ITBA

Instituto Tecnológico de Buenos Aires

28/AGOSTO

MEDIDAS DE CORRELACIÓN

Correlación

Sentido

creciente o positiva / decreciente o negativa

Forma

lineal / no lineal

Fuerza

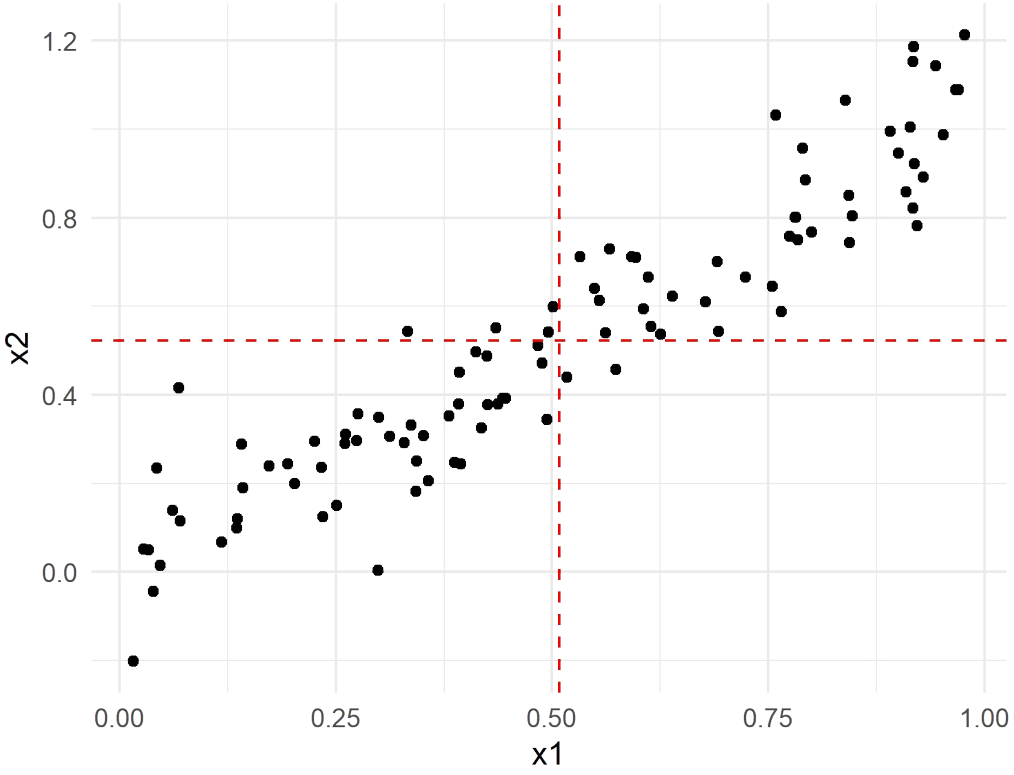
(dispersión en relación al patrón)

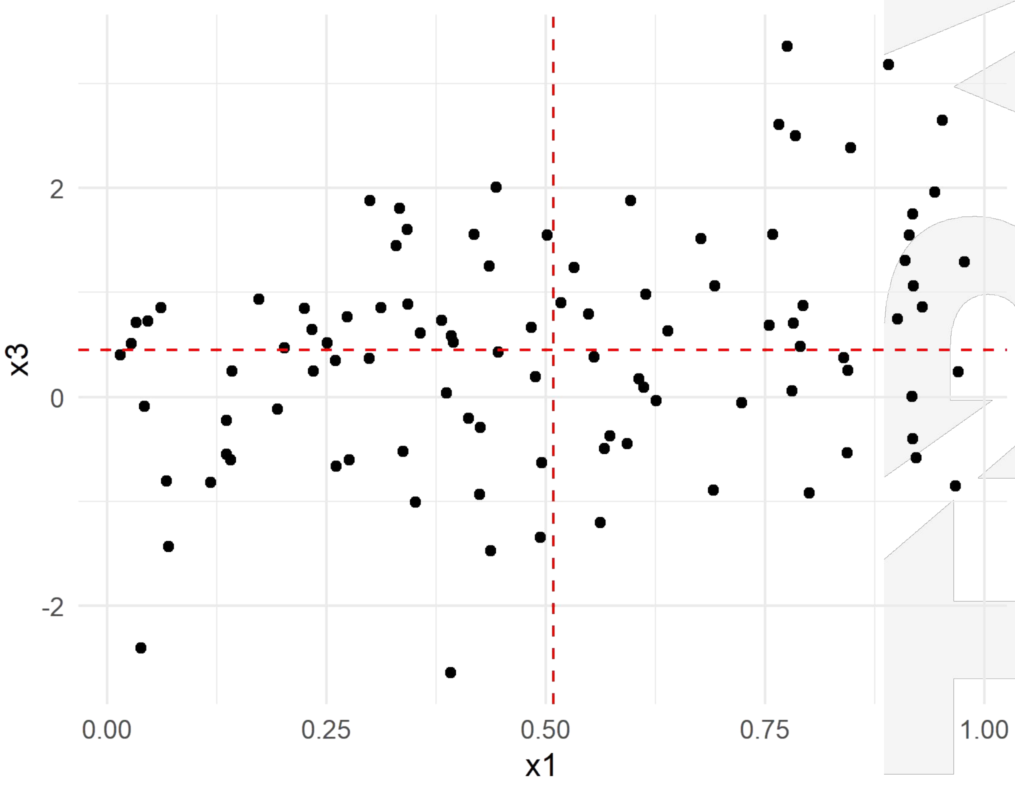
fuerte / moderada / débil

Correlación entre variables numéricas

Covarianza

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





Covarianza

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

- Captura **relaciones lineales**
- El **signo** indica el sentido

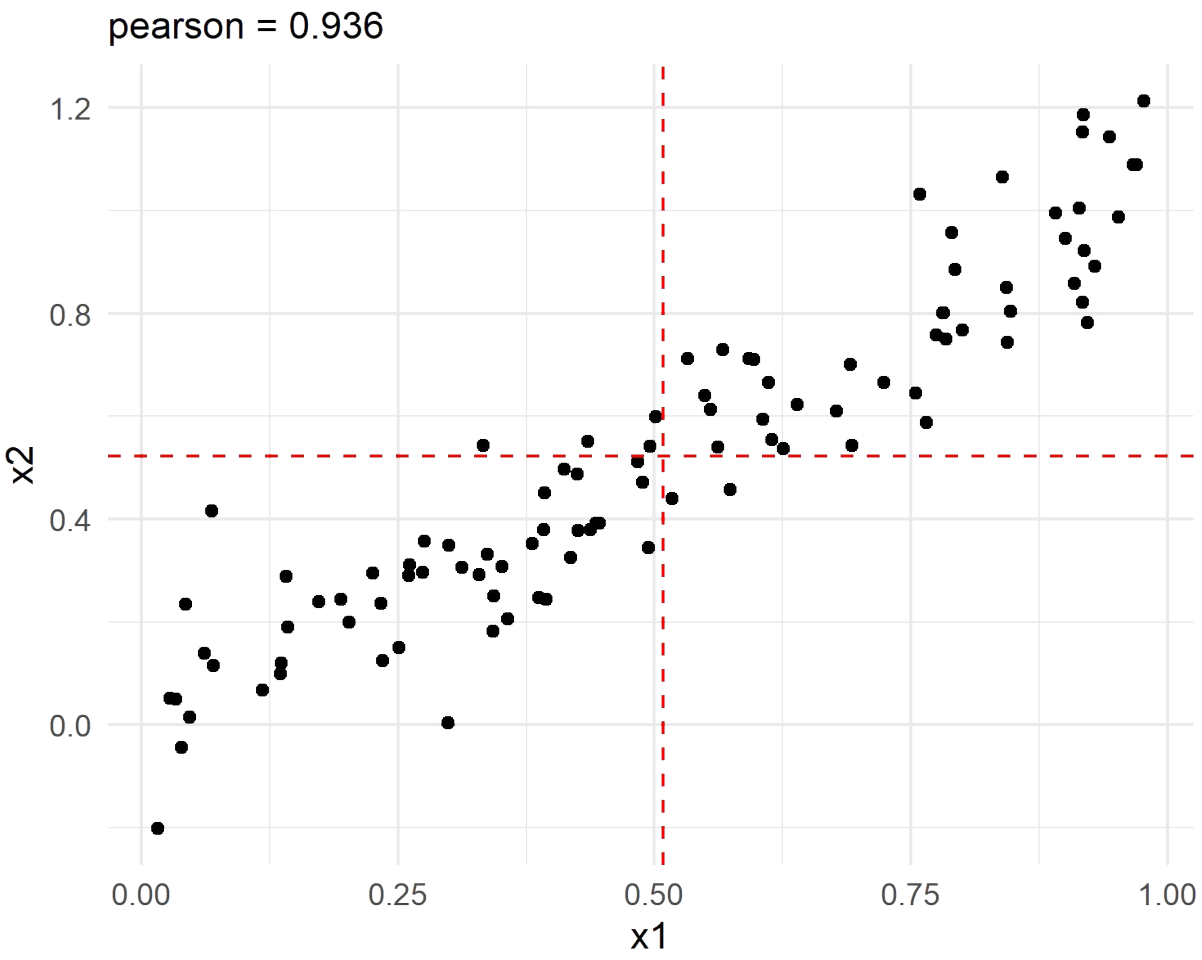
Pero no está normalizada
(depende de las unidades de medida de las variables…)

¿Cuál es la versión normalizada de la covarianza?

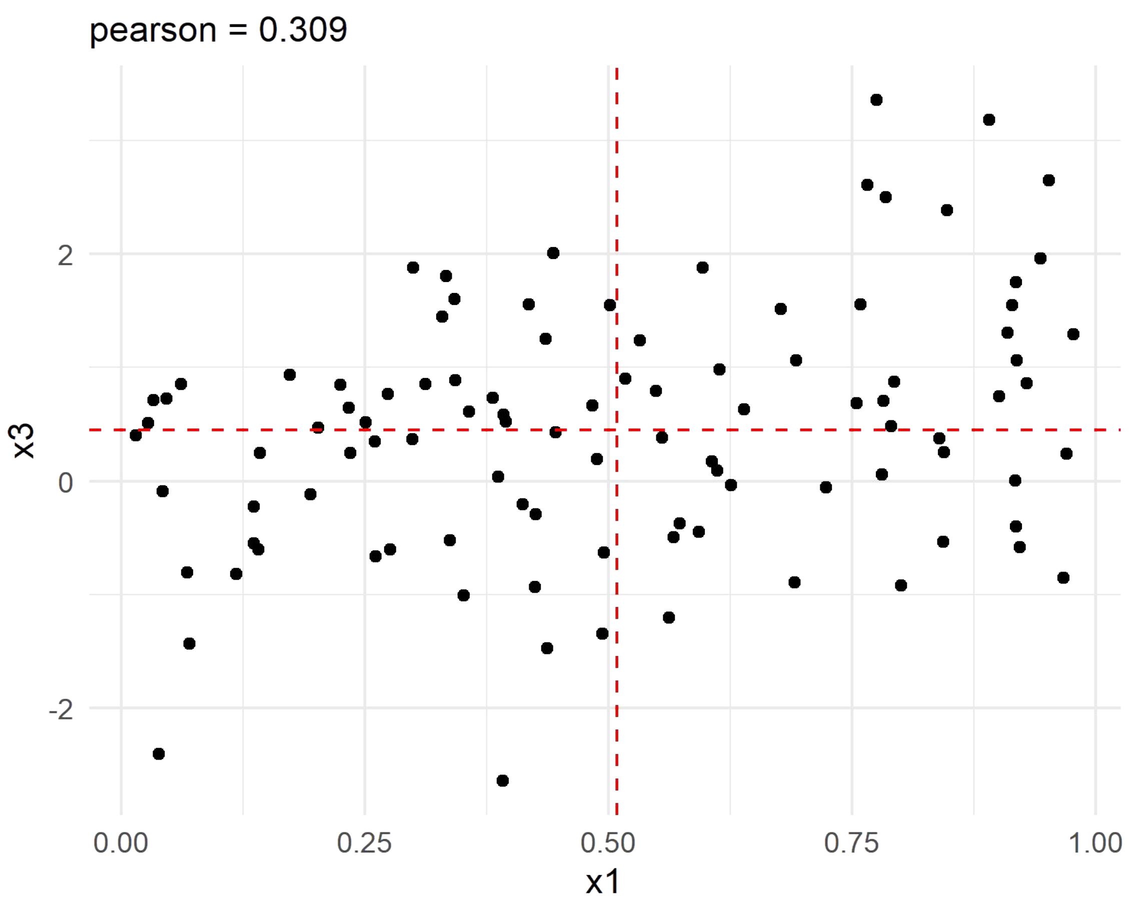
Correlación de Pearson

$$r_{x,y} = \frac{cov_{x,y}}{S_x S_y}$$

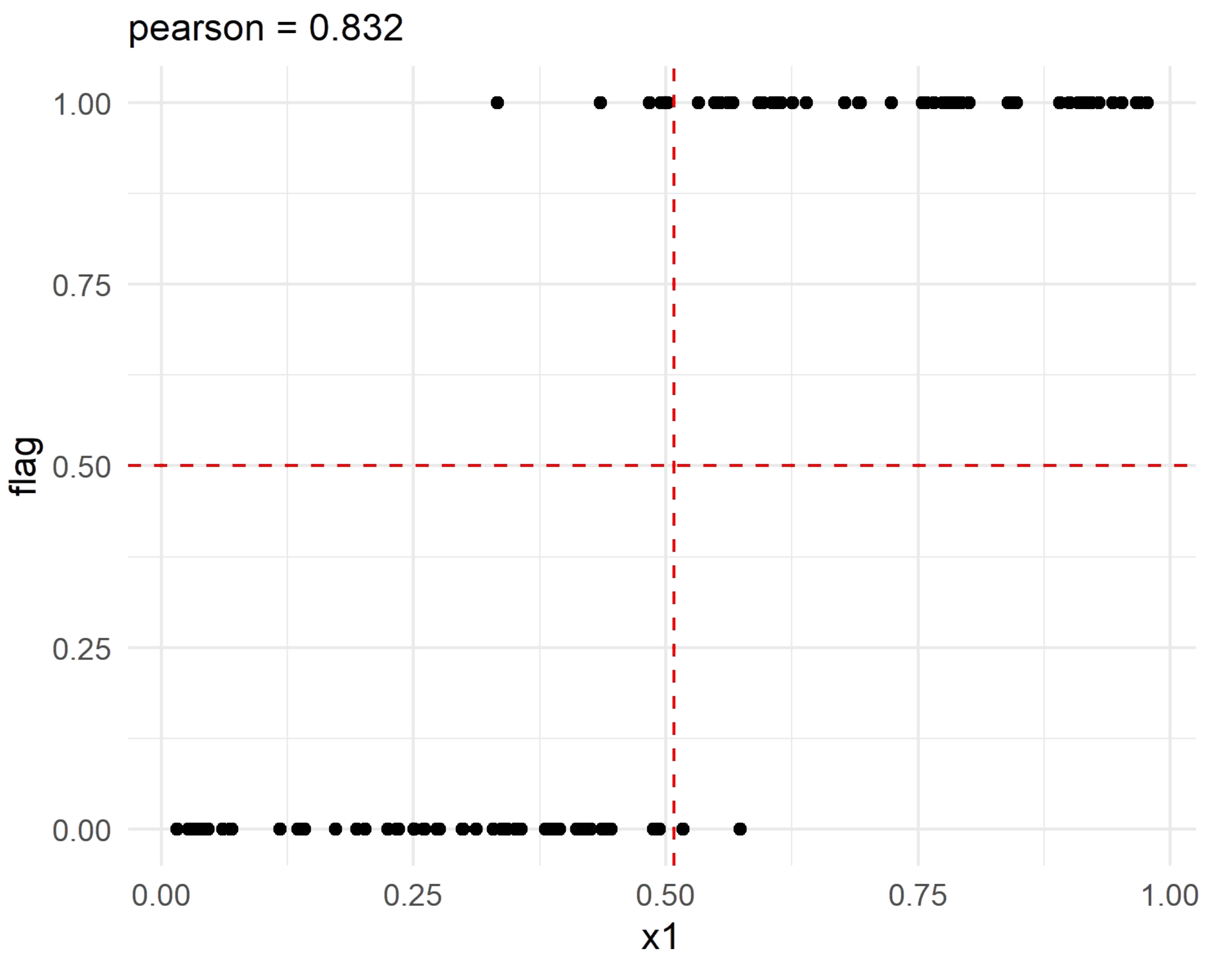
$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$



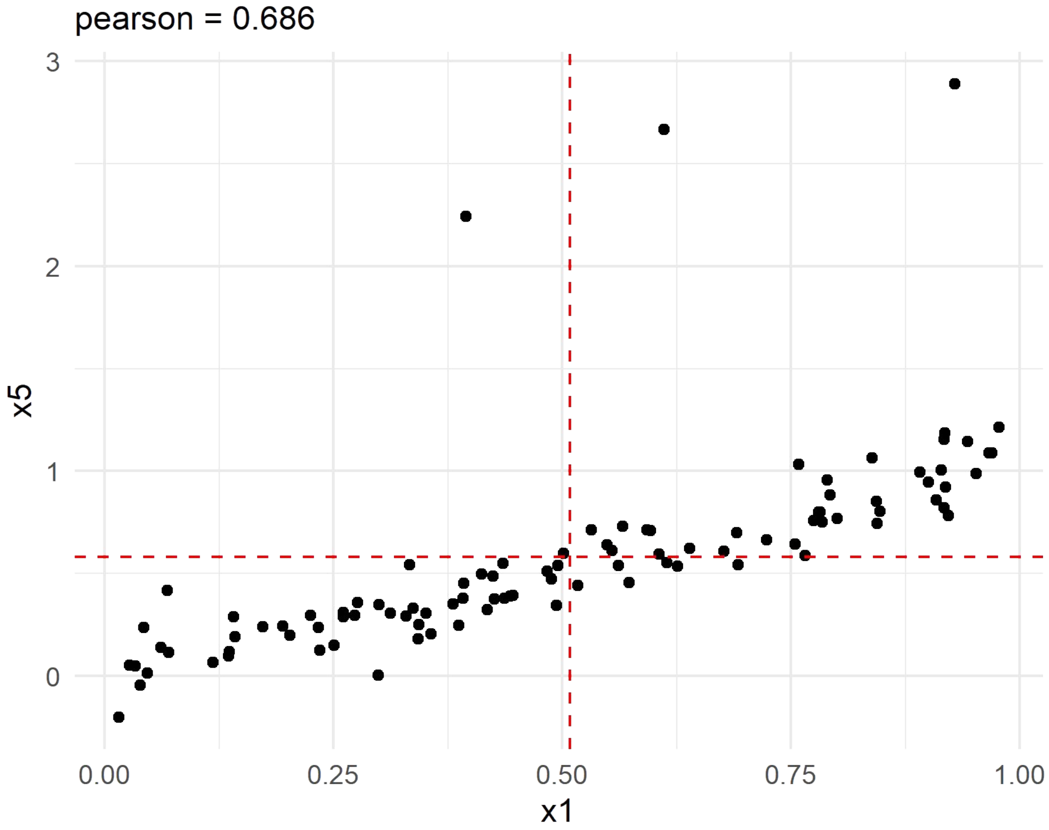
Correlación de Pearson

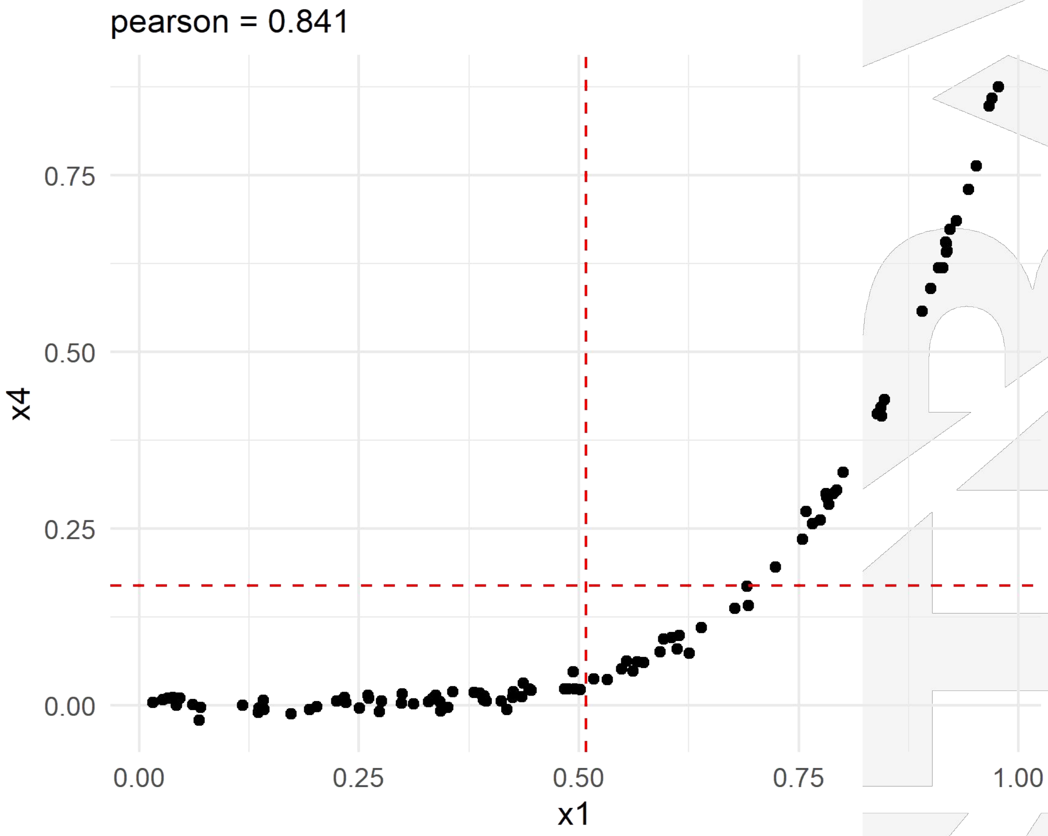


Correlación de Pearson



Correlación de Pearson





Correlación de Pearson

$$r_{x,y} = \frac{cov_{x,y}}{S_x S_y}$$

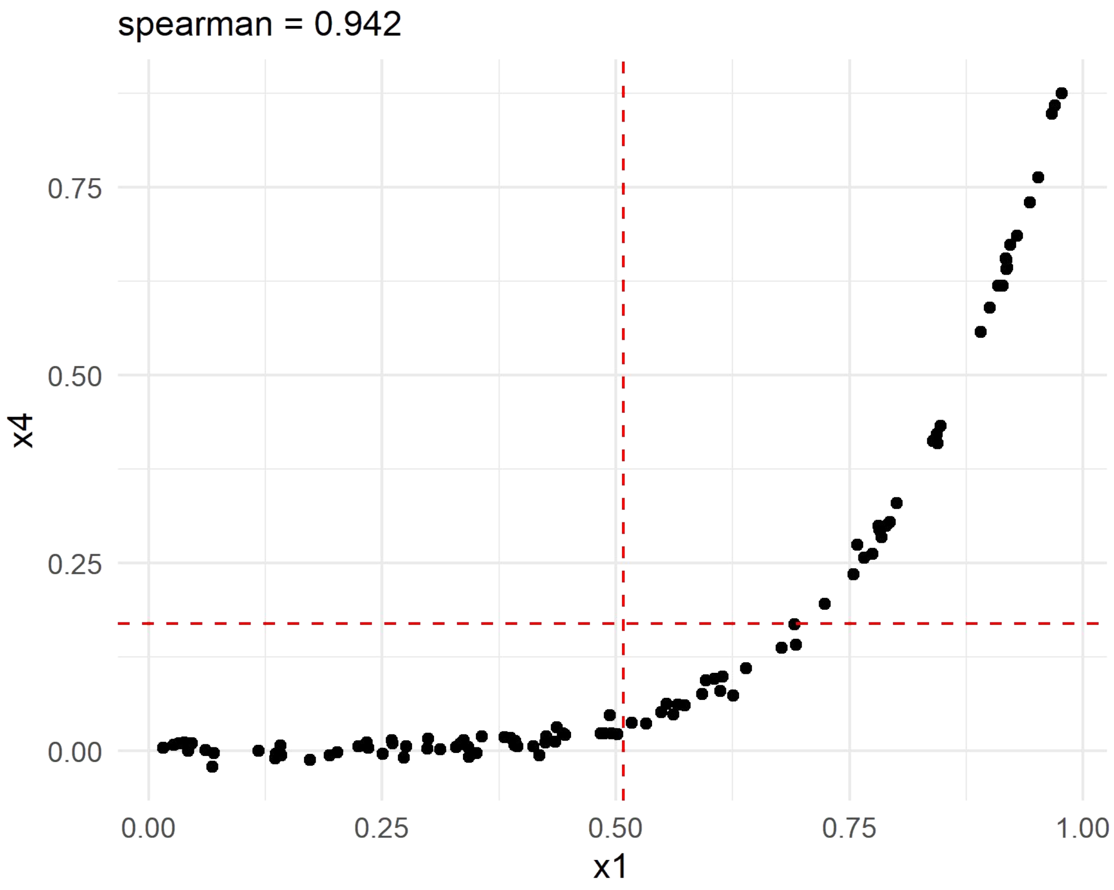
- Mide **asociación lineal**
- El **signo** indica sentido
- La **magnitud** indica la fuerza de relación lineal
- Está **normalizada** entre [-1,+1] (no depende de unidades de medida de las variables)

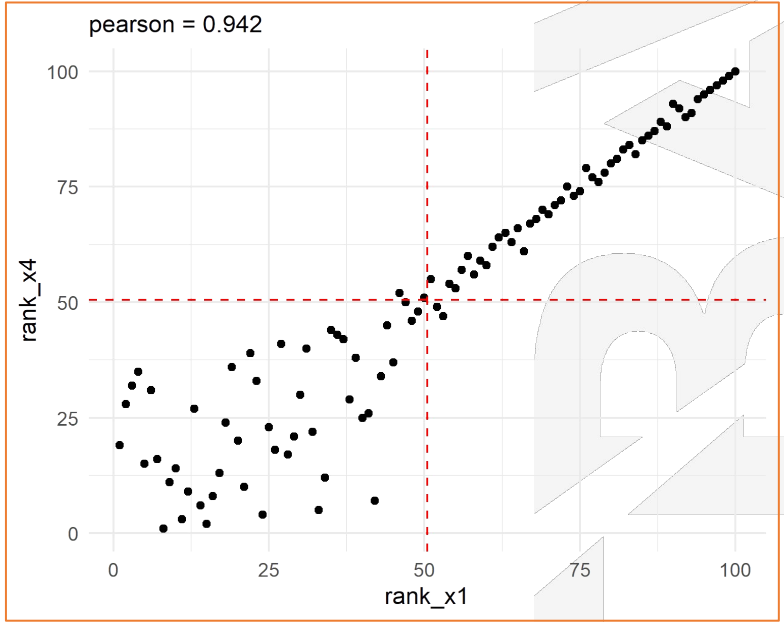
Puede medir asociaciones entre var. binaria y var. continua
Es sensible a observaciones atípicas

Correlación de Spearman

$$sp_{x,y} = r_{rank(x),rank(y)}$$

x	rank_x
46	4
39	3
13	1
53	5
36	2
89	6

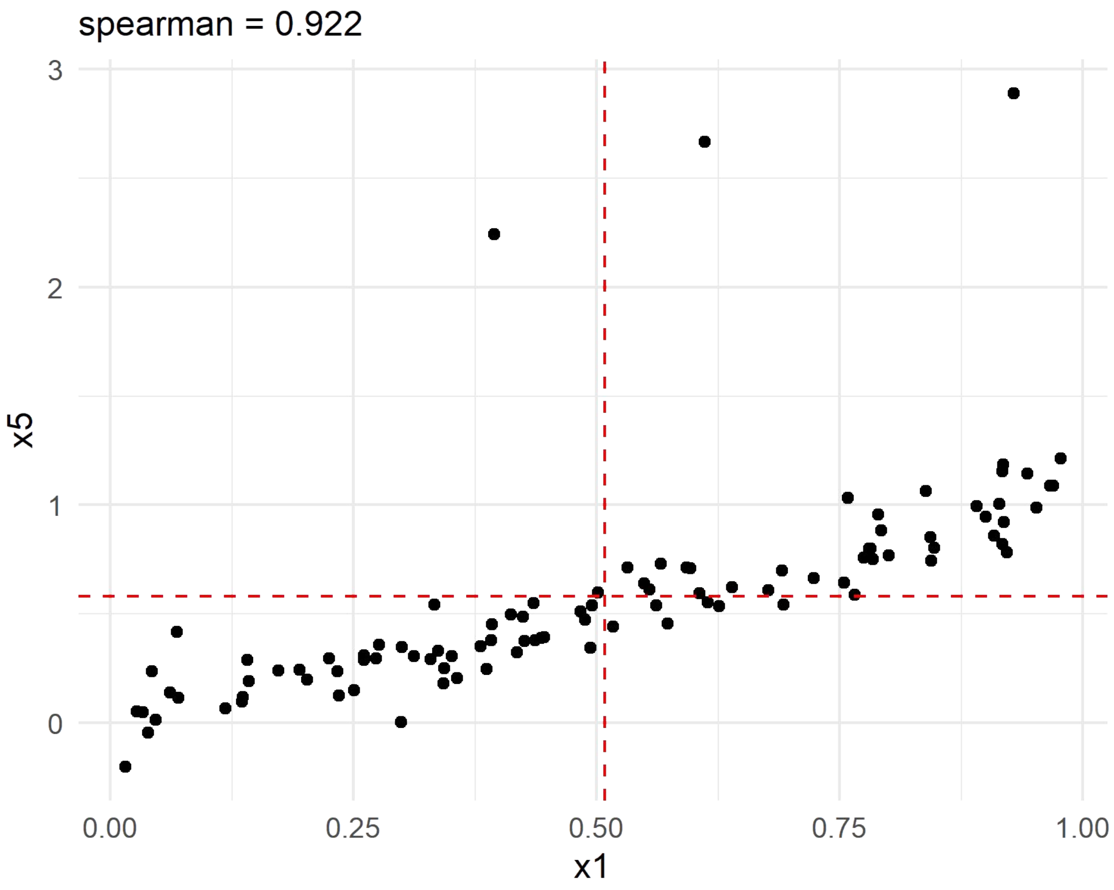


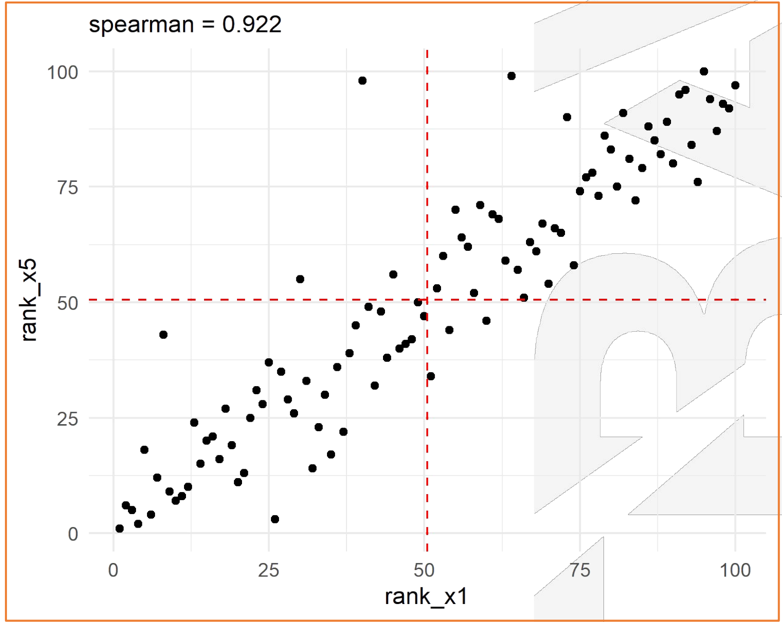


Correlación de Spearman

$$sp_{x,y} = r_{rank(x),rank(y)}$$

x	rank_x
46	4
39	3
13	1
53	5
36	2
89	6





Correlación de Spearman

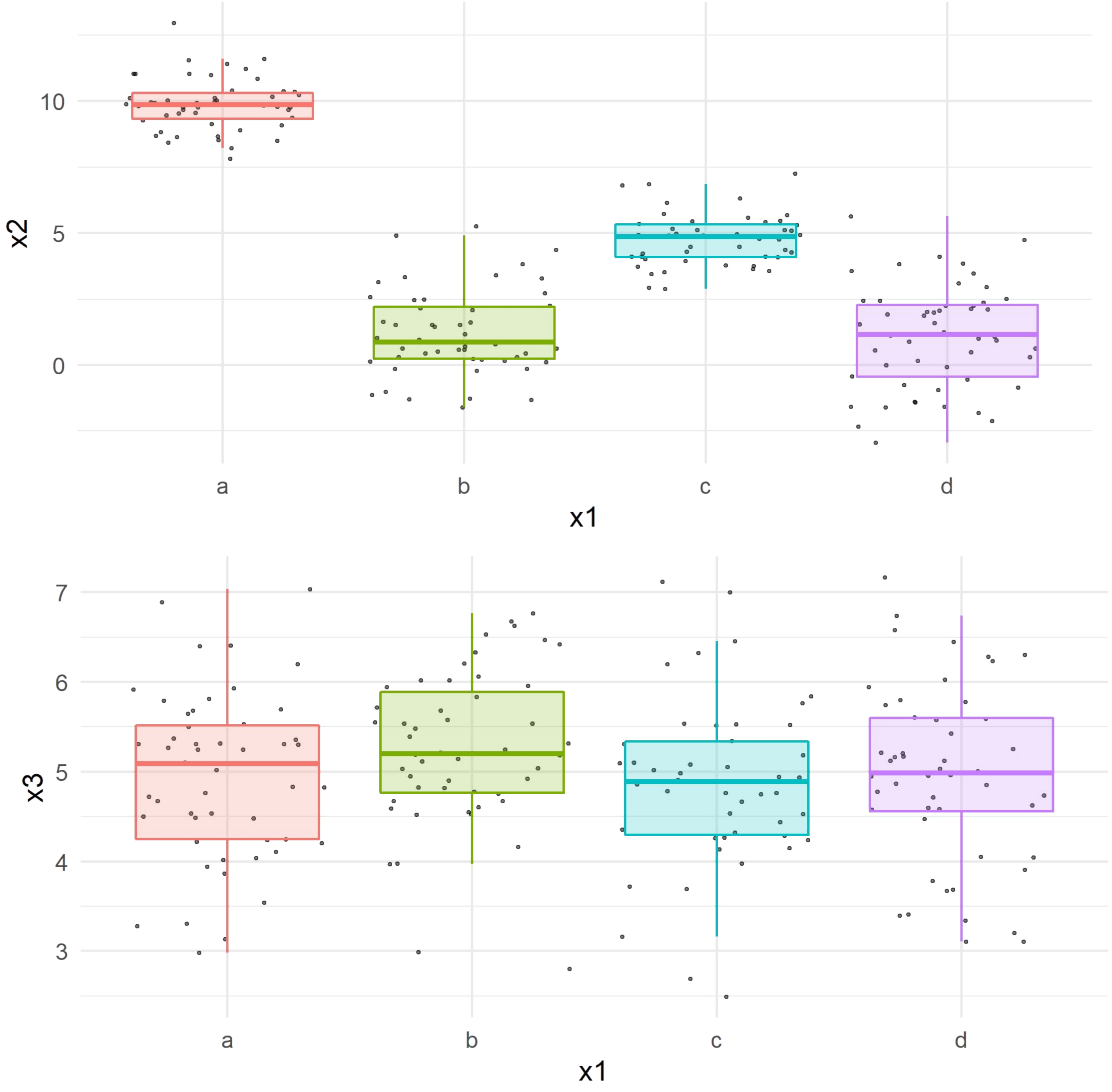
$$sp_{x,y} = r_{rank(x),rank(y)}$$

Es la correlación de Pearson del ranking de las variables.

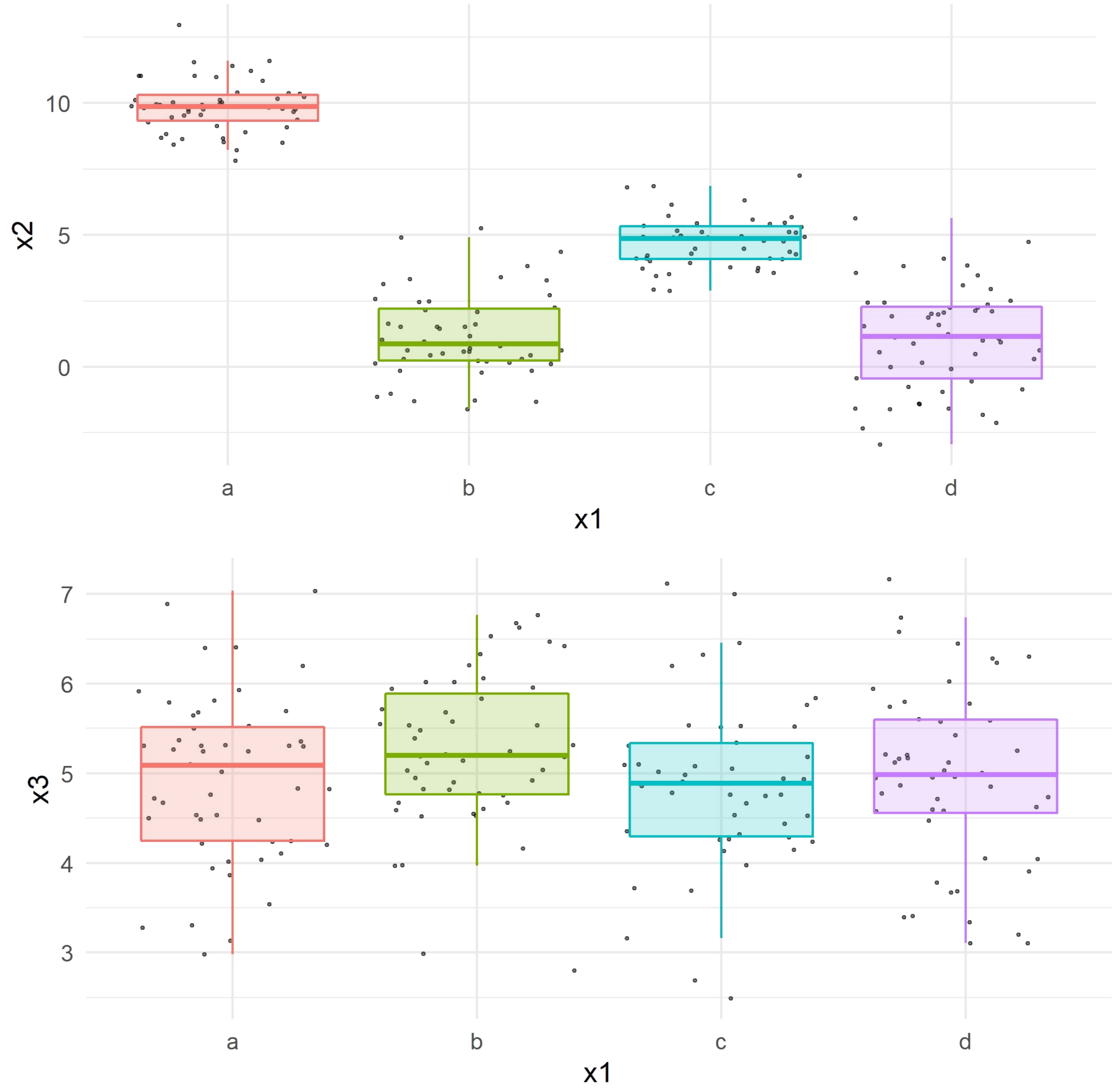
- Mide **asociaciones monótonas** (captura no linealidades)
- El **signo** indica sentido
- La **magnitud** indica la fuerza de la relación
- Está **normalizada** entre [-1,+1]

Puede medir asociaciones entre var. binaria y var. continua
Es más robusta a observaciones atípicas

Correlación entre variables numéricas y categóricas


10
5
0
x2
a
b
c
d
x1
7
6
5
4
3
x3
a
b
c
d
x1

Medidas de ANOVA



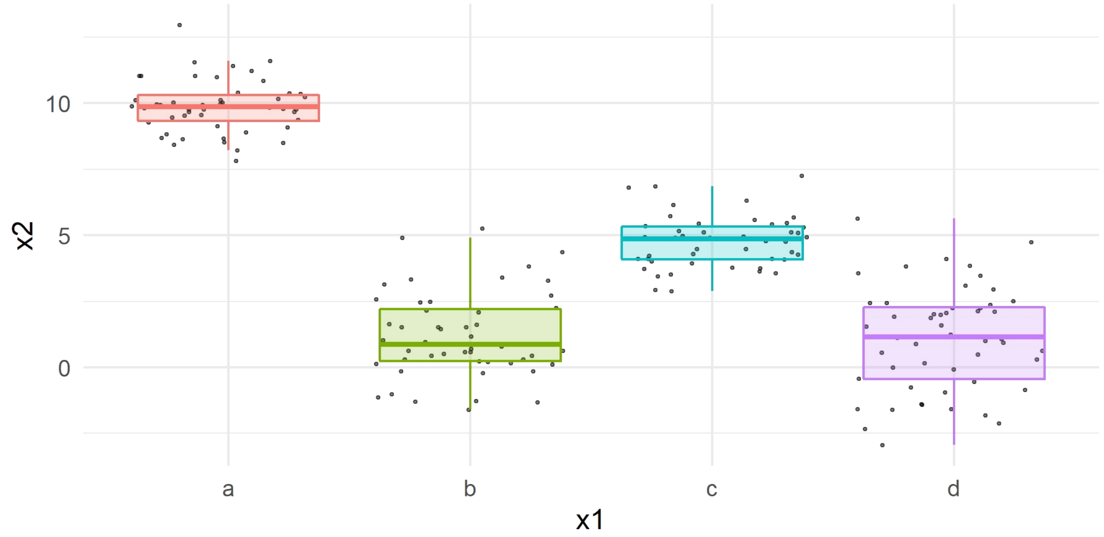
$$SS_{total} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2$$

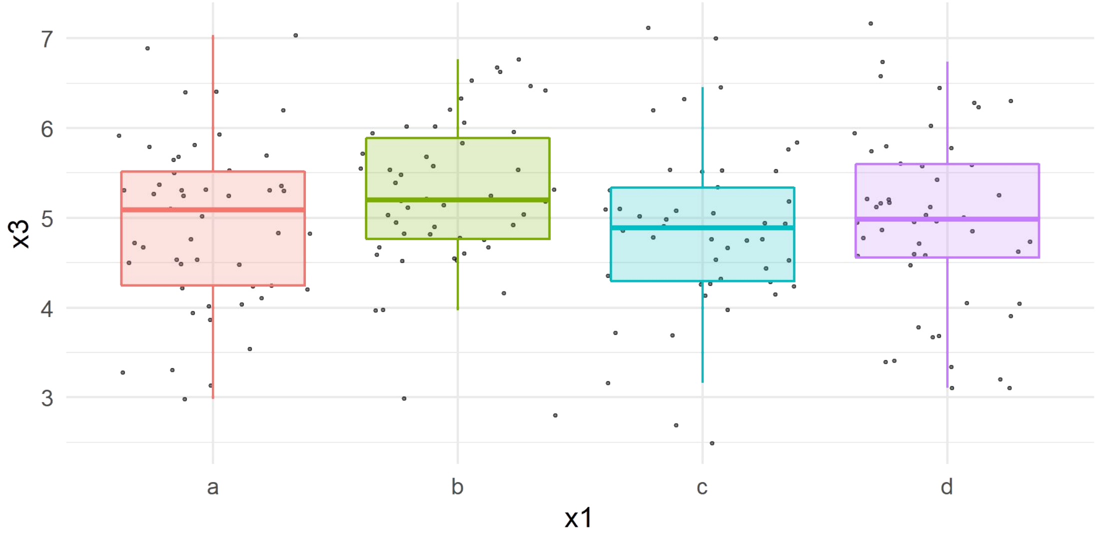
$$SS_{total} = SS_{between} + SS_{within}$$

$$SS_{between} = \sum_{j=1}^{p} n_j (\overline{x}_j - \overline{x})^2$$

$$SS_{within} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

Medidas de ANOVA



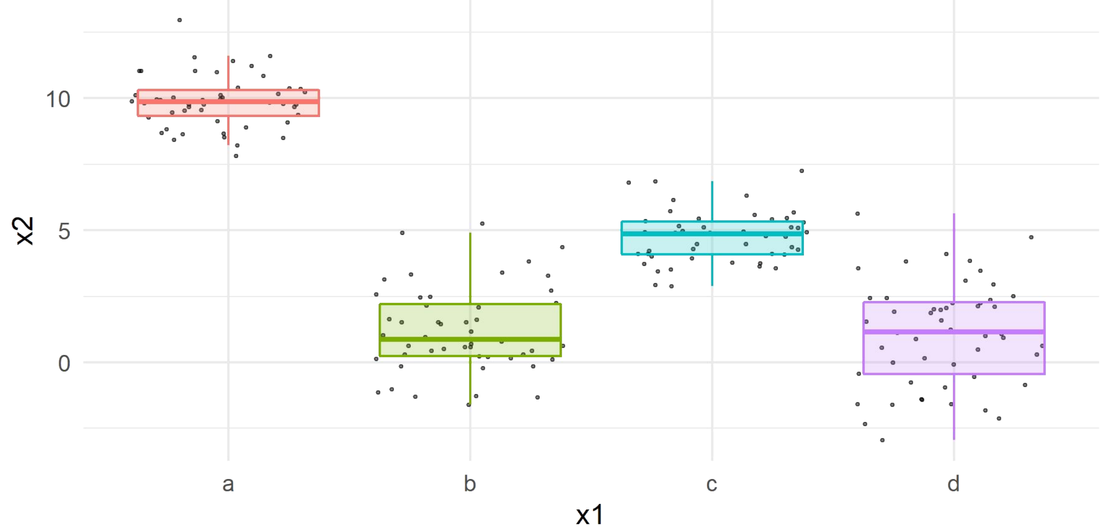


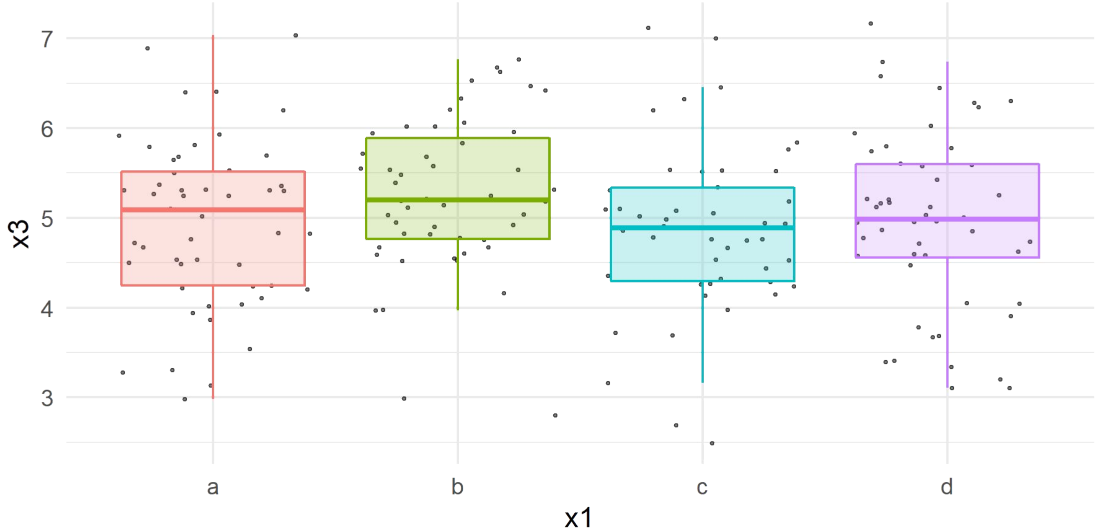
$$F = \frac{SS_{between}/(p-1)}{SS_{within}/(n-p)}$$

term	df	sumsq	meansq	statistic	p.value
x1	3	2624.9209	874.973630	410.979	0
Residuals	196	417.2837	2.128999	NA	NA

term	df	sumsq	meansq	statistic	p.value
x1	3	4.423538	1.4745128	1.695823	0.169229
Residuals	196	170.421436	0.8694971	NA	NA

Medidas de ANOVA





$$\omega^2 = \frac{SS_B - (p-1)(\frac{SS_W}{n-p})}{SS_T + \frac{SS_B}{p-1}}$$

omega-squared

$$\omega^2 = 0.86$$

$$\omega^2 = 0.01$$

Sobre los rangos de valores: https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize

ANOVA

$$\omega^2 = \frac{SS_B - (p-1)(\frac{SS_W}{n-p})}{SS_T + \frac{SS_B}{p-1}}$$

omega-squared

ANOVA generaliza el *test t* más allá de dos medias:

- H. nula: todas las medias son iguales

- H. alt.: al menos una media difiere

Podemos usar un **effect size** del test para medir la fuerza de la **asociación entre una variable categórica y una continua**.

Los p-valores nos hablan de la probabilidad de que exista un efecto.
¿Pero cuán relevante es ese efecto si existe?
$\longrightarrow$ El **effect size** es una **medida normalizada de la magnitud del efecto**

Medidas de ANOVA

Atención

Si las poblaciones se alejan mucho de la **normalidad** y hay **heterocedasticidad**:

(a) Transformar con **log(x)**, o bien

(b) Usar **Kruskal-Wallis + epsilon-squared** (effect size)

Sobre los rangos de valores:
los mismos que ANOVA

Correlación entre variables categóricas

region/equipo	Boca	River	Otros	Total
Norte	11	150	15	176
Sur	190	16	18	224
Total	201	166	33	400

region/equipo	Boca	River	Otros	Total
Norte	6.2% (11)	85.2% (150)	8.5% (15)	100.0% (176)
Sur	84.8% (190)	7.1% (16)	8.0% (18)	100.0% (224)

region/equipo	Boca	River	Otros
Norte	5.5% (11)	90.4% (150)	45.5% (15)
Sur	94.5% (190)	9.6% (16)	54.5% (18)
Total	100.0% (201)	100.0% (166)	100.0% (33)

V de Cramér

region/equipo	Boca	River	Otros	Total
Norte	11	150	15	176
Sur	190	16	18	224
Total	201	166	33	400

region/equipo	Boca	River	Otros	Total
Norte	88.4	73	14.5	175.9
Sur	112.6	93	18.5	224.1
Total	201.0	166	33.0	400.0

test chi-cuadrado

$$\mathrm{P}(x,y) = \mathrm{P}(x)\mathrm{P}(y)$$

independencia

$$p_{\cdot j} = \frac{O_{\cdot j}}{N} = \sum_{i=1}^{r} \frac{O_{i,j}}{N}$$

$$p_{i \cdot} = \frac{O_{i \cdot}}{N} = \sum_{j=1}^{c} \frac{O_{i,j}}{N},$$

frecuencias esperadas bajo independencia

$$E_{i,j} = N p_{i\cdot} p_{\cdot j},$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Fuente: https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

V de Cramér

$$V = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}}$$

region/equipo	Boca	River	Otros	Total
Norte	6.2% (11)	85.2% (150)	8.5% (15)	100.0% (176)
Sur	84.8% (190)	7.1% (16)	8.0% (18)	100.0% (224)

region/equipo	Boca	River	Otros
Norte	5.5% (11)	90.4% (150)	45.5% (15)
Sur	94.5% (190)	9.6% (16)	54.5% (18)
Total	100.0% (201)	100.0% (166)	100.0% (33)

$$\mathrm{V} = 0.82$$

V de Cramér

$$V = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}}$$

El test chi-cuadrado mide independencia de atributos:

- *H. nula: atributos independientes ("igualdad de perfiles")*
- *H. alt: lo contrario*

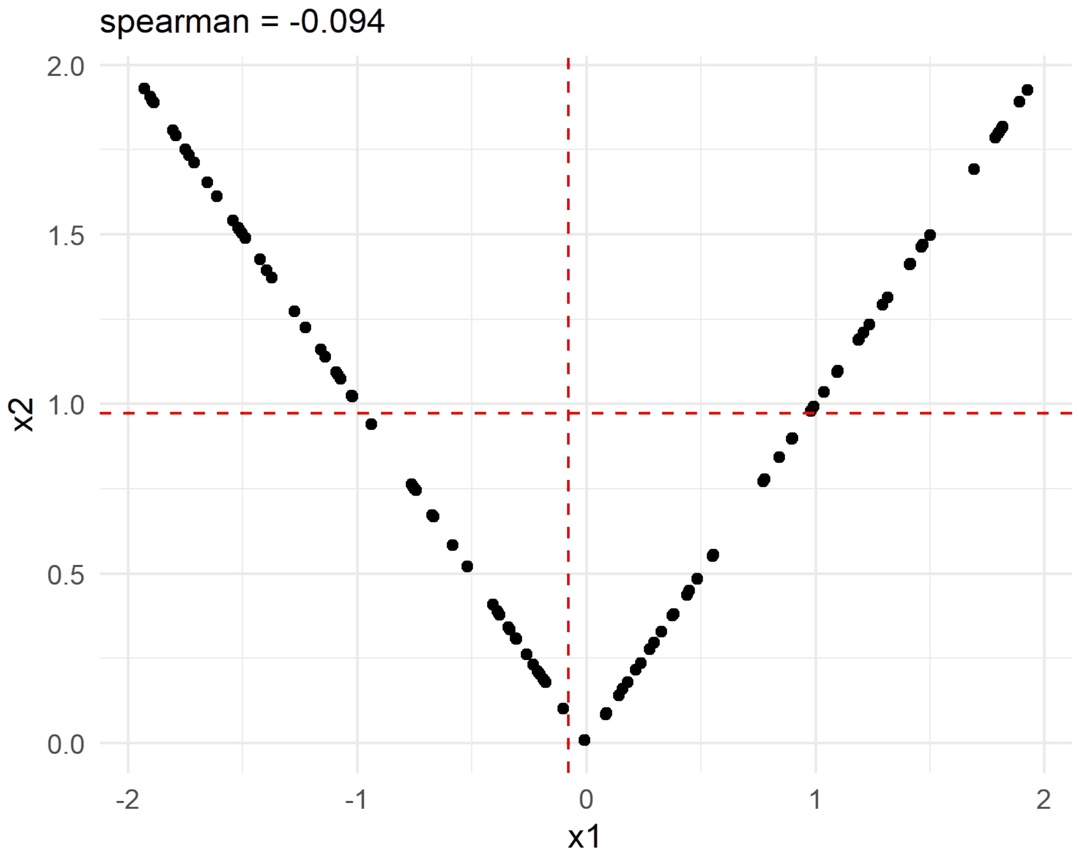
El estadístico mide la diferencia entre las frecuencias observadas y esperadas bajo independencia.

Podemos usar el **effect size** (**V de Cramér**) para medir **asociación entre variables categóricas**. Varía entre 0 y 1.

Sobre los rangos de valores: https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize

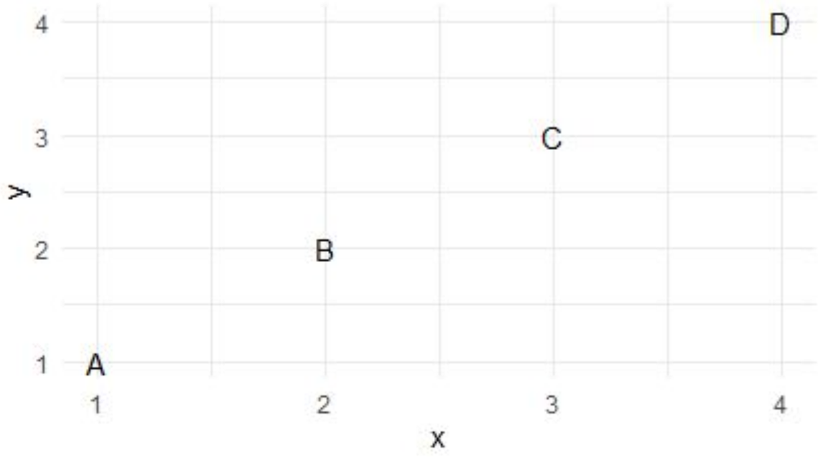
Un problema

Pearson/spearman pueden ser ≃ 0 pero esto no implica independencia:



Tal vez nos interesa capturar patrones más raros…
(e.g. no monótonos)

Distance Correlation (dCor)



dCor = 1

X

	A	B	C	D
A	0	1	2	3
B	1	0	1	2
C	2	1	0	1
D	3	2	1	0

Y

	A	B	C	D
A	0	1	2	3
B	1	0	1	2
C	2	1	0	1
D	3	2	1	0

matrices de distancias

	A	B	C	D
A	-1.75	-0.25	0.75	1.25
B	-0.25	-0.75	0.25	0.75
C	0.75	0.25	-0.75	-0.25
D	1.25	0.75	-0.25	-1.75

	A	B	C	D
A	-1.75	-0.25	0.75	1.25
B	-0.25	-0.75	0.25	0.75
C	0.75	0.25	-0.75	-0.25
D	1.25	0.75	-0.25	-1.75

matrices centradas

A	B	C	D
0.328125	0.078125	0.078125	0.328125

“distance covariance” por obs.

Distance Correlation (dCor)

$$a_{j,k} = \|X_j - X_k\|, \qquad j,k = 1,2,\ldots,n,$$
$$b_{j,k} = \|Y_j - Y_k\|, \qquad j,k = 1,2,\ldots,n,$$

matrices de distancias

$$A_{j,k} := a_{j,k} - \bar{a}_{j\cdot} - \bar{a}_{\cdot k} + \bar{a}_{\cdot\cdot}, \qquad B_{j,k} := b_{j,k} - \bar{b}_{j\cdot} - \bar{b}_{\cdot k} + \bar{b}_{\cdot\cdot},$$

matrices centradas

$$\mathrm{dCov}_n^2(X,Y) := \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}\,B_{j,k}.$$

distance covariance

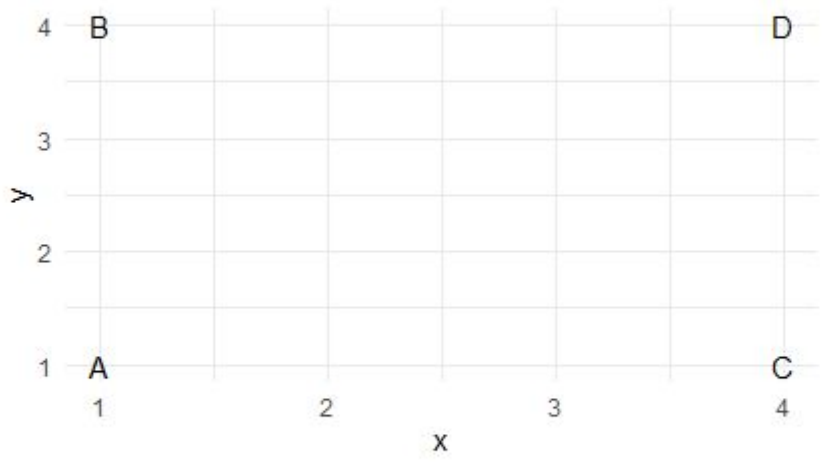
$$\mathrm{dVar}_n(X) := \mathrm{dCov}_n^2(X,X) = \tfrac{1}{n^2}\sum_{k,\ell} A_{k,\ell}^2,$$

distance variance

$$\boxed{\mathrm{dCor}(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dVar}(X)\,\mathrm{dVar}(Y)}},}$$

Fuente: https://en.wikipedia.org/wiki/Distance_correlation

Distance Correlation (dCor)



$dCor = 0$

X

	A	B	C	D
A	0	0	3	3
B	0	0	3	3
C	3	3	0	0
D	3	3	0	0

	A	B	C	D
A	0	3	0	3
B	3	0	3	0
C	0	3	0	3
D	3	0	3	0

Y

matrices de distancias

	A	B	C	D
A	-1.5	-1.5	1.5	1.5
B	-1.5	-1.5	1.5	1.5
C	1.5	1.5	-1.5	-1.5
D	1.5	1.5	-1.5	-1.5

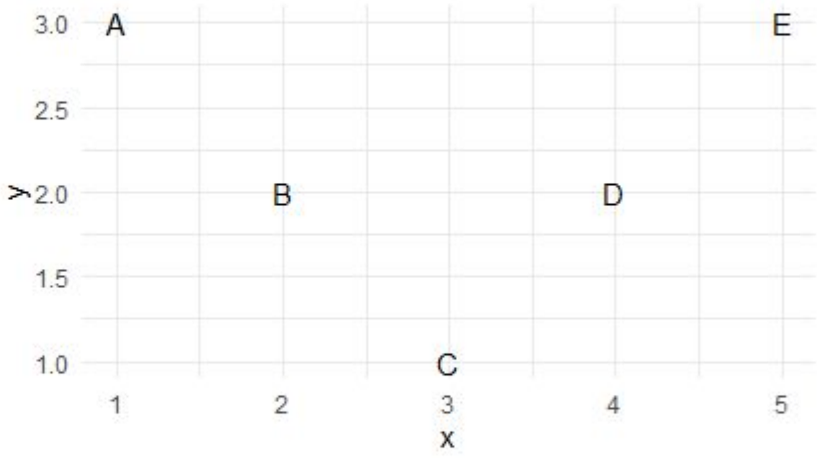
	A	B	C	D
A	-1.5	1.5	-1.5	1.5
B	1.5	-1.5	1.5	-1.5
C	-1.5	1.5	-1.5	1.5
D	1.5	-1.5	1.5	-1.5

matrices centradas

A	B	C	D
0	0	0	0

dCov por obs.

Distance Correlation (dCor)



dCor = 0.53

X

	A	B	C	D	E
A	0	1	2	3	4
B	1	0	1	2	3
C	2	1	0	1	2
D	3	2	1	0	1
E	4	3	2	1	0

Y

	A	B	C	D	E
A	0	1	2	1	0
B	1	0	1	0	1
C	2	1	0	1	2
D	1	0	1	0	1
E	0	1	2	1	0

matrices de distancias

	A	B	C	D	E
A	-2.4	-0.8	0.4	1.2	1.6
B	-0.8	-1.2	0.0	0.8	1.2
C	0.4	0.0	-0.8	0.0	0.4
D	1.2	0.8	0.0	-1.2	-0.8
E	1.6	1.2	0.4	-0.8	-2.4

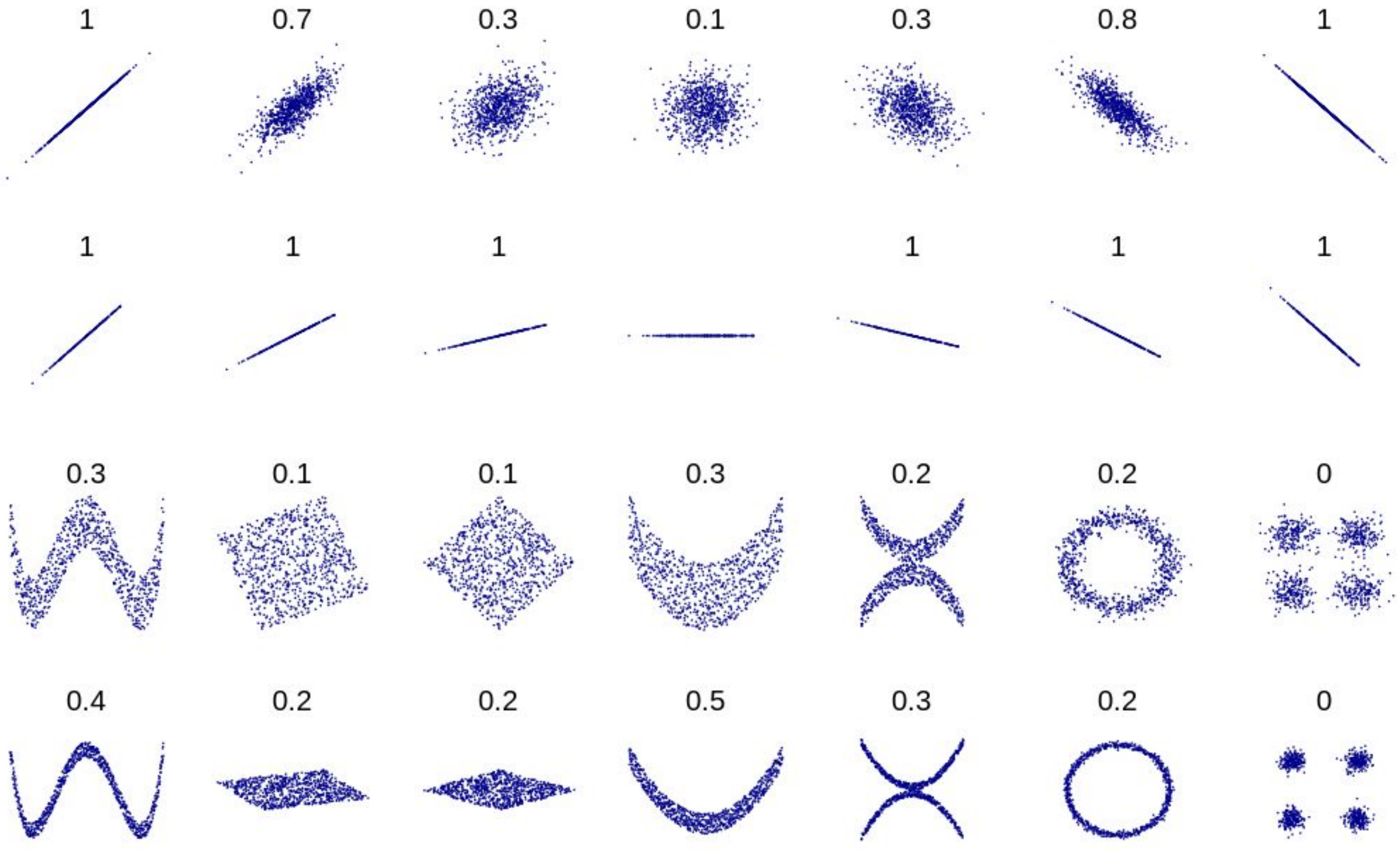
	A	B	C	D	E
A	-0.8	0.4	0.8	0.4	-0.8
B	0.4	-0.4	0.0	-0.4	0.4
C	0.8	0.0	-1.6	0.0	0.8
D	0.4	-0.4	0.0	-0.4	0.4
E	-0.8	0.4	0.8	0.4	-0.8

matrices centradas

A	B	C	D	E
0.0448	0.0128	0.0768	0.0128	0.0448

dCov por obs.

Distance Correlation (dCor)



***Otros*:**

→ HSIC (*Hilbert-Schmidt Independence Criterion*)

→ MIC (Maximal Mutual Information)

(Murphy 6.3)

Fuente: https://en.wikipedia.org/wiki/Distance_correlation

Lecturas recomendadas

- *Statistical Reasoning in the Behavioral Sciences (King et al, 2018)*
- *Handbook of Parametric and Nonparametric Statistical Procedures (Sheskin, 2011)* Table 1.20
- *Probabilistic Machine Learning (Murphy, 2022)* Cap. 3.1
- *Measures of Association: How to Choose? (Khamis, 2008)*
- *The need to report effect size estimates revisited (Tomczak y Tomczak, 2014)*