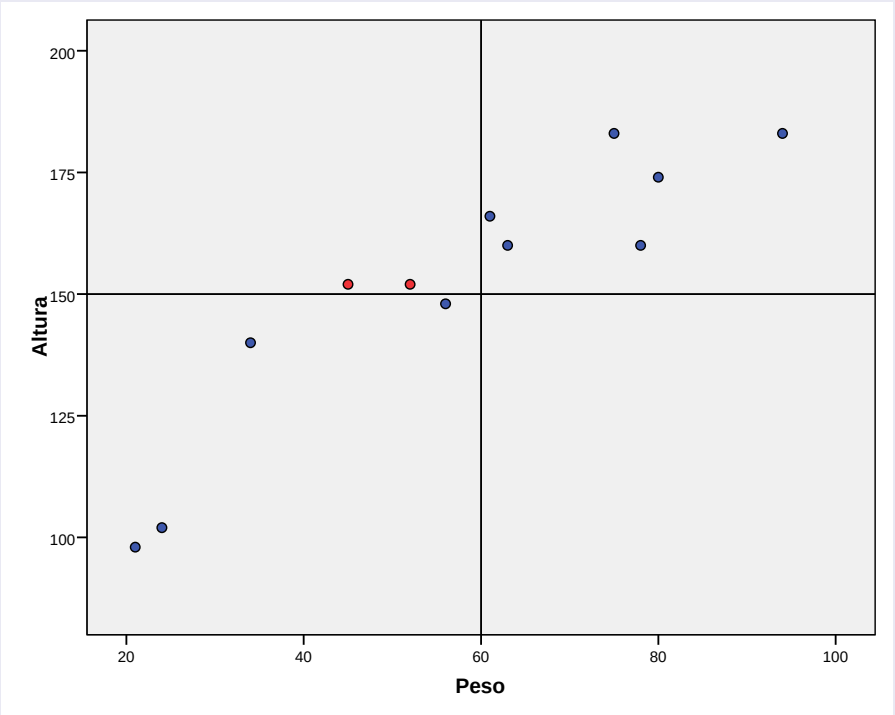
$$-s_x \cdot s_y \leq s_{xy} \leq +s_x \cdot s_y .$$

Interpretación gráfica

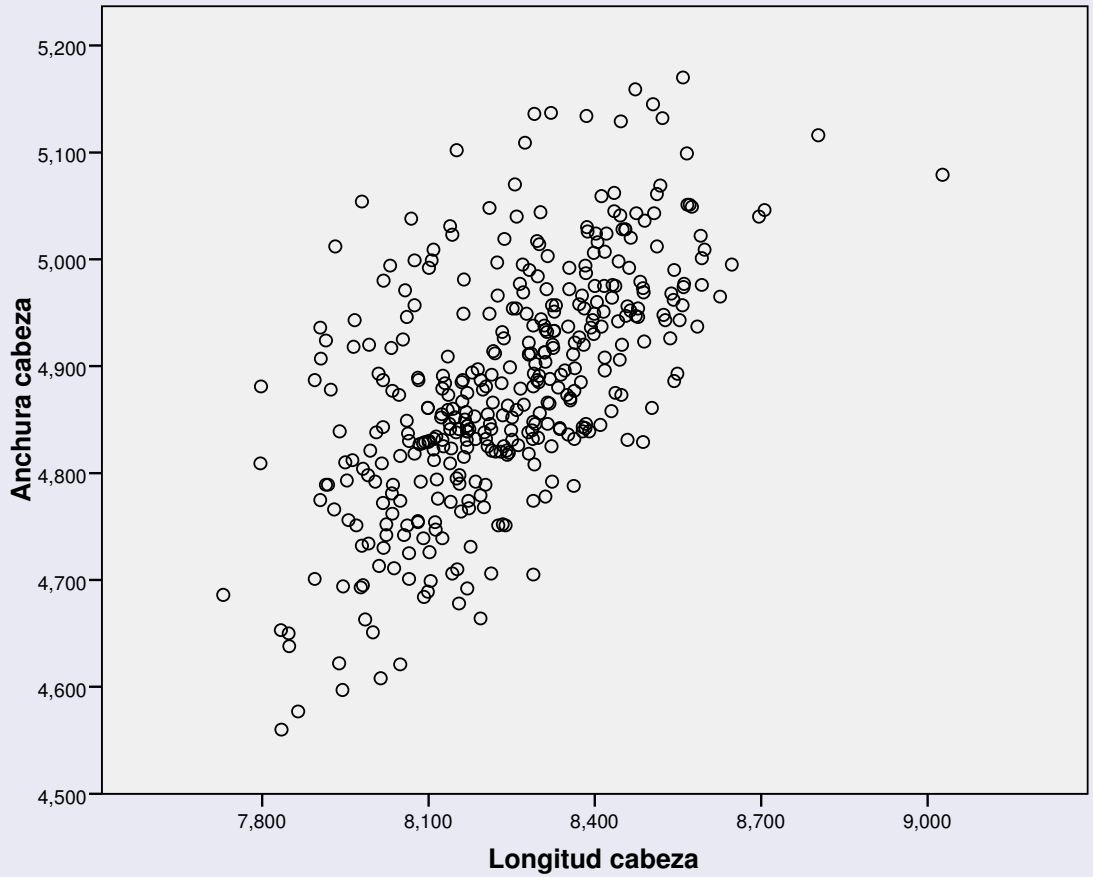


Interpretación gráfica

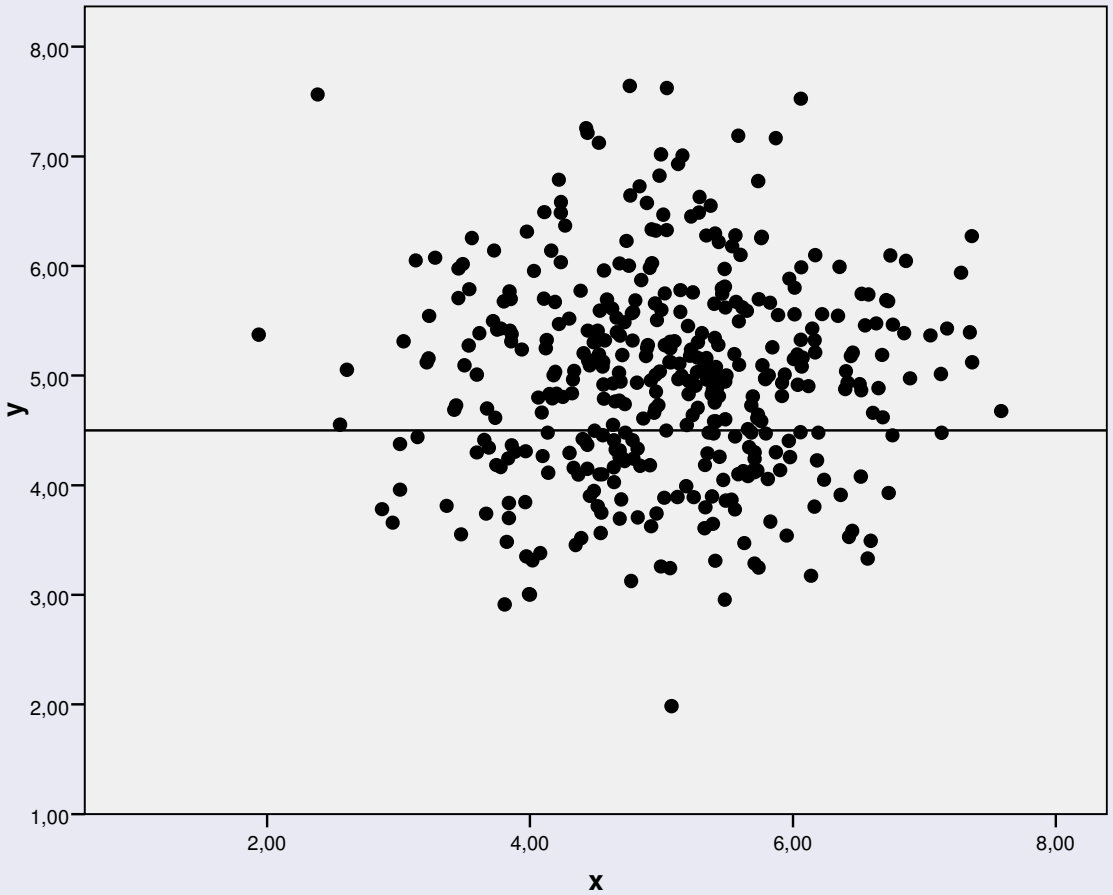
$$-630{,}71 \leq s_{xy} \leq +630{,}71 \qquad s_{xy} = 577{,}86$$



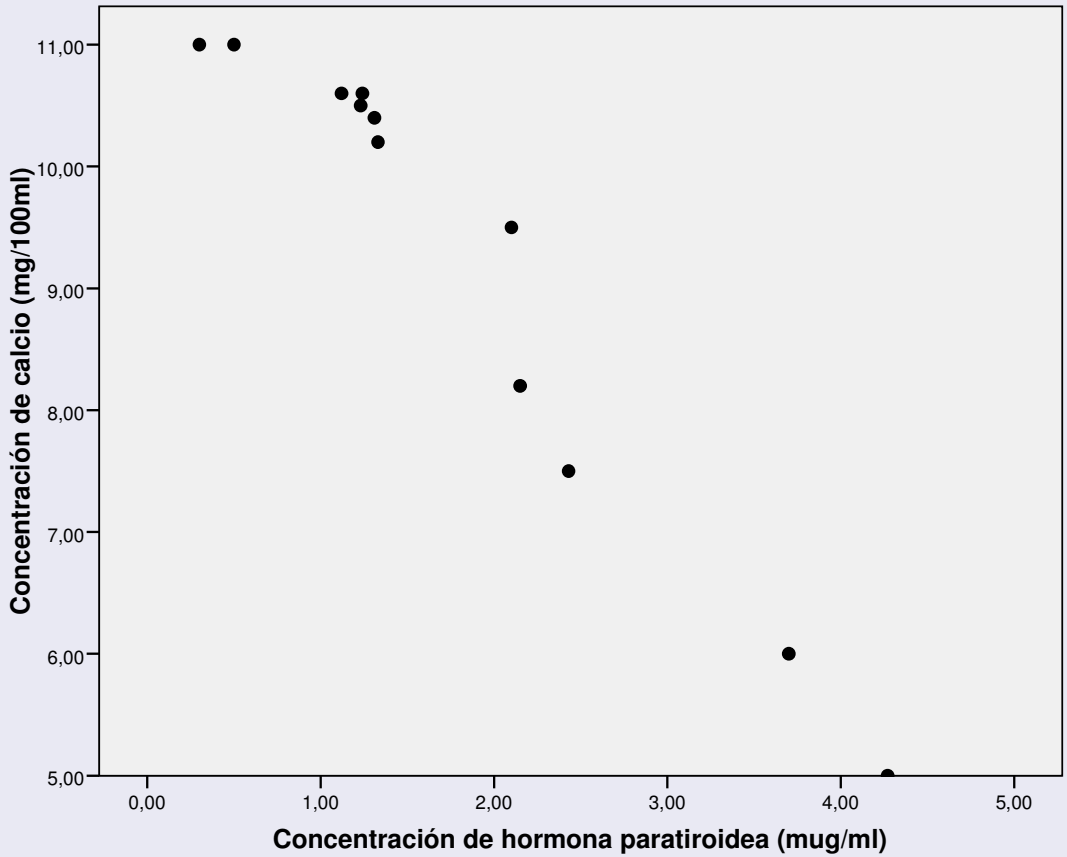
Covarianza positiva



Covarianza próxima a cero



Covarianza negativa



Coeficiente de correlación lineal r

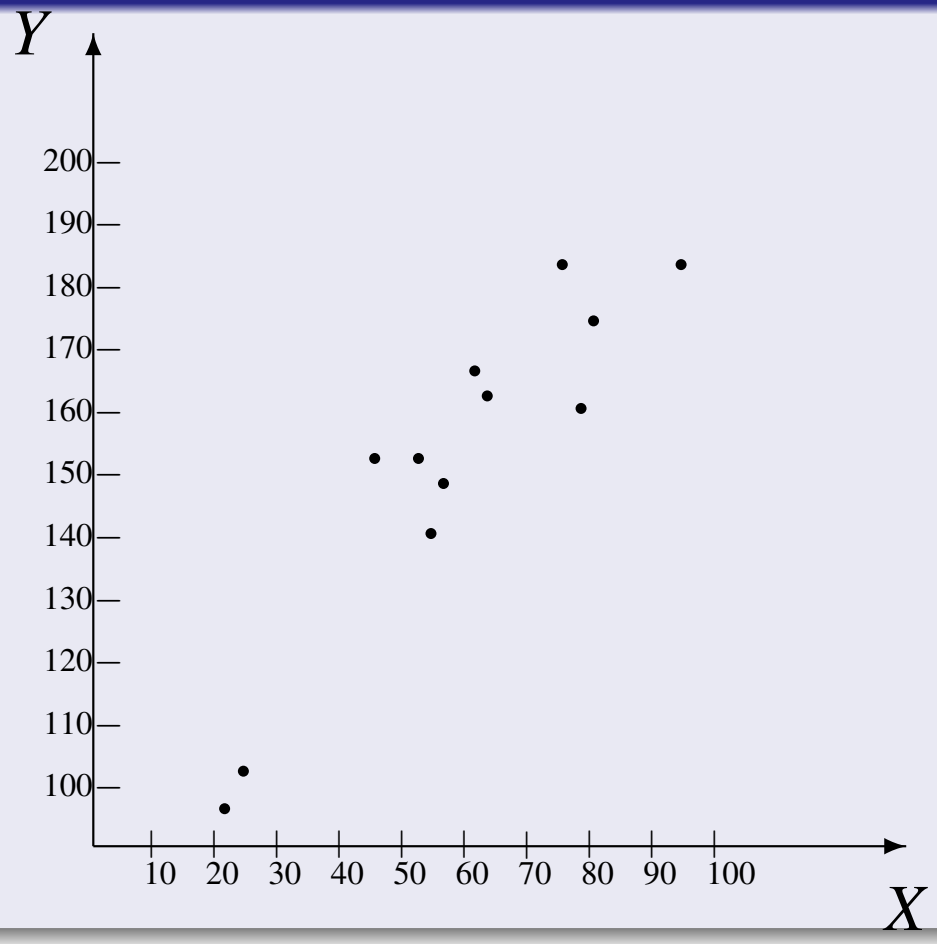
Medida adimensional del grado de correlación

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

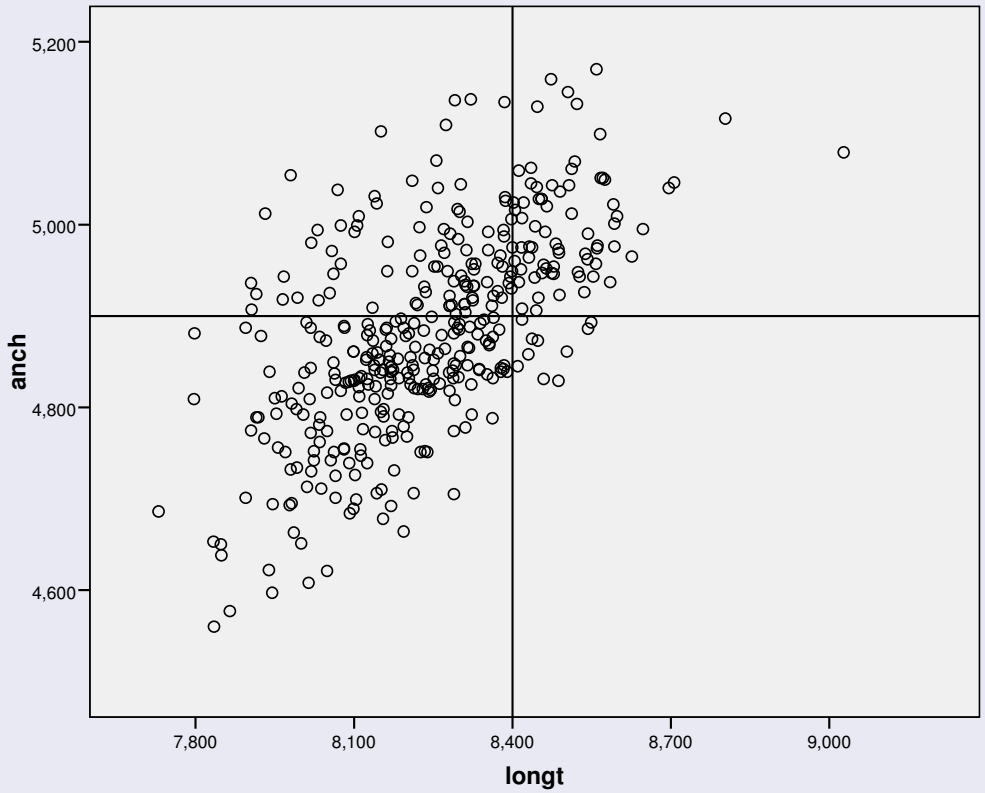
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

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

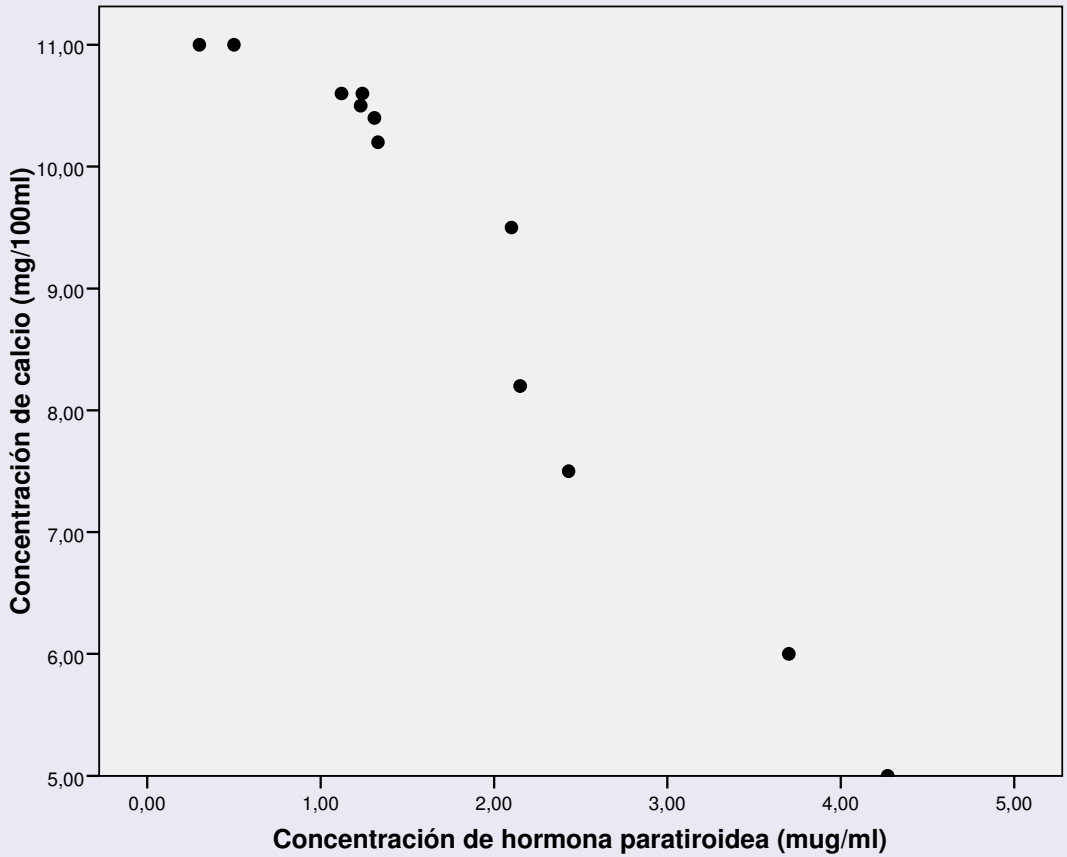
$r = 0{,}91$


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

$r = 0{,}625$



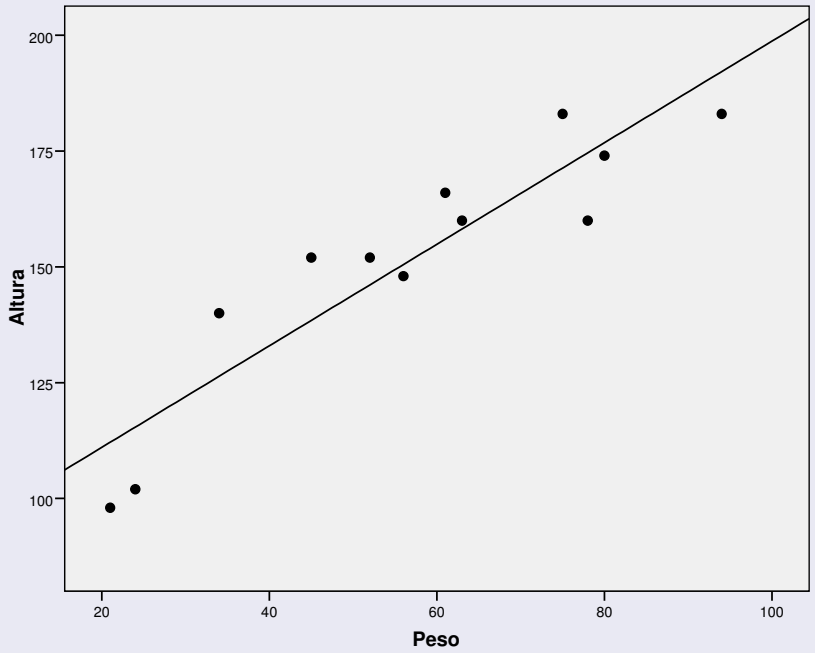
$r = -0{,}97$


11,00
10,00
9,00
8,00
7,00
6,00
5,00
Concentración de calcio (mg/100ml)
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: $x = 62kg \rightarrow \hat{y} = 89{,}11 + 1{,}10 \cdot 60 = 155{,}11cm$



Cálculo de la recta

El individuo i-ésimo tiene valores (x_i, y_i).

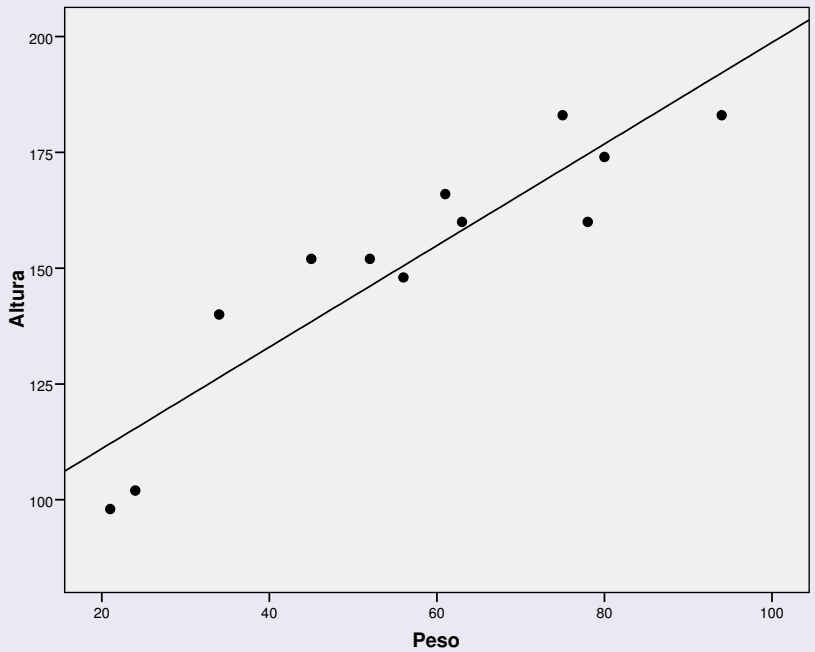
Una recta $y = a + b \cdot x$ predice para x_i el valor $a + bx_i$.

La recta es *apropiada* si la diferencia $y_i - (a + bx_i)$ es *pequeña*.

Solución mínimo-cuadrática

$$\texttt{min} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

$$y = 89{,}11 + 1{,}10x$$



Varianza residual

Mide el error cometido por la recta de regresión

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

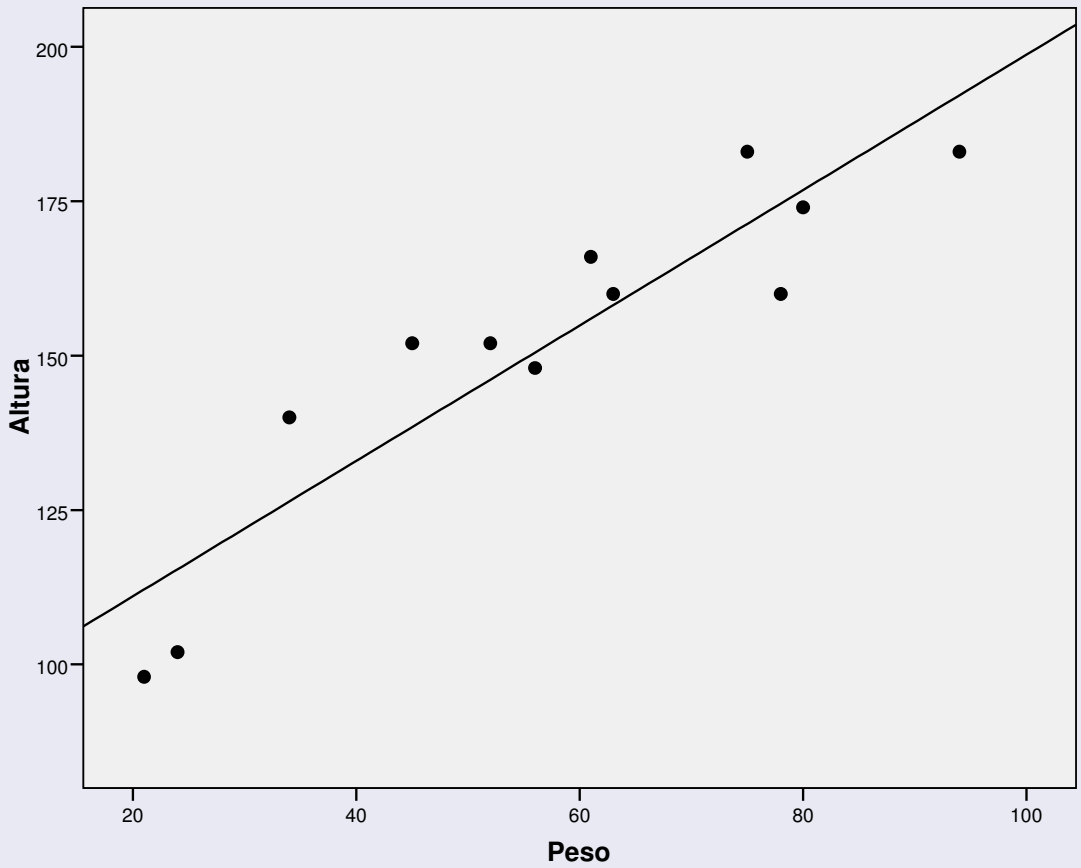
x_i	y_i	$(a+bx_i)$	$[y_i - (a+bx_i)]^2$
80	174	176.80	7.86
45	152	138.44	183.94
63	160	158.17	3.36
94	183	192.15	83.70
24	102	115.42	180.05
75	183	171.32	136.37
56	148	150.50	6.23
52	152	146.11	34.69
61	166	155.98	100.48
34	140	126.38	185.51
21	98	112.12	199.66
78	160	174.61	213.47
			1335.32

Coeficiciente de determinación r^2

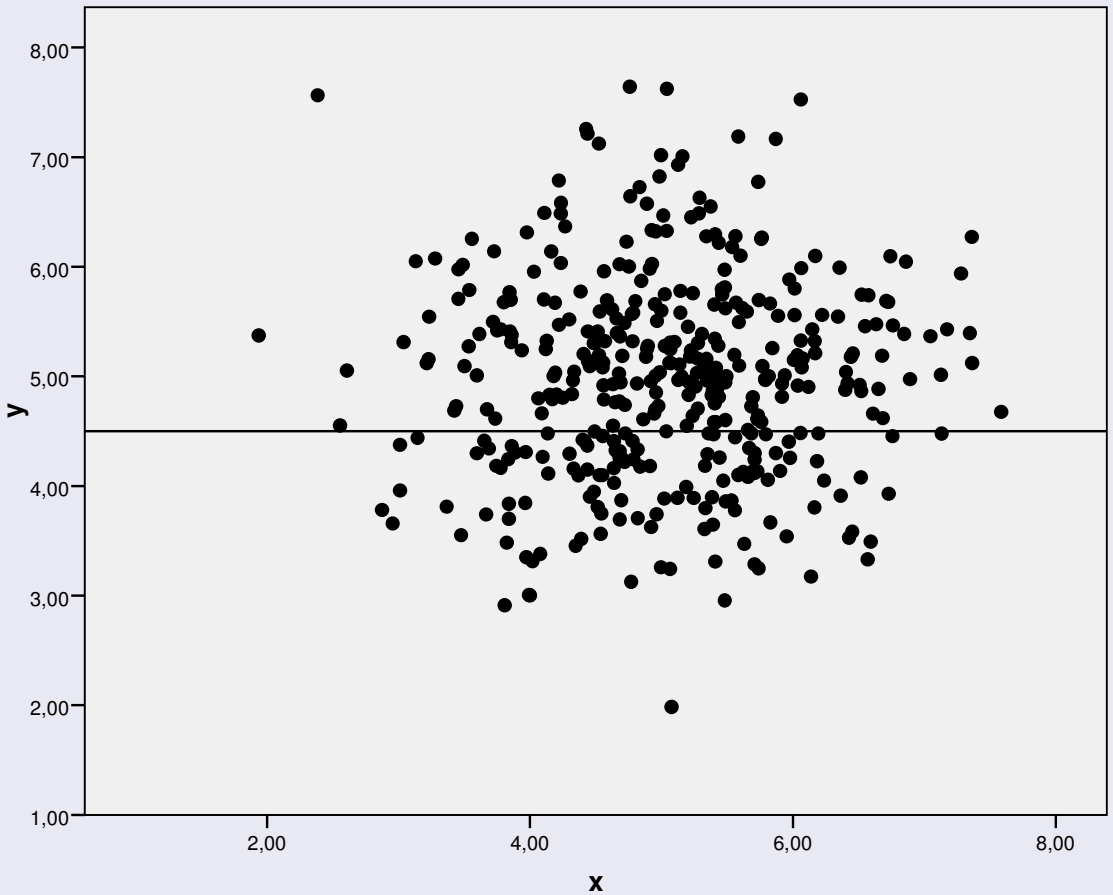
$$\frac{s^2_{y\leftarrow x}}{s^2_y} = 1 - r^2_{xy}$$

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

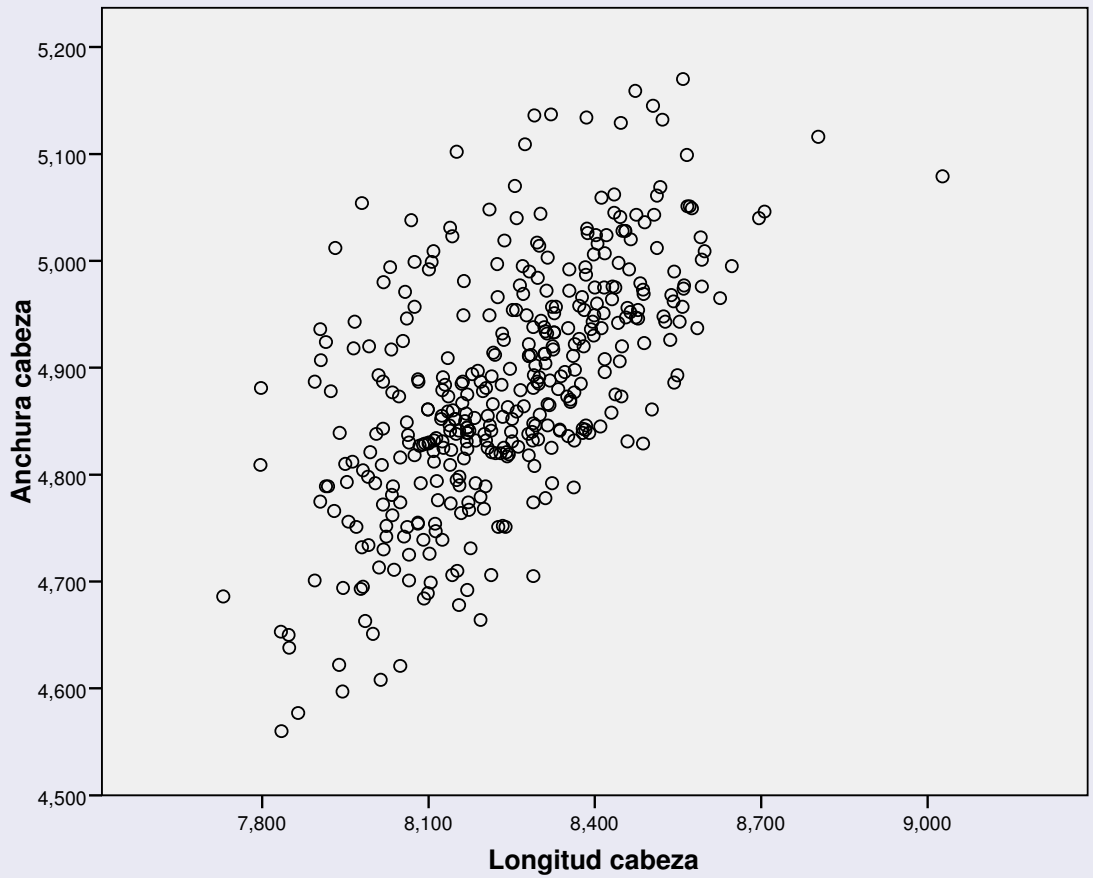
$r^2 = 0{,}82$



$r^2 \simeq 0$

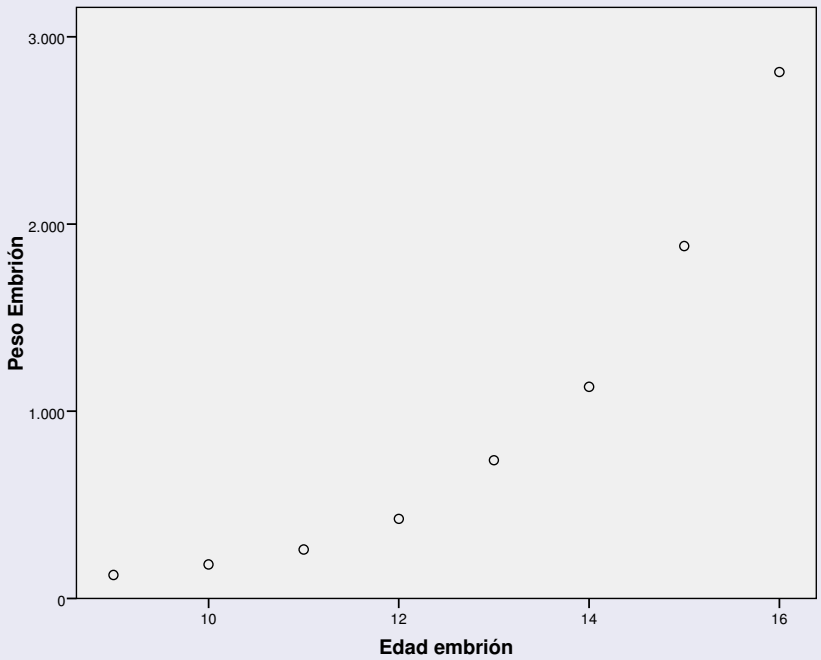


$r^2 = 0{,}39$


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

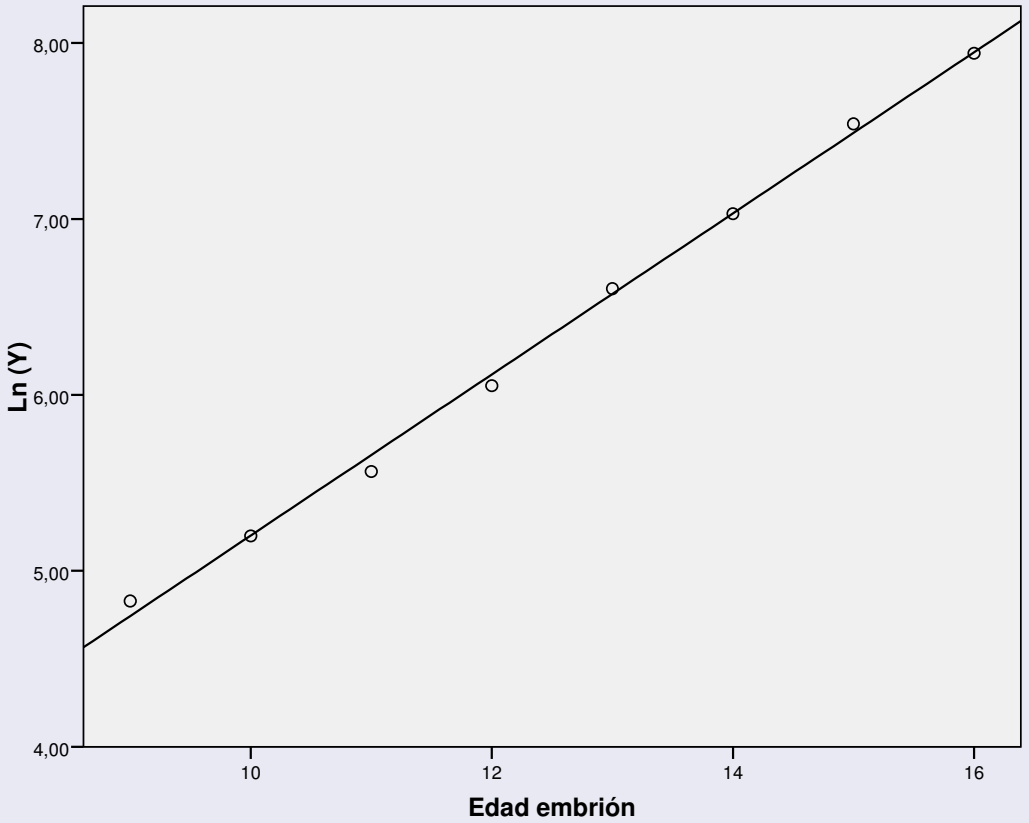
Regresión no lineal

Edad días-Peso embrión: Transformar variables



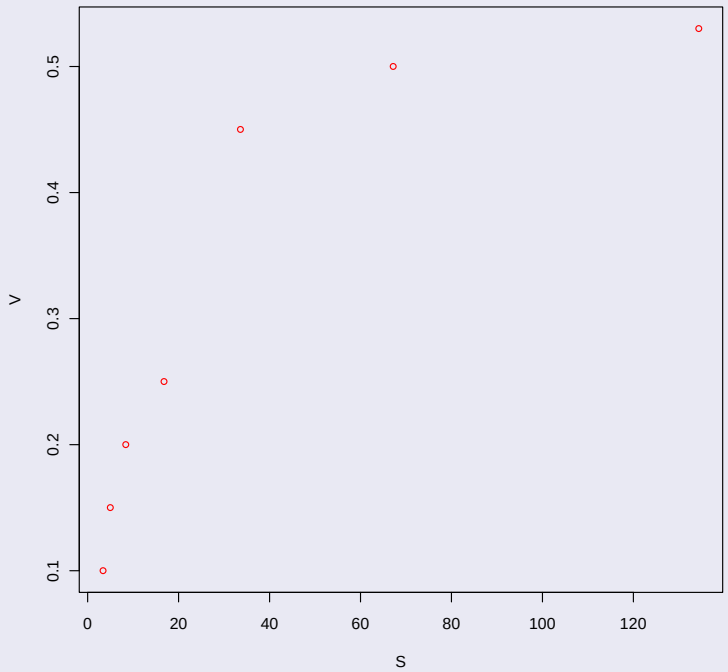
$\tilde{y} = \ln y, \tilde{x} = x \quad r = 0{,}99 \quad \tilde{y} = 0{,}62 + 0{,}46\tilde{x}$

Deshacemos la transformación: $y = 1{,}86 \cdot 1{,}58^x$



Ecuación de Michaellis-Menten

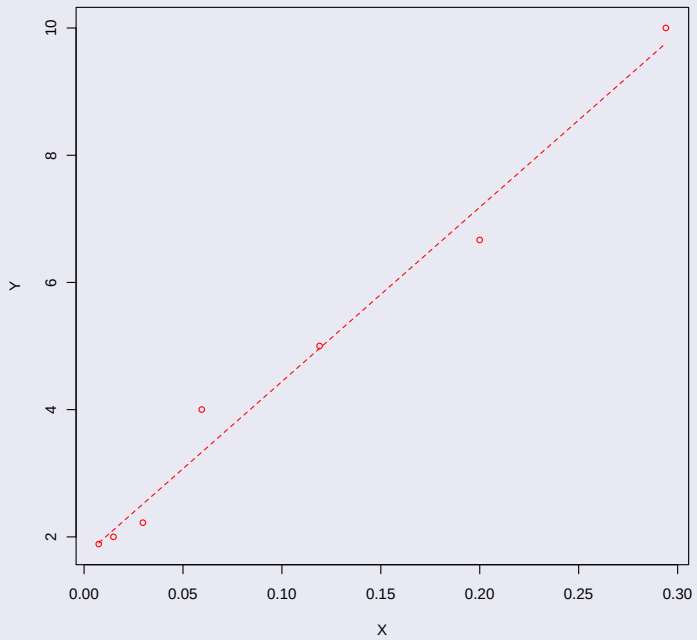
$[S]$	3.4	5.0	8.4	16.8	33.6	67.2	134.4
V	0.10	0.15	0.20	0.25	0.45	0.50	0.53



$y = 1/V; x = 1/S$

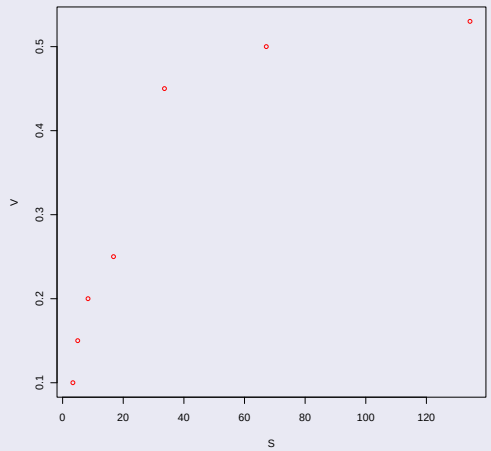
$y = 1{,}65 + 27{,}52x; \quad r = 0{,}99$

$[S]$	3.4	5.0	8.4	16.8	33.6	67.2	134.4
V	0.10	0.15	0.20	0.25	0.45	0.50	0.53



Deshacemos el cambio

$$V = \frac{0{,}60[S]}{1{,}67 + [S]}$$



$$V_{max} = 0{,}60 \qquad K_M = 16{,}67$$

Relación entre dos 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

Agente radiactivo - Cáncer tiroides

Exposición

Tumor	(2×2)	Sí	No	Total
	Sí	25	30	55
	No	4975	94970	99945
	Total	5000	95000	100000

Gen-tumor

Gen \ Tumor	Sí	No	Total
Sí	610	360	970
No	390	640	1030
Total	1000	1000	2000

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

Proporciones filas y columnas

$$\hat{P}(A_i) = \frac{O_{i\cdot}}{n} \qquad \hat{P}(SO_2 \texttt{ alto}) = \frac{20}{60} = 0.33$$

$$\hat{P}(B_j) = \frac{O_{\cdot j}}{n} \qquad \hat{P}(\texttt{Cloroplastos medio}) = \frac{25}{60} = 0.42$$

Proporciones celdas

$$\hat{P}(A_i \cap B_j) = \frac{O_{ij}}{n}$$

$$\hat{P}(SO_2 \texttt{ alto y Cloroplastos medio}) = \frac{4}{60} = 0.067$$

Proporciones condicionadas

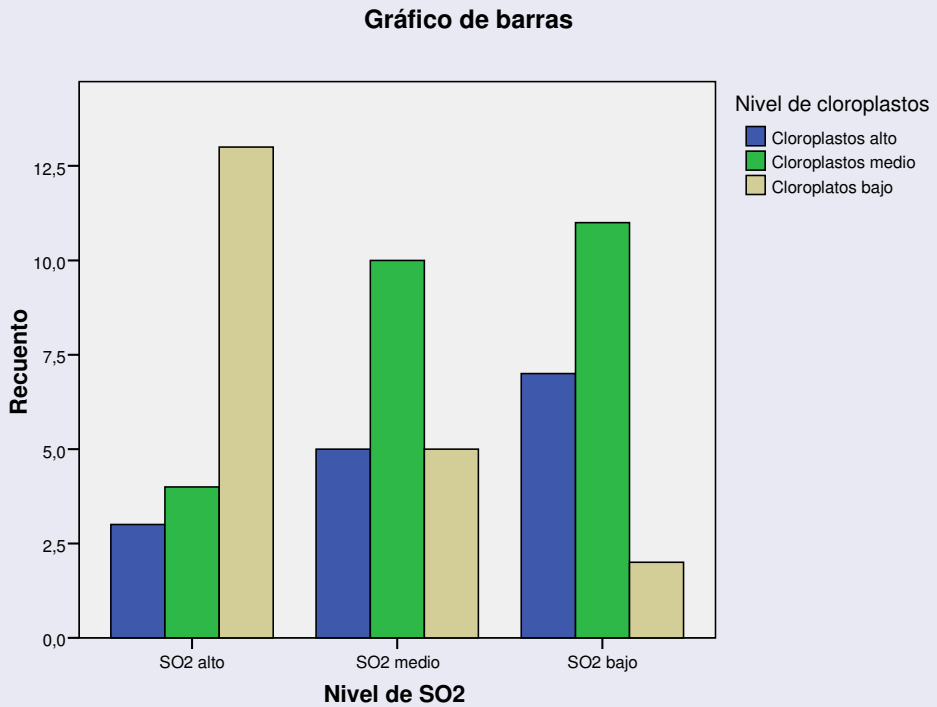
$$\hat{P}(A_i|B_j) = \frac{O_{ij}}{O_{\cdot j}} = \frac{\hat{P}(A_i \cap B_j)}{\hat{P}(B_j)}$$

$$\hat{P}\left(SO_2 \texttt{ alto} \middle| \texttt{Cloroplastos medio}\right) = \frac{4}{25} = 0.16$$

$$\hat{P}(B_j|A_i) = \frac{O_{ij}}{O_{i\cdot}} = \frac{\hat{P}(A_i \cap B_j)}{\hat{P}(A_i)}$$

$$\hat{P}\left(\texttt{Cloroplastos bajo} \middle| SO_2 \texttt{ alto}\right) = \frac{13}{20} = 0.65$$

Gráfico de barras agrupadas



Medidas del grado de dependencia

¿En qué consiste la ausencia de dependencia?

$$\hat{P}(B_j|A_i) = \hat{P}(B_j)$$

El hecho de que se verifique A_i no altera la proporción que se daba en general para B_j, o viceversa, para todo i y j.

$$\frac{\hat{P}(A_i \cap B_j)}{\hat{P}(A_i)} = \hat{P}(B_j)$$

Independencia (sobre la muestra)

$$\hat{P}(A_i \cap B_j) = \hat{P}(A_i) \cdot \hat{P}(B_j)$$

¿Qué deberíamos esperar en caso de independencia?

$$\hat{P}(A_i \cap B_j) = \hat{P}(A_i) \cdot \hat{P}(B_j)$$

$$\frac{O_{ij}?}{n} = \frac{O_{i\cdot}}{n} \times \frac{O_{\cdot j}}{n}$$

$$O_{ij}? = \frac{O_{i\cdot} \times O_{\cdot j}}{n}$$

En definitiva

$$E_{ij} = \frac{O_{i\cdot} \times O_{\cdot j}}{n}$$

Ejemplo

En el caso de los cloroplastos, de no existir relación alguna cabría esperar las siguientes observaciones:

`Cloroplastos`

	(3×3)	Alto	Medio	Bajo	Total
SO_2	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

Observados vs esperados

Cuando mayor sea la diferencia entre estas tablas más fuerte será la correlación:

`Cloroplastos`

	(3×3)	Alto	Medio	Bajo	Total
SO_2	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

`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

Observados vs Esperados: 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

	(3×3)	Alto	Medio	Bajo	Total
SO_2	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

Caso especialmente sencillo: tablas 2×2

Vacunación

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

$0 \leq C \leq 0{,}707 \qquad C = 0{,}206$

Otra medida de asociación: ϕ

$$\phi^2 = \frac{\chi^2_{exp}}{n}, \qquad 0 \leq \phi \leq 1$$

$$\phi = \sqrt{\frac{(11 \cdot 464 - 70 \cdot 538)^2}{81 \cdot 1002 \cdot 549 \cdot 534}} = 0{,}211$$

Vacunación

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

Caso extremo: $\phi = 1$

Vacunación

Hepatitis	Sí	No	Total
Sí	0	81	81
No	1002	0	1002
Total	1002	81	1083

$$\hat{P}(\texttt{hepatitis}|\texttt{vacunados}) = 0$$
$$\hat{P}(\texttt{hepatitis}|\texttt{no vacunados}) = 1$$

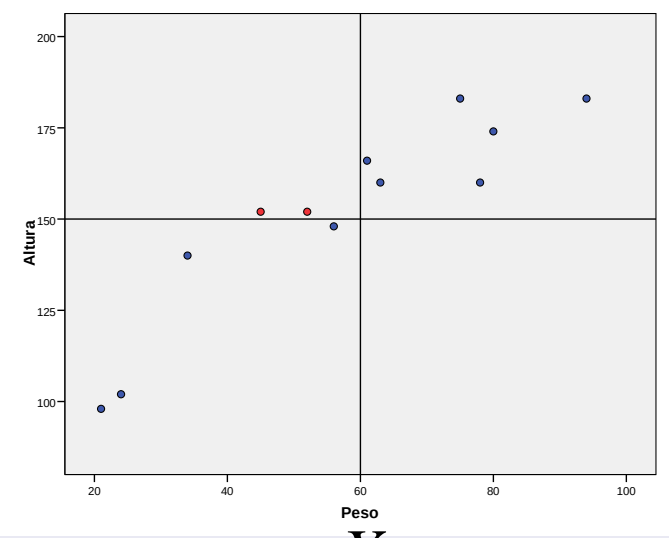
Caso extremo: $\phi = 0$

Vacunación

Hepatitis	Sí	No	Total
Sí	334	27	361
No	668	54	722
Total	1002	81	1083

$$\hat{P}(\texttt{hepatitis}|\texttt{vacunados}) = 0{,}33$$
$$\hat{P}(\texttt{hepatitis}|\texttt{no vacunados}) = 0{,}33$$

Analogía correlación: $\phi \leftrightarrow r$



X

Y		−	+	Tot
	+	2	6	8
	−	4	0	4
	Tot	6	6	12

Factores de riesgo

Agente radiactivo - Cáncer tiroides

Exposición

Tumor	(2×2)	Sí	No	Total
	Sí	25	30	55
	No	4975	94970	99945
	Total	5000	95000	100000

(2×2)	FR	$\overline{FR}$	Total
E	a	b	a+b
$\overline{E}$	c	d	c+d
Total	a+c	b+d	n

Tipos de estudios

- **Transversal o de prevalencia:** Muestra aleatoria amplia estudiada con el objeto de estimar la prevalencia de la enfermedad.
- **De seguimiento o de cohortes:** Se seleccionan un grupo de expuestos al factor y otro de no expuestos para seguir su evolución. Tiene por objeto estimar la incidencia de la enfermedad por grupos. No permite estimar prevalencia.
- **Retrospectivos o de casos-control:** Se selecciona un grupo de enferemos y otro de sanos para indagar si han estado expuestos al factor de riesgo. No permite estimar prevalencia ni incidencia.

Ejemplos

Exposición

Tumor	(2×2)	Sí	No	Total
	Sí	25	30	55
	No	4975	94970	99945
	Total	5000	95000	100000

Tumor

Gen		Sí	No	Total
	Sí	610	360	970
	No	390	640	1030
	Total	1000	1000	2000

Estimaciones

Prevalencia

$P(E)$. P denota la proporción (desconocida) en toda la población. La estimamos mediante la proporción $\hat{P}$ en la tabla (muestra). Sólo en estudios de prevalencia:

(2×2)	FR	$\overline{FR}$	Total
E	a	b	a+b
$\overline{E}$	c	d	c+d
Total	a+c	b+d	n

$$\hat{P}(E) = \frac{a+b}{n}$$

Incidencias por cohortes

$P(\texttt{E}|\texttt{FR})$; $P(\texttt{E}|\overline{\texttt{FR}})$. Ejemplo: $P(\texttt{Hep}|\texttt{No vac})$; $P(\texttt{Hep}|\texttt{Vac})$
Sólo en estudios de cohortes (no caso-control)

(2×2)	FR	$\overline{FR}$	Total
E	a	b	a+b
$\overline{E}$	c	d	c+d
Total	a+c	b+d	n

$$\hat{P}(\texttt{E}|\texttt{FR}) = \frac{a}{a+c} \qquad \hat{P}(\texttt{E}|\overline{\texttt{FR}}) = \frac{b}{b+d}$$

$$\hat{P}(\texttt{Tumor}|\texttt{Exposición}) = \frac{25}{5000} = 0{,}500\,\%$$
$$\hat{P}(\texttt{Tumor}|\texttt{No exposición}) = \frac{30}{95000} = 0{,}032\,\%$$

Riesgo atribuible

Diferencia absoluta entre incidencias (cohortes)

$$RA = P(\mathtt{E}|\mathtt{FR}) - P(\mathtt{E}|\overline{\mathtt{FR}})$$

En nuestro ejemplo (exposición agente radioactivo):

$$\begin{aligned} \hat{RA} &= \hat{P}(\mathtt{Hep}|\mathtt{No\ \ vac}) - \hat{P}(\mathtt{Hep}|\mathtt{Vac}) \\ &= 0{,}500\,\% - 0{,}032\,\% \\ &= 0{,}468\,\% \end{aligned}$$

Fracción atribuible a la exposición

Diferencia relativa. Indica la parte de riesgo de los expuestos que se debe al factor en sí.

$$FA = \frac{P(\mathtt{E}|\mathtt{FR}) - P(\mathtt{E}|\overline{\mathtt{FR}})}{P(\mathtt{E}|\mathtt{FR})}$$

En nuestro ejemplo (exposición agente radioactivo):

$$\begin{aligned} \hat{FA} &= \frac{\hat{P}(\mathtt{Tumor}|\mathtt{Exp}) - \hat{P}(\mathtt{Tumor}|\mathtt{No\ exp})}{\hat{P}(\mathtt{Tumor}|\mathtt{Exp})} \\ &= \frac{0{,}468}{0{,}500} \\ &= 0{,}936\ (93{,}6\,\%) \end{aligned}$$

Riesgo relativo

$$RR = \frac{P(\mathtt{E}|\mathtt{FR})}{P(\mathtt{E}|\overline{\mathtt{FR}})}$$

En nuestro ejemplo (exposición radioactividad):

$$RR = \frac{P(\mathtt{Tumor}|\mathtt{Exp})}{P(\mathtt{Tumor}|\mathtt{No\ exp})}$$

Estimación:

$$\hat{RR} = \frac{\hat{P}(\mathtt{Tumor}|\mathtt{Exp})}{\hat{P}(\mathtt{Tumor}|\mathtt{No\ exp})} = \frac{0{,}500}{0{,}032} = 15{,}6$$

Se estima que es 15.6 veces más probable desarrollar ese tipo de tumor si se está expuesto al agente radioactivo que si no se está expuesto.

Odds Ratio

Aunque puede utilizarse también en los estudios de cohortes, es la única medida apropiada para los de caso-control porque en éstos no es posible estimar correctamente las incidencias al haber escogido una cantidad determinada de enfermos en la muestra, normalmente muy por encima de lo que corresponde a la prevalencia de la enfermedad.

Ejemplo: tumor-gen

		Tumor		
		Sí	No	Total
Gen	Sí	610	360	970
	No	390	640	1030
	Total	1000	1000	2000

Rojo: El gen es factor de riesgo $(610 \cdot 640)$

Azul: El gen no es factor de riesgo $(360 \cdot 390)$

Razón de productos cruzados

Gen \ Tumor	Sí	No	Total
Sí	610	360	970
No	390	640	1030
Total	1000	1000	2000

$$\hat{OR} = \frac{610 \cdot 640}{360 \cdot 390} = 2{,}70$$

Ejemplo exposición radioactividad

Exposición

Tumor	(2×2)	Sí	No	Total
	Sí	25	30	55
	No	4975	94970	99945
	Total	5000	95000	100000

$$\hat{OR} = \frac{25 \cdot 94970}{30 \cdot 4975} = 15{,}9$$

Recordemos $\hat{RR} = 15{,}6$

Interpretación formal de $\hat{OR}$

A partir de la denominada Regla de Bayes podemos probar que la razón de productos cruzados $\hat{OR}$ es una estimación válida del siguiente parámetro poblacional denominado Odds ratio (de ahí la notación OR):

$$OR = \frac{\frac{P(\mathrm{E}|\mathrm{FR})}{P(\overline{\mathrm{E}}|\mathrm{FR})}}{\frac{P(\mathrm{E}|\overline{\mathrm{FR}})}{P(\overline{\mathrm{E}}|\overline{\mathrm{FR}})}}$$

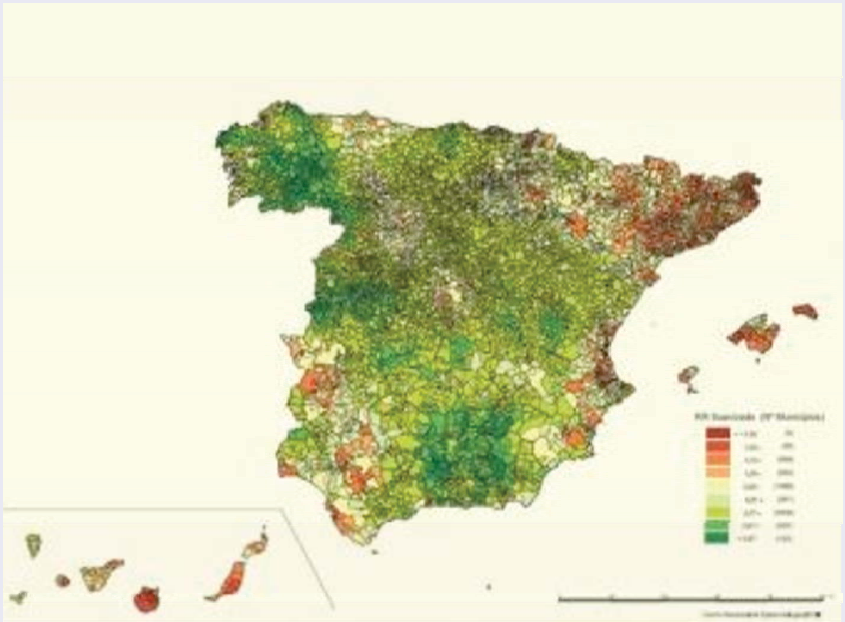
De ahí que nuestro parámetro se denomine con frecuencia Odds Ratio.

Tipos de tablas (estudios)

- Prevalencia (exposición aleatoria, incidencia aleatoria): `prev`, *RA*, *FA*, *RR*, *OR*.
- Cohortes (exposición controlada, incidencia aleatoria): *RA*, *FA*, *RR*, *OR*
- Caso-control (exposición aleatoria, incidencia controlada): *OR*

Otras medidas relacionadas

Riesgo relativo suavizado



Tests diagnóstico

Resultado del test

Enfermedad (2×2)	+	-	Total
Sí	120	80	200
No	90	710	800
Total	210	790	1000

Tabla 2×2. Enfermedad controlada. Se supone conocida la prevalencia $P(E)$.

Propiedades del test

- Sensibilidad: $P(+|\mathtt{E})$. $\mathtt{E} \rightarrow -$ Falso negativo.
- Especificidad: $P(-|\overline{\mathtt{E}})$. $\overline{\mathtt{E}} \rightarrow +$ Falso positivo.
- Valor predictivo positivo $VP+ = P(\mathtt{E}|+)$. No tabla.
- Valor predictivo negativo $VP- = P(\overline{\mathtt{E}}|-)$. No tabla.

Sensibilidad y especificidad

Resultado del test

Enfermedad	(2 × 2)	+	-	Total
	Sí	120	80	200
	No	90	710	800
	Total	210	790	1000

$$\texttt{sens} = 120/200 = 0{,}60$$
$$\texttt{esp} = 710/800 = 0{,}89$$

Test ideal: sens=1; esp=1

Resultado del test

Enfermedad	(2×2)	+	-	Total
	Sí	200	0	200
	No	0	8000	800
	Total	200	800	1000

$$\texttt{sens} = 200/200 = 1$$
$$\texttt{esp} = 800/800 = 1$$

VP+ y VP-

No pueden estimarse directamente a partir de la tabla porque, normalmente, la cantidad de enfermos que recoge la tabla es muy superior a la que cabría esperar en función de la prevalencia de la enfermedad. No obstante, si la prevalencia es conocida, podemos hacer uso de la Regla de Bayes para obtener una estimación válida de VP+ y VP-.

VP+ y VP- según Regla de Bayes

$$VP+ = \frac{\texttt{sens} \times \texttt{prev}}{\texttt{sens} \times \texttt{prev} + (1 - \texttt{esp}) \times (1-\texttt{prev})}$$

$$VP- = \frac{\texttt{esp} \times (1 - \texttt{prev})}{(1 - \texttt{sens}) \times \texttt{prev} + \texttt{esp} \times (1-\texttt{prev})}$$

$VP+$, $VP-$

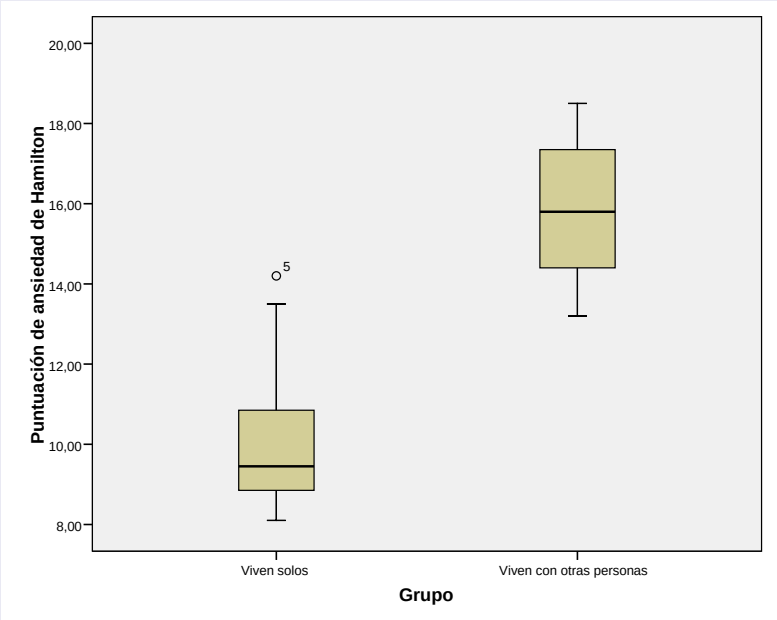
Dato conocido: `prev` $= 0{,}02$

$$VP+ = \frac{0{,}60 \times 0{,}02}{0{,}60 \times 0{,}02 + 0{,}113 \times 0{,}98} = 0{,}097$$

$$VP- = \frac{0{,}887 \times 0{,}98}{0{,}40 \times 0{,}02 + 0{,}887 \times 0{,}98} = 0{,}990$$

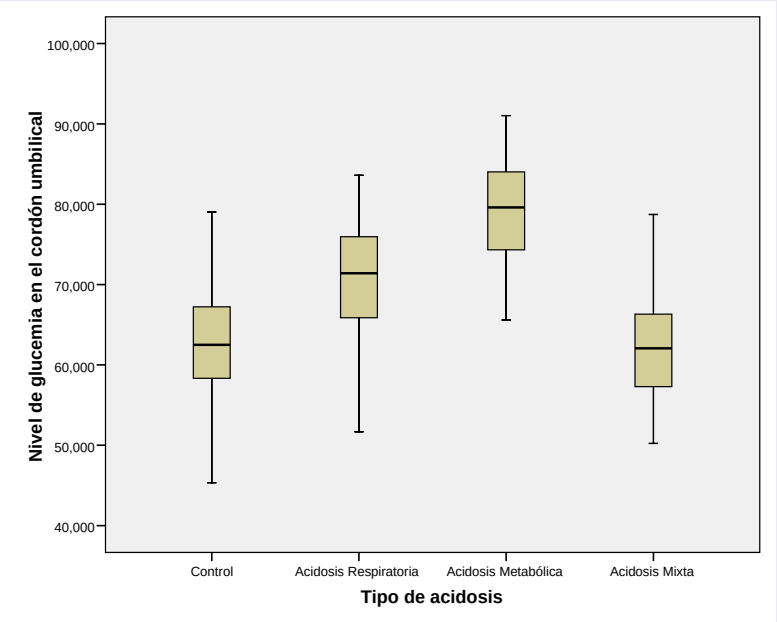
Introducción a las comparaciones de medias

Factor→ Cuantitativa



¿Influye el estilo de vida en la ansiedad?

¿Influye la acidosis en la glucemia?



Contrastes de hipótesis

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

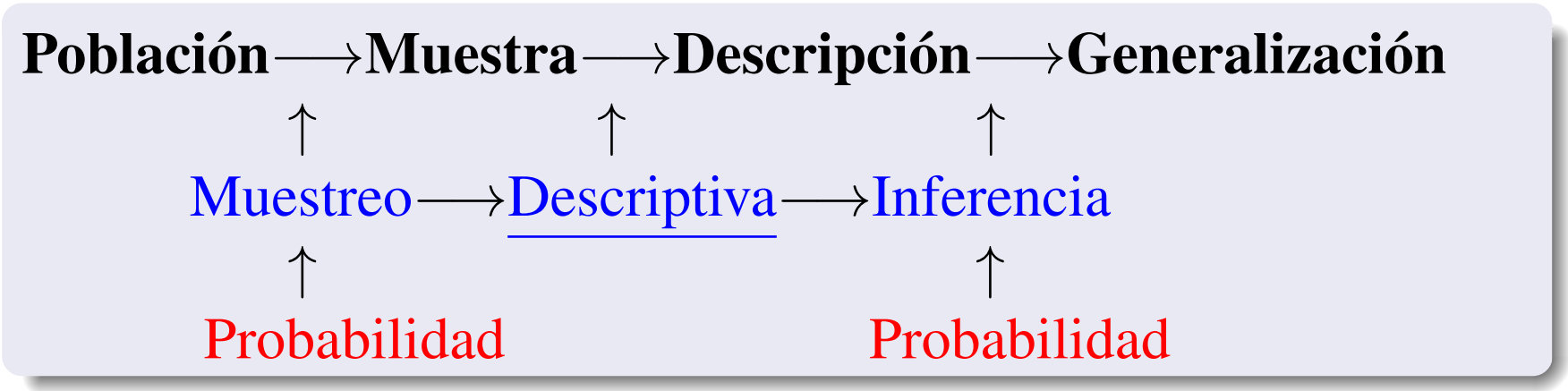
Parámetros poblacionales

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.

Proceso estadístico



Diseño de experimentos

- Las muestras deben seleccionarse aleatoriamente controlando el factor o los factores a estudiar.
- A partir de los valores típicos muestrales se efectúan una serie de cálculos de carácter probabilístico (test de hipótesis) que conducen a una decisión.