

Análisis de datos Bivariante

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA



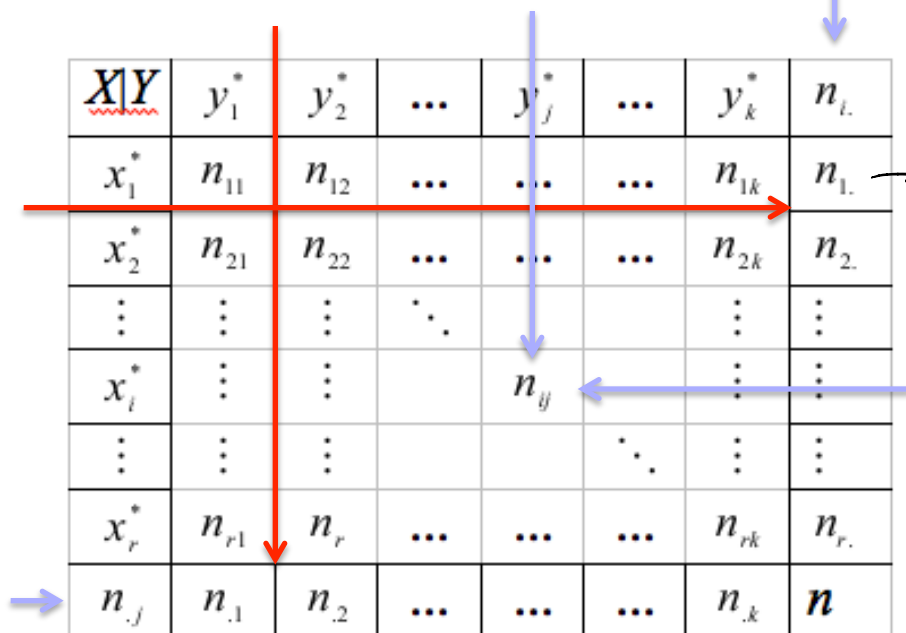
Tipología de los datos

$$(X, Y) \rightarrow (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

$x_i \in S_X = \{x_1^*, x_2^*, \dots, x_r^*\}$ Posibles valores de X (o sus marcas de clase).

$y_i \in S_Y = \{y_1^*, y_2^*, \dots, y_k^*\}$ Posibles valores de Y (o sus marcas de clase).

Tabla de contingencia (clasificación)



The table shows the joint and marginal frequencies for variables X and Y. A red arrow points down the first column, and a blue arrow points right across the first row. A blue arrow points down the j-th column, and a blue arrow points right across the i-th row. The bottom-right cell is labeled 'n'.

<u>$X \backslash Y$</u>	y_1^*	y_2^*	...	y_j^*	...	y_k^*	$n_{i.}$
x_1^*	n_{11}	n_{12}	n_{1k}	$n_{1.}$
x_2^*	n_{21}	n_{22}	n_{2k}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots			\vdots	\vdots
x_i^*	\vdots	\vdots		n_{ij}		\vdots	\vdots
\vdots	\vdots	\vdots			\ddots	\vdots	\vdots
x_r^*	n_{r1}	$n_{r.}$	n_{rk}	$n_{r.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.k}$	n

n_{ij} Frecuencias absolutas (frecuencia conjunta de ocurrencias de la categoría i de la variable X y la categoría j de la variable Y).

Distribuciones marginales

Frecuencias marginales absolutas.

$$n_{i.} = \sum_{j=1}^k n_{ij} \quad i = 1, \dots, r \quad (\text{variable } X)$$

$$n_{.j} = \sum_{i=1}^r n_{ij} \quad j = 1, \dots, k \quad (\text{variable } Y)$$

Supongamos que se pretende realizar un estudio para una muestra de una población en la que a cada individuo se le pregunta sobre su estado civil y se anota el sexo. Los datos obtenidos sobre un total de 15 individuos son:

(Soltero,Hombre), (C,H), (S,M), (S,M), (V,M), (C,M), (S,H), (C,H), (C,H), (C,M), (V,M), (C,H), (V,H), (S,H), (C,M).

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$n_{i.}$
<i>Hombre</i>	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
<i>Mujer</i>	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

Tablas de clasificación

$$(X, Y) \rightarrow (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

$$x_i \in S_X = \{x_1^*, x_2^*, \dots, x_r^*\} \quad \text{Posibles valores de } X \text{ (o sus marcas de clase).}$$

$$y_i \in S_Y = \{y_1^*, y_2^*, \dots, y_k^*\} \quad \text{Posibles valores de } Y \text{ (o sus marcas de clase).}$$

Tabla de contingencia

<u>$X \backslash Y$</u>	y_1^*	y_2^*	...	y_j^*	...	y_k^*	$n_{i.}$
x_1^*	n_{11}	n_{12}	n_{1k}	$n_{1.}$
x_2^*	n_{21}	n_{22}	n_{2k}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots			\vdots	\vdots
x_i^*	\vdots	\vdots		n_{ij}		\vdots	\vdots
\vdots	\vdots	\vdots			\ddots	\vdots	\vdots
x_r^*	n_{r1}	n_{r2}	n_{rk}	$n_{r.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.k}$	n

$$f_{ij} = \frac{n_{ij}}{n} \quad \text{Frecuencias relativas (al total de individuos).}$$

Distribuciones marginales

Frecuencias marginales relativas.

$$f_{i.} = \frac{n_{i.}}{n} \quad i = 1, \dots, r \quad (\text{variable } X)$$

$$f_{.j} = \frac{n_{.j}}{n} \quad j = 1, \dots, k \quad (\text{variable } Y)$$

Supongamos que se pretende realizar un estudio para una muestra de una población en la que a cada individuo se le pregunta sobre su estado civil y se anota el sexo. Los datos obtenidos sobre un total de 15 individuos son:

(Soltero,Hombre), (C,H), (S,M), (S,M), (V,M), (C,M), (S,H), (C,H), (C,H), (C,M), (V,M), (C,H), (V,H), (S,H), (C,M).

$S. \setminus E.C.$	Soltero	Casado	Viudo	$n_{i.}$
Hombre	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
Mujer	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

$$f_{ij} = \frac{n_{ij}}{n}$$

$S. \setminus E.C.$	Soltero	Casado	Viudo	f_i
Hombre	$f_{11} = 0,2$	$f_{12} = 0,266$	$f_{13} = 0,066$	$f_{1.} = 0,532$
Mujer	$f_{21} = 0,133$	$f_{22} = 0,2$	$f_{23} = 0,133$	$f_{2.} = 0,466$
$f_{.j}$	$f_{.1} = 0,333$	$f_{.2} = 0,466$	$f_{.3} = 0,199$	$f_{..} = 0,998 \simeq 1$

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA



Distribuciones condicionadas

$$f_{Y|X} = \frac{n_{ij}}{n_{i.}} \quad ; j = 1, \dots, k, \quad ; \quad i = 1, \dots, r$$

$$f_{X|Y} = \frac{n_{ij}}{n_{.j}} \quad ; i = 1, \dots, r, \quad ; j = 1, \dots, k$$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$n_{i.}$
<i>Hombre</i>	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
<i>Mujer</i>	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	
<i>Hombre</i>	0,376	0,5	0,124	1
<i>Mujer</i>	0,285	0,43	0,285	1

Frecuencias condicionadas del sexo para cada categoría del estado civil.

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	
<i>Hombre</i>	0,6	0,57	0,332	
<i>Mujer</i>	0,4	0,43	0,668	
	1	1	1	

Frecuencias condicionadas del estado civil para cada categoría del sexo.

Medidas gráficas

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$n_{i.}$
<i>Hombre</i>	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
<i>Mujer</i>	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

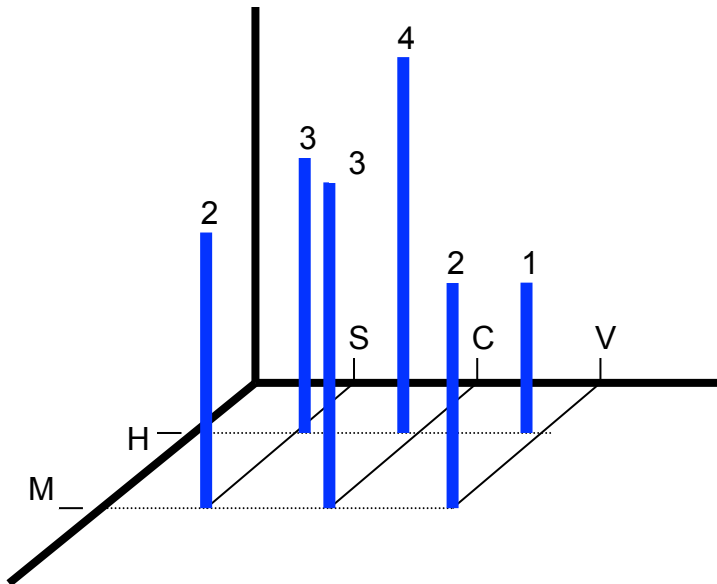
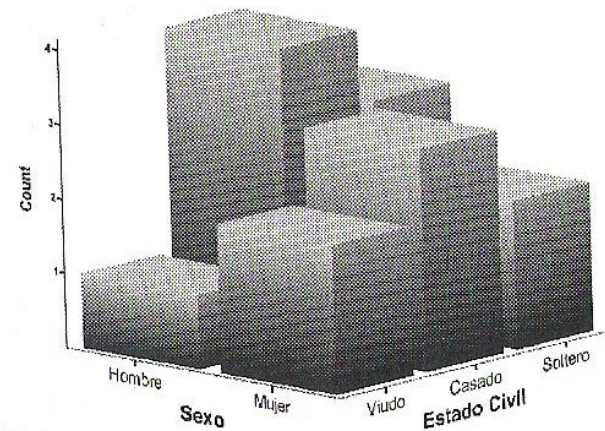


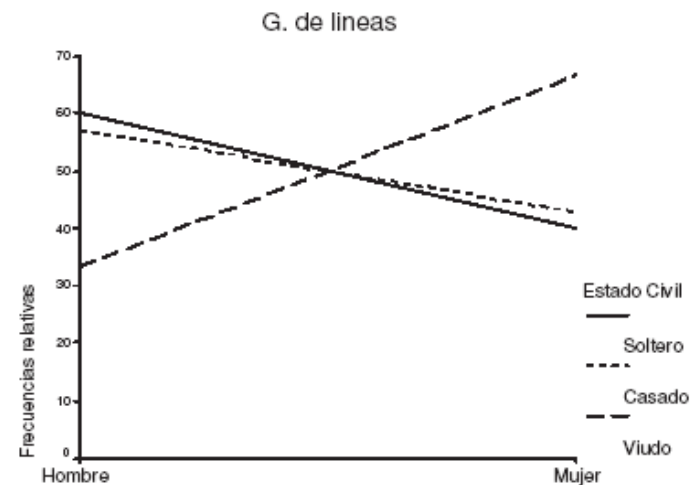
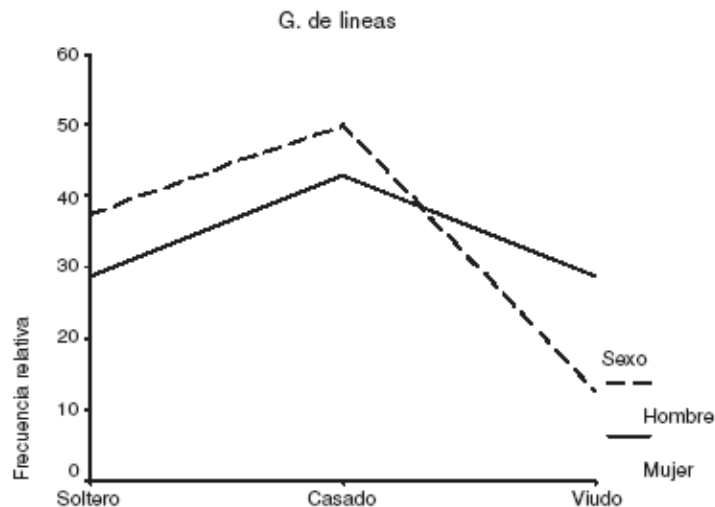
Diagrama de Barras



Medidas gráficas

<i>S. \ E.C.</i>	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	
<i>Hombre</i>	0,376	0,5	0,124	1
<i>Mujer</i>	0,285	0,43	0,285	1

<i>S. \ E.C.</i>	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	
<i>Hombre</i>	0,6	0,57	0,332	
<i>Mujer</i>	0,4	0,43	0,668	
	1	1	1	



Medidas de asociación:

Escala nominal

$$fe_{ij} = \frac{n_{i.} n_{.j}}{n}, \quad \forall ij \quad \text{frecuencias absolutas esperadas en caso de ausencia de asociación}$$

$$\chi^2 \text{ de Pearson} \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^r \sum_{j=1}^k \frac{(e_{ij})^2}{fe_{ij}}$$

$$\chi^2 \in [0, n \cdot t]; t = \min\{(r-1), (k-1)\}$$

$$\text{Coeficiente } C \text{ de contingencia} \quad C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad C \in [0, 1)$$

$$V \text{ de Cramer} \quad V = \sqrt{\frac{\chi^2}{nt}} \quad V \in [0, 1]$$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$n_{i.}$
<i>Hombre</i>	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
<i>Mujer</i>	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} - fe_{ij})^2}{fe_{ij}} = \sum_{i=1}^r \sum_{j=1}^k \frac{(e_{ij})^2}{fe_{ij}}$$

$$fe_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$fe_{(i,.)}$
<i>Hombre</i>	$fe_{e(1,1)} = 2,6$	$fe_{e(1,2)} = 3,7$	$fe_{e(1,3)} = 1,6$	$fe_{e(1,.)} = 7,9 \approx 8$
<i>Mujer</i>	$fe_{e(2,1)} = 2,3$	$fe_{e(2,2)} = 3,2$	$fe_{e(2,3)} = 1,4$	$fe_{e(2,.)} = 6,9 \approx 7$
$fe_{e(.,j)}$	$fe_{e(.,1)} = 4,9 \approx 5$	$fe_{e(.,2)} = 6,9 \approx 7$	$fe_{e(.,3)} = 3$	$fe_{e(.,.)} \approx 15$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>
<i>Hombre</i>	$e_{11} = 0,4$	$e_{12} = 0,3$	$e_{13} = -0,6$
<i>Mujer</i>	$e_{21} = -0,3$	$e_{22} = -0,2$	$e_{23} = 0,6$

$$\chi^2 = \sum_{\forall i,j} \frac{e_{ij}^2}{fe_{(i,j)}} = 0,59$$

$$\max \chi^2 = n \cdot t = n \cdot \min\{(2-1), (3-1)\} = 15 \cdot 1 = 15$$

$$V = \sqrt{\frac{\chi^2}{n \cdot t}} = \sqrt{\frac{0,59}{15}} = 0,198$$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>
<i>Hombre</i>	$\frac{e_{11}^2}{fe_{(i,j)}} = 0,06$	0,02	0,22
<i>Mujer</i>	0,03	0,01	0,25

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{0,59}{0,59 + 15}} = 0,195$$

Escala ordinal

Coeficientes predictivos λ :

$$\lambda_{Y/X} = \frac{\sum_{i=1}^r \max_{j=1..k} n_{ij} - \max_{j=1..k} n_{.j}}{n - \max_{j=1..k} n_{.j}} \in [0,1]$$

$$\lambda_{X/Y} = \frac{\sum_{j=1}^k \max_{i=1..r} n_{ij} - \max_{i=1..r} n_{i.}}{n - \max_{i=1..r} n_{i.}} \in [0,1]$$

$S. \setminus E.C.$	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	$n_{i.}$
<i>Hombre</i>	$n_{11} = 3$	$n_{12} = 4$	$n_{13} = 1$	$n_{1.} = 8$
<i>Mujer</i>	$n_{21} = 2$	$n_{22} = 3$	$n_{23} = 2$	$n_{2.} = 7$
$n_{.j}$	$n_{.1} = 5$	$n_{.2} = 7$	$n_{.3} = 3$	$n_{..} = 15$

$$\lambda_{Y/X} = \frac{4 + 3 - 7}{15 - 7} = 0$$

Indica que la variable X (sexo) no tiene ningún poder predictivo sobre Y (estado civil).

$$\lambda_{X/Y} = \frac{3 + 4 + 2 - 8}{15 - 8} = 0,142$$

Escaso poder predictivo de Y (estado civil) sobre X (sexo).

Escala numérica

Covarianza
$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y} = \left(\frac{1}{n} \sum_{i=1}^r \sum_{j=1}^k n_{ij} x_i y_i \right) - \bar{x} \bar{y}$$

Coeficiente de correlación
$$r_{xy} = \frac{S_{xy}}{S_x S_y} \in [-1, 1]$$

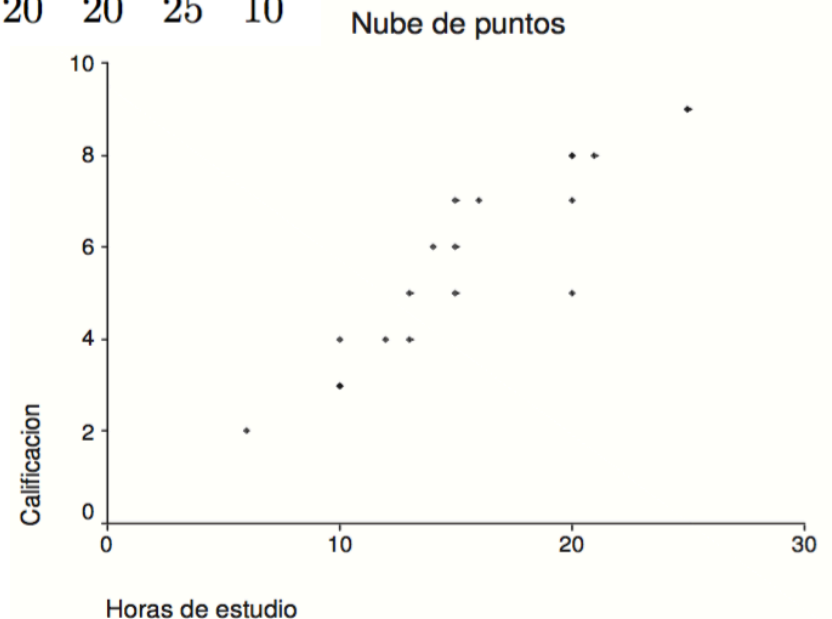
<i>X : Calificación</i>	5	7	3	8	4	6	9	8	3	5
<i>Y : Horas</i>	15	20	10	21	12	15	25	20	10	13

<i>X : Calificación</i>	7	4	3	2	6	7	8	5	9	4
<i>Y : Horas</i>	15	13	10	6	14	16	20	20	25	10

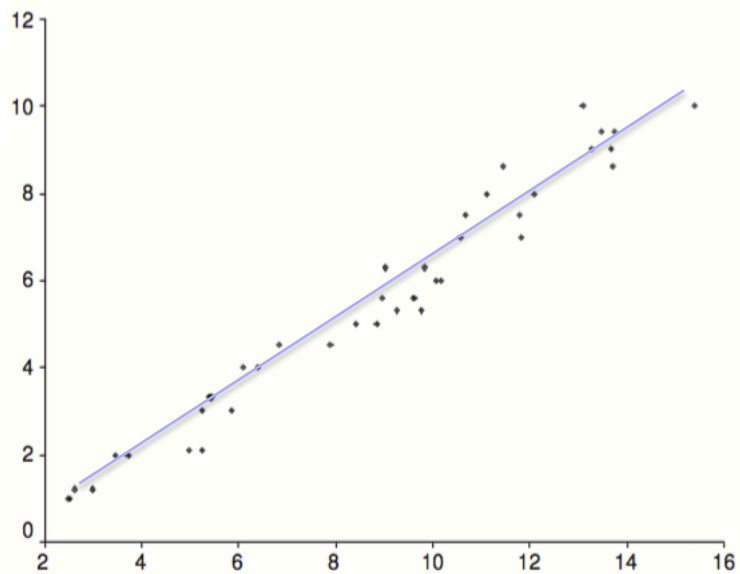
$$S_x^2 = 4.4275$$

$$S_y^2 = 26.55$$

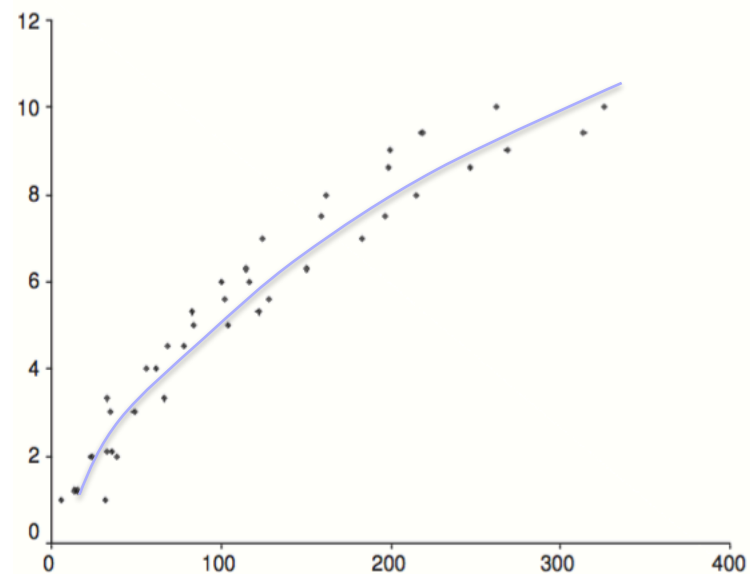
$$S_{xy} = 9.975 \quad r_{xy} = \frac{S_{xy}}{S_x S_y} = 0.92$$



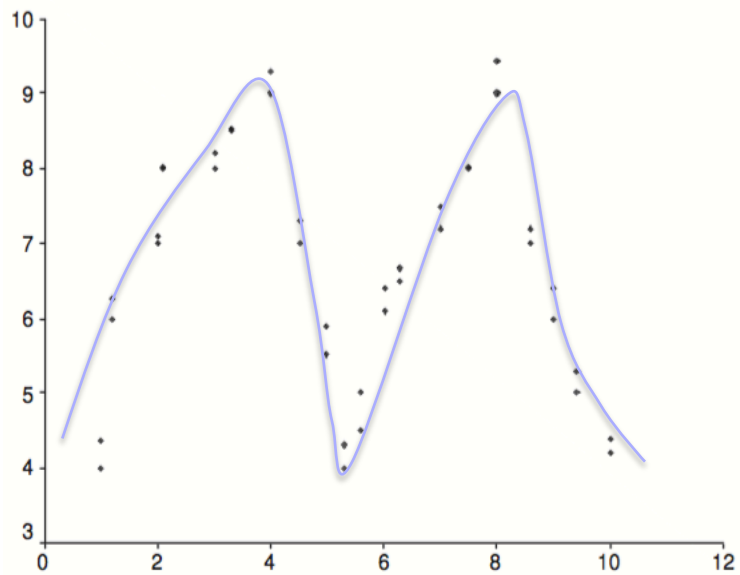
R. lineal creciente



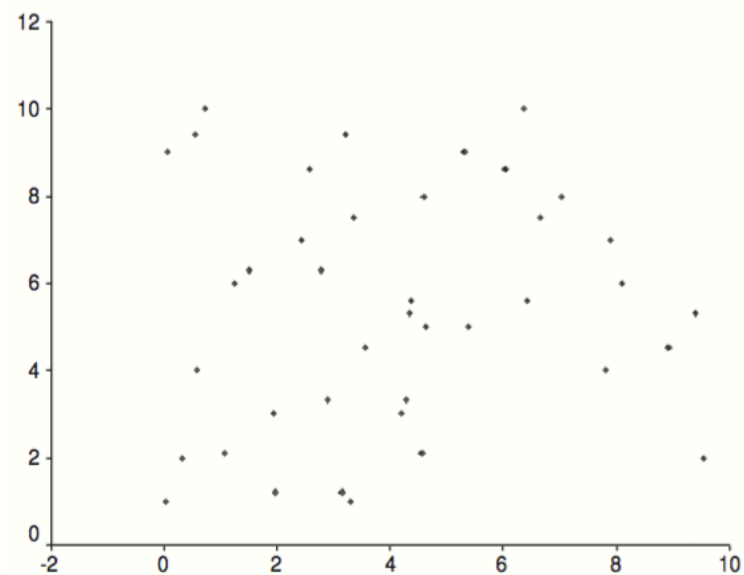
R. Monotona



R. no lineal



Ausencia de R.



Análisis de datos Bivariante

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA

