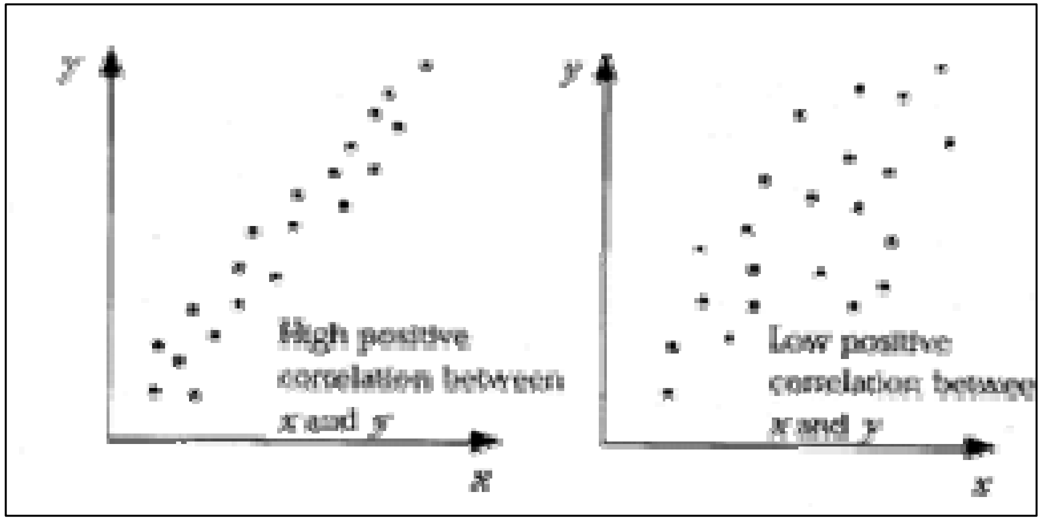
Tema VI. El parecido entre variables

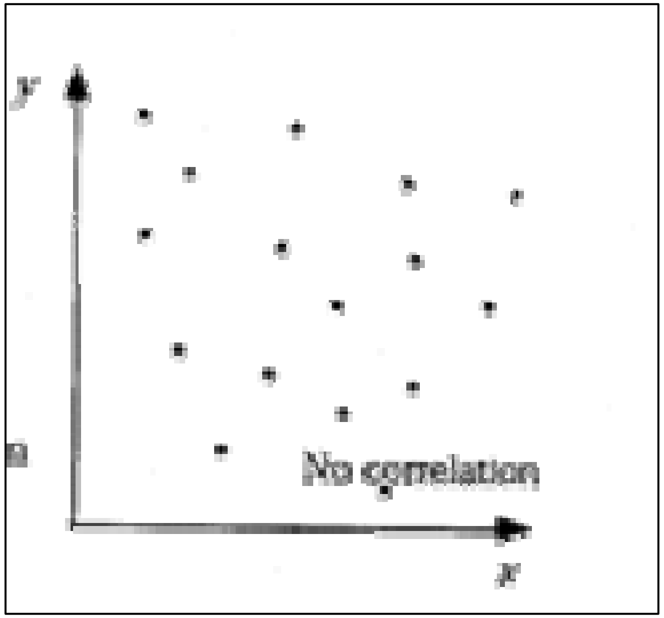
6.1. Las medidas de parecido/asociación:

- La relación visual entre dos variables (aleatorias?):

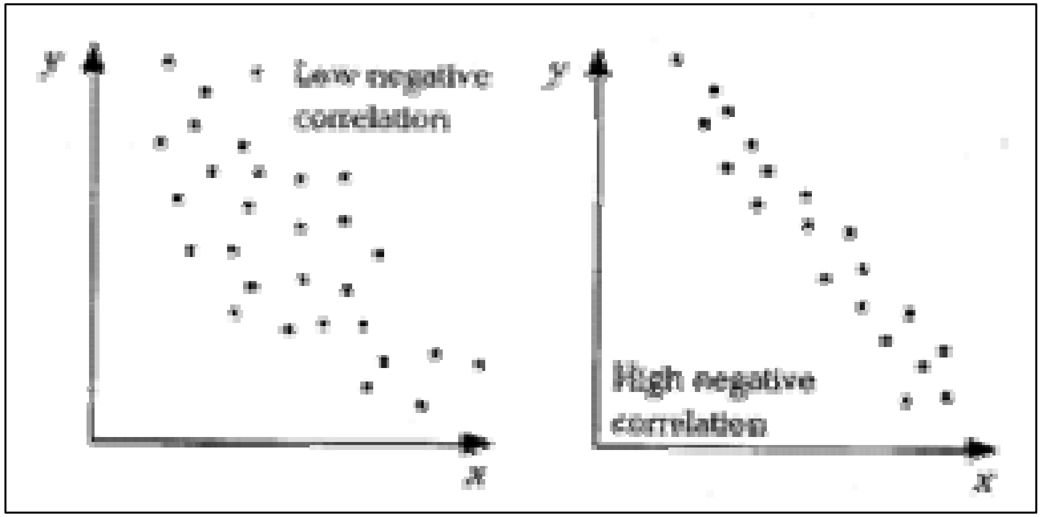
Asociación +



SIN asociación



Asociación -



Tema VI. El parecido entre variables

6.1. Las medidas de parecido/asociación:

- La medida del parecido:
  - . La covarianza entre dos variables x,y (COV_{XY})
  - . La covarianza consigo misma
  - . El problema de las comparaciones

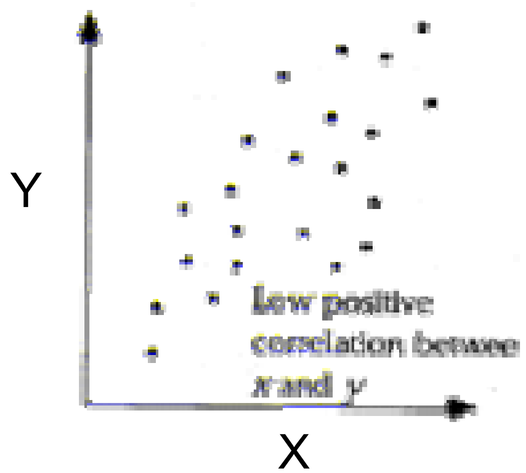
$$C(xy) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \qquad -\infty \leq C(xy) \leq +\infty$$

Covariance & Variance

$$C(xx) = V(x)$$

$$C(xx) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)$$

$$C(xx) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = V(x)$$



Ejemplo 1:

4	12
5	13
6	11
7	15
8	23

COV = 4,8

Ejemplo 2 (E1/10):

0,4	1,2
0,5	1,3
0,6	1,1
0,7	1,5
0,8	2,3

COV = 0,048

Tema VI. El parecido entre variables

6.1. Las medidas de parecido/asociación:

- La covarianza de una muestra (n-1)
- La estandarización de la covarianza: el coeficiente de correlación (r)

Ejemplo 1:

4	12
5	13
6	11
7	15
8	23

$\sum x = 30$
$\sum y = 74$
$\sum x^2 = 190$
$\sum y^2 = 1188$
$\sum xy = 468$

$$r_{XY} = \frac{COVxy}{\sigma x \sigma y} = \frac{\left(\frac{\sum(x_i - x)(y_i - y)}{n-1}\right)}{\left(\frac{\sqrt{(x_i - x)^2 (y_i - y)^2}}{\sqrt{(n-1) \times (n-1)}}\right)}$$

COV = 4,8
SD_1 = 1,414
SD_2 = 4,308
r = 0,788

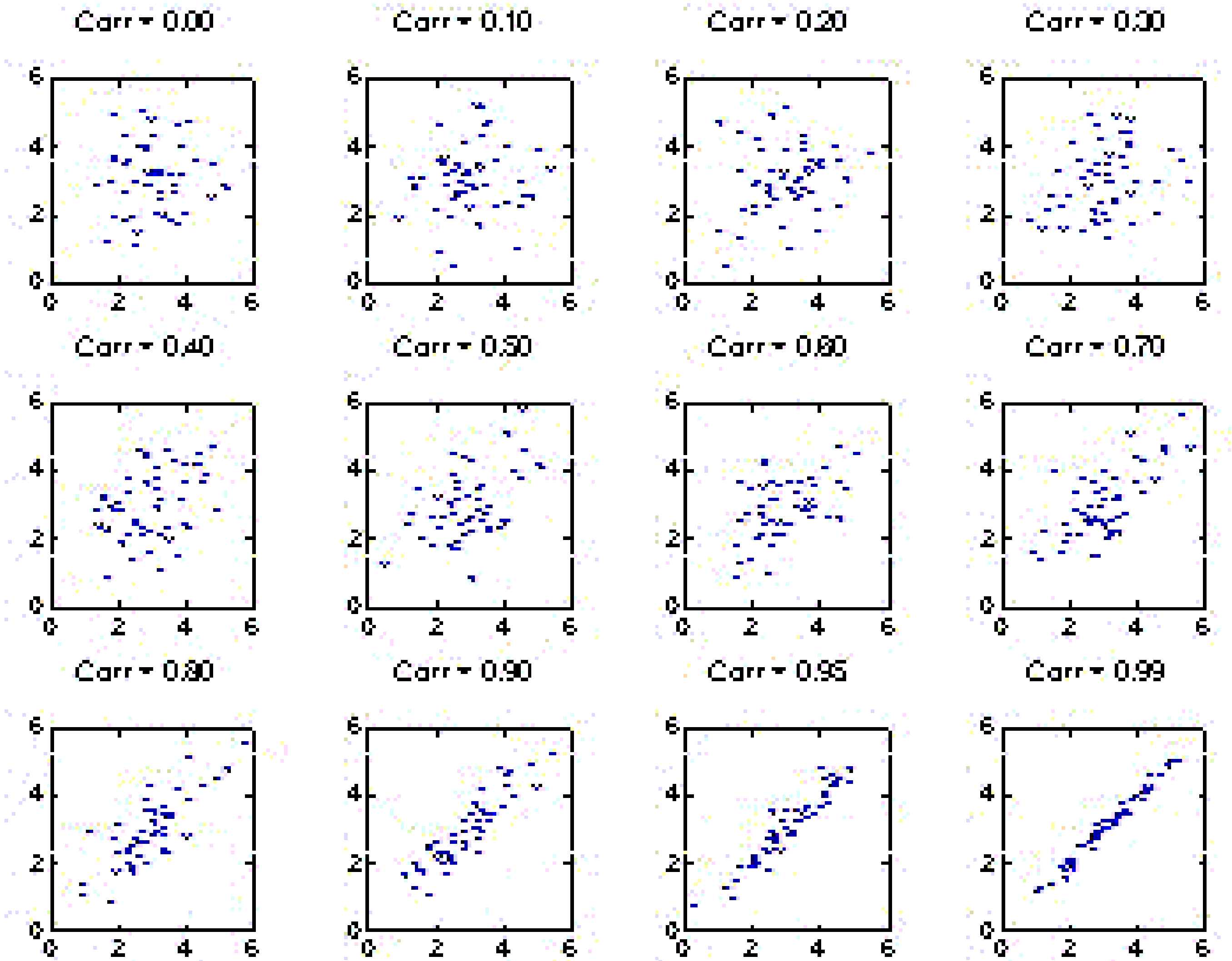
Fórmula simplificada

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{5\,(468) - (30 \times 74)}{\sqrt{((5 \times 190) - 30^2) \times ((5 \times 1188) - 74^2)}} = 0{,}788$$

6.1. Las medidas de parecido/asociación:

- El coeficiente de correlación:
  . Ejemplos



Tema VI. El parecido entre variables

6.2. El test de significación del coeficiente de correlación:

- La distribución muestral de r, y su varianza muestral
- El test t clásico (cuando r es moderado o bajo?)
- Para valores de r muy cercanos a 1 se puede uar la transformación Z (cuando n > 50!)

Varianza muestral

$$s^2_r = \frac{1 - r^2}{n - 2}$$

Test t clásico

$$t_r = \frac{r - 0}{\sqrt{(1-r^2)/(n-2)}} = r\sqrt{(n-2)/(1-r^2)}$$

Estadístico

$$z = ½ \ln \left((1+r)/1-r\right)$$

Varianza muestral

$$\sigma^2_z = \frac{1}{(n - 3)}$$

Test t clásico

$$t_z = \frac{z - 0}{1/\sqrt{(n - 3)}} = Z\sqrt{(n - 3)}$$

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$r = 0,56$

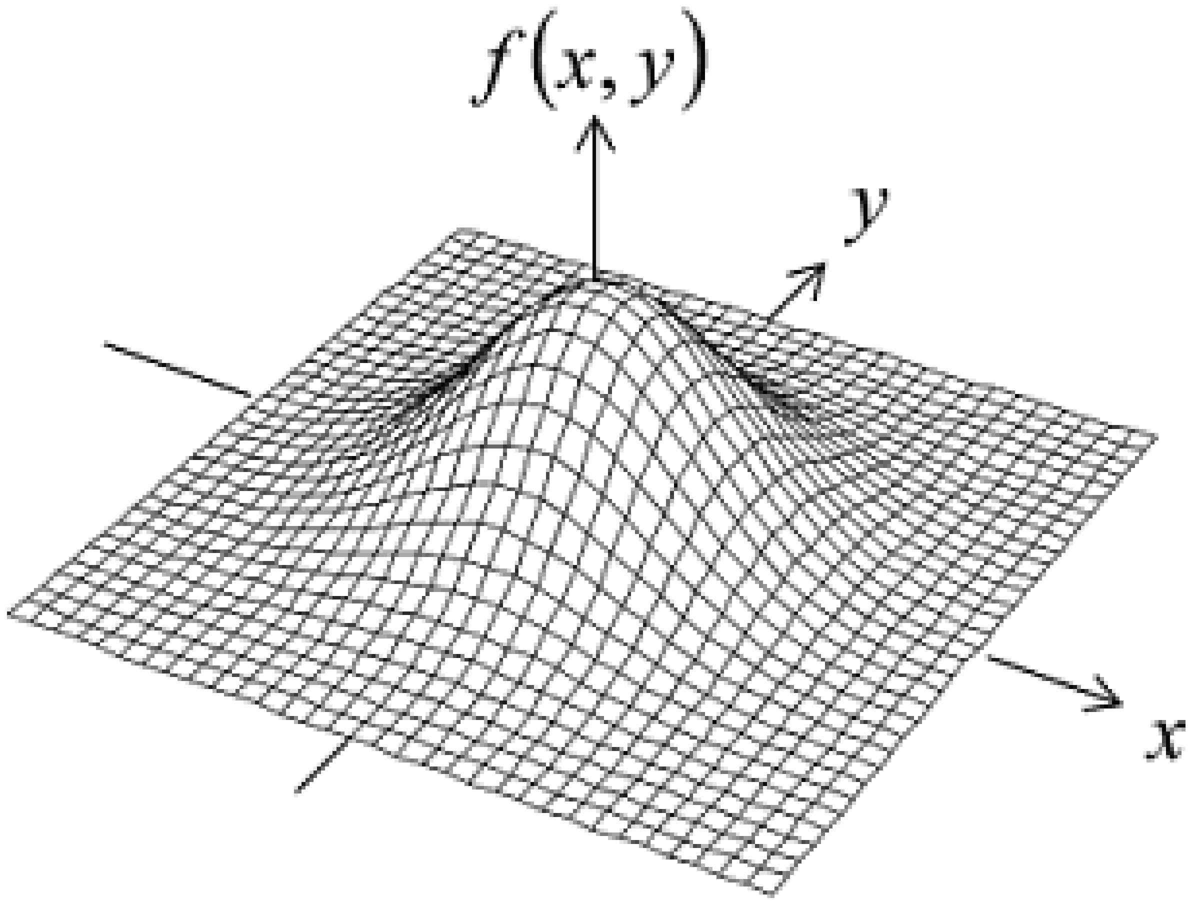
$t_r = 0,56 \sqrt{11,65} = 1,91$

Repetir para n = 50

$z = 0,5 \ln (3,545) = 0,63$
$t_r = 0,63 \sqrt{(7)} = 0,24$

6.3. Las asunciones del modelo de correlación :

- Se usa con dos variables aleatorias
- Las dos variables siguen una distribución Normal Bivariada (al menos normal)
- Si no se cumplen; test de asociación no paramétrico



6.4. Los test de asociación NO PARAMÉTRICOS:

- No asumen ninguna distribución

Enfermos : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
Sanos : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

Enfermos : 10, 1, 2, 8, 6,5, 3, 5, 9, 6,5, 4
Sanos : 6,5, 1, 5, 9, 3,5, 9, 6,5, 9, 2, 3,5

Acuerdos: +, +, -, -, +, -, -, +, +, +
$n_C = 6$
$n_D = 4$
$S = n_C + n_D$
$\sum t_i(t_i-1)/2 = 2(2-1)/2 = 1$
$\sum u_i(u_i-1)/2 = (2(2-1) + 3(3-1) + (2(2-1))/2 = 5$

$$\tau = \frac{2}{\sqrt{((10x9)/2) - 1) \text{ x } ((10x9)/2) - 5)}} = 0{,}05$$

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i(t_i-1)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i(u_i-1)/2\right]}}$$

6.4. Los test de asociación NO PARAMÉTRICOS:

- Evaluación de probabilidad:
  - . > 10 aproximación normal: varianza muestral y test Z
  - . < 10 Tablas de permutación

Enfermos : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
Sanos : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$$\tau = \frac{2}{\sqrt{((10x9)/2) - 1) \times ((10x9)/2) - 5)}} = 0,05$$

$$\sigma^2_T = \frac{2(2N + 5)}{9N(N-1)} = 0,062$$

$$z = \frac{\tau - \mu_T(0)}{\sigma_T} = 0,20$$

No se rechaza H_0

TABLE R₁
Upper-tail probabilities for T, the Kendall rank-order correlation coefficient ($N \leq 10$)*

Entries are $p = P[T \geq \text{tabled value}]$.

N	T	p	N	T	p	N	T	p	N	T	p
4	.000	.625	7	.048	.500	9	.000	.540	10	.022	.500
	.333	.375		.143	.386		.056	.460		.067	.431
	.667	.167		.238	.281		.111	.381		.111	.364
	1.000	.042		.333	.191		.167	.306		.156	.300
				.429	.119		.222	.238		.200	.242
5	.000	.592		.524	.068		.278	.179		.244	.190
	.200	.408		.619	.035		.333	.130		.289	.146
	.400	.242		.714	.015		.389	.090		.333	.108
	.600	.117		.810	.005		.444	.060		.378	.078
	.800	.042		.905	.001		.500	.038		.422	.054
	1.000	.008		1.000	.000		.556	.022		.467	.036
							.611	.012		.511	.023
6	.067	.500	8	.000	.548		.667	.006		.556	.014
	.200	.360		.071	.452		.722	.003		.600	.008
	.333	.235		.143	.360		.778	.001		.644	.005
	.467	.136		.214	.274		.833	.000		.689	.002
	.600	.068		.286	.199		.889	.000		.733	.001
	.733	.028		.357	.138		.944	.000		.778	.000
	.867	.008		.429	.089		1.000	.000		.822	.000
	1.000	.001		.500	.054					.867	.000
				.571	.031					.911	.000
				.643	.016					.956	.000
				.714	.007					1.000	.000
				.786	.003						
				.857	.001						
				.929	.000						
				1.000	.000						

* Adapted and reproduced by permission of the publishers Charles Griffin & Co. Ltd., 16 Pembridge Road, London W11 3HL, from Appendix Table 5 of Kendall, M. G. (1970). *Rank correlation methods* (fourth edition).

Referencias Bibliográficas

LIBROS:

Siegel, S., Castellan, N.J.1988. Nonparametric Statistics. McGraw Hill, New York

Sokal,R.R., Rohlf, F.J. 1995. Biometry. Freeman and co., New York

Underwood, A.J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Oceanogr. Mar. Biol. Ann. Rev*., 19: 513-605.

PÁGINAS WEB:

http://www.statsdirect.com/help/statsdirect.htm#regression_and_correlation/covgr.htm
(Pagina Web donde se explica el test de Kendal tau)