Datos Multivariados

Estadística Computacional

Juan Zamora Osorio
juan.zamora@pucv.cl

Instituto de Estadística
Pontificia Universidad Católica de Valparaíso

28 de agosto de 2023

Datos bivariados

id Pasajero	Clase	Satisfacción
1	E	2
2	E	4
3	E	1
4	B	3
5	E	1
6	B	2
7	P	4
8	E	3
9	E	2
10	B	4
11	E	3
12	B	3
⋮	⋮	⋮

Datos bivariados categóricos – tabla de contingencia

id Pasajero	Clase	Satisfacción
1	E	2
2	E	4
3	E	1
4	B	3
5	E	1
6	B	2
7	P	4
8	E	3
9	E	2
10	B	4
11	E	3
12	B	3
⋮	⋮	⋮

		Satisfacción				Total
		1	2	3	4	
Clase	E	2	2	2	1	7
	B	0	1	2	1	4
	P	0	0	0	1	1
Total		2	3	4	3	12

Datos bivariados categóricos – tabla de contingencia

		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100

Clases bivariadas

- $(E, 1)$, $(E, 2)$, $(E, 3)$, $(E, 4)$, $(B, 2)$, $(B, 3)$, $(B, 4)$, $(P, 3)$, $(P, 4)$.
- ¡Pueden ser muchas!

Datos bivariados categóricos – tabla de contingencia

				Y			Total
		y_1	$\cdots$	y_j	$\cdots$	y_J	
	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kJ}	$n_{k\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_K	n_{K1}	$\cdots$	n_{Kj}	$\cdots$	n_{KJ}	$n_{K\cdot}$
Total		$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot J}$	n

Frecuencias absolutas marginales

$$n = \sum_{k=1}^{K} n_{k\cdot} = \sum_{j=1}^{j} n_{\cdot j} = \sum_{k=1}^{K} \sum_{j=1}^{J} n_{kj}.$$

Datos bivariados categóricos – tabla de contingencia

		Y					Total
		y_1	$\cdots$	y_j	$\cdots$	y_J	
	x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1J}	$f_{1\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X	x_k	f_{k1}	$\cdots$	f_{kj}	$\cdots$	f_{kJ}	$f_{k\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_K	f_{K1}	$\cdots$	f_{Kj}	$\cdots$	f_{KJ}	$f_{K\cdot}$
Total		$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot J}$	1

Frecuencias relativas marginales

$$1 = \sum_{k=1}^{K} f_{k\cdot} = \sum_{j=1}^{j} f_{\cdot j} = \sum_{k=1}^{K} \sum_{j=1}^{J} f_{kj}.$$

Datos bivariados categóricos – frecuencias condicionales

		Y					Total
		y_1	$\cdots$	y_j	$\cdots$	y_J	
	x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1J}	$f_{1\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X	x_k	f_{k1}	$\cdots$	f_{kj}	$\cdots$	f_{kJ}	$f_{k\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_K	f_{K1}	$\cdots$	f_{Kj}	$\cdots$	f_{KJ}	$f_{K\cdot}$
Total		$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot J}$	1

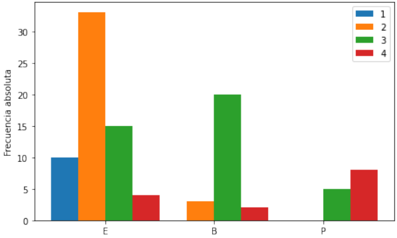
Frecuencias condicionales

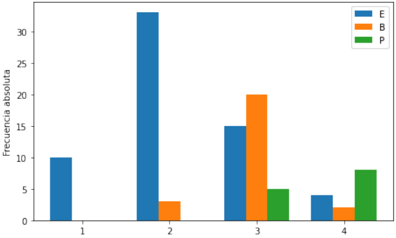
$$f_{k|j} = \frac{f_{kj}}{f_{\cdot j}} = \frac{n_{kj}}{n_{\cdot j}},$$

$$f_{j|k} = \frac{f_{kj}}{f_{k\cdot}} = \frac{n_{kj}}{n_{k\cdot}}.$$

Ejemplo – satisfacción de vuelos

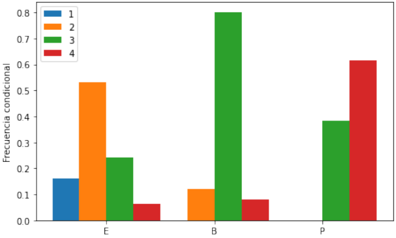
		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100

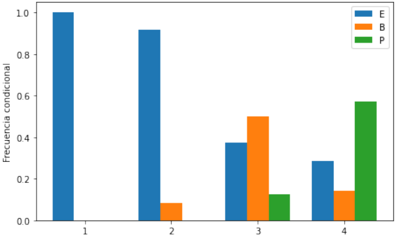




Ejemplo – satisfacción de vuelos

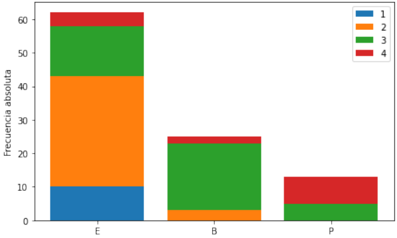
		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100

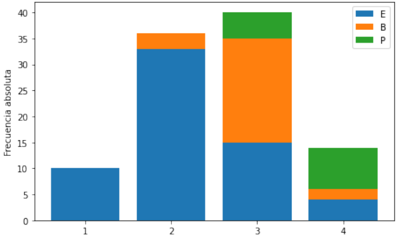




Ejemplo – satisfacción de vuelos

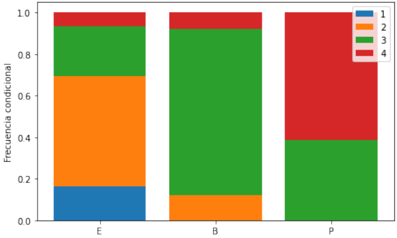
		Satisfacción				Total
		1	2	3	4	
	E	10	33	15	4	62
Clase	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100

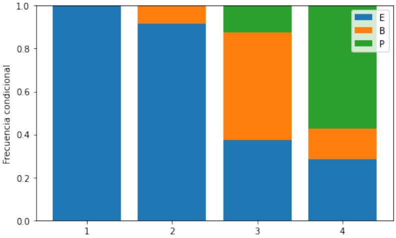




Ejemplo – satisfacción de vuelos

		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100





Independencia

		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100

Se busca

- $f_{k|j} = f_{k\cdot}$ y $f_{j|k} = f_{\cdot j}$.

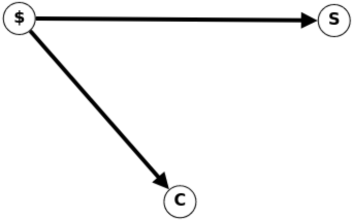
Luego

$$f_{kj} = f_{k\cdot} f_{\cdot j}.$$

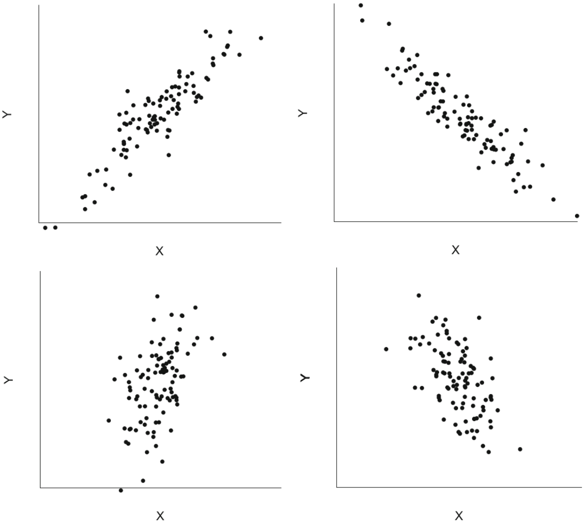
- Esperamos frecuencia absoluta
  $n_{kj} = nf_{kj} = nf_{k\cdot} f_{\cdot j} = n\frac{n_{k\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{k\cdot} n_{\cdot j}}{n}$.

Independencia

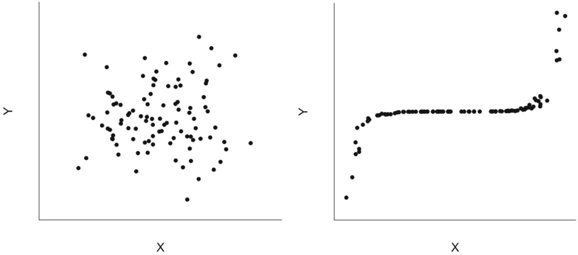
		Satisfacción				Total
		1	2	3	4	
Clase	E	10	33	15	4	62
	B	0	3	20	2	25
	P	0	0	5	8	13
Total		10	36	40	14	100



Datos bivariados continuos

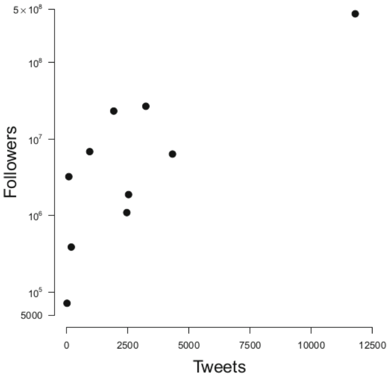


Gráficos de dispersión (*scatter*)

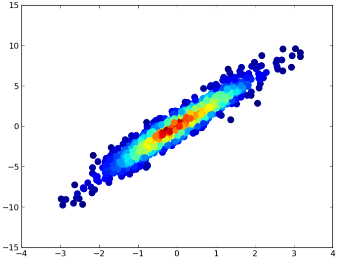


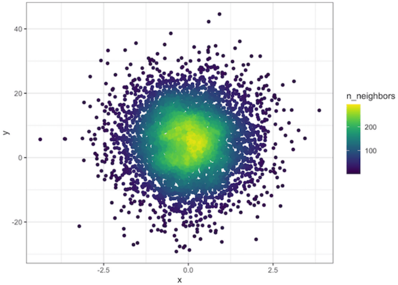
Gráficos de dispersión (*scatter*)

Name	Tweets	Followers
Angela Merkel	25	7194
Barack Obama	11,800	43,400,000
Jacob Zuma	99	324,000
Dilma Rousseff	1934	2,330,000
Sauli Niinistö	199	39,000
Vladimir Putin	2539	189,000
Francois Hollande	4334	639,000
David Cameron	952	688,000
Enrique P. Nieto	3245	2,690,000
John Key	2468	110,000

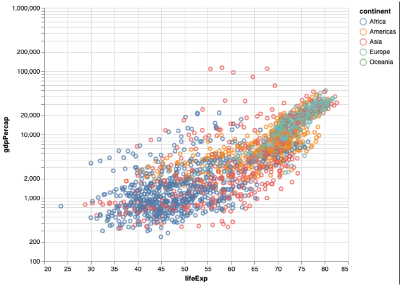


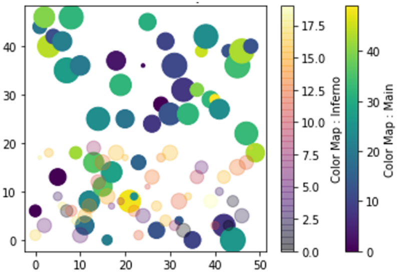
Gráficos de dispersión (*scatter*)



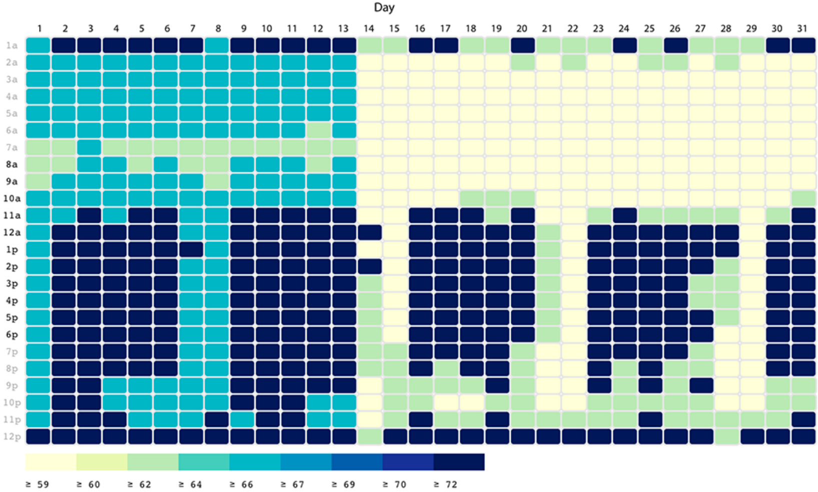


Gráficos de dispersión (*scatter*)

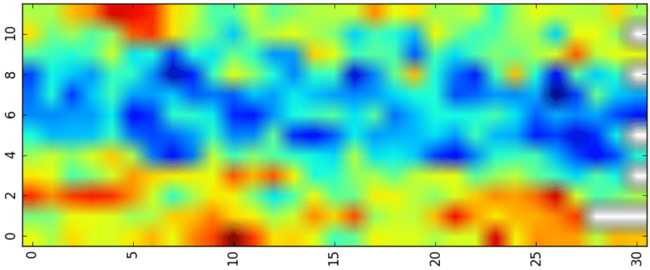




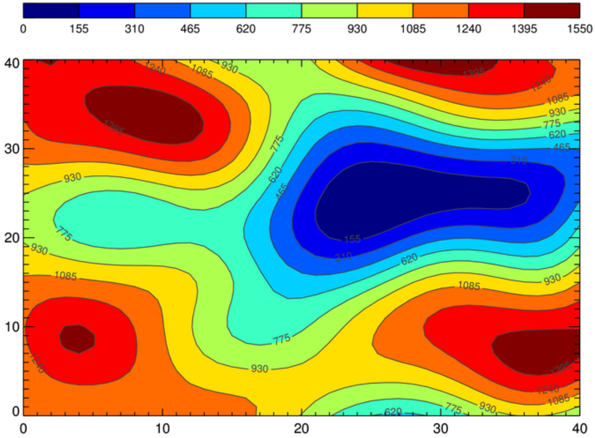
Mapas de calor



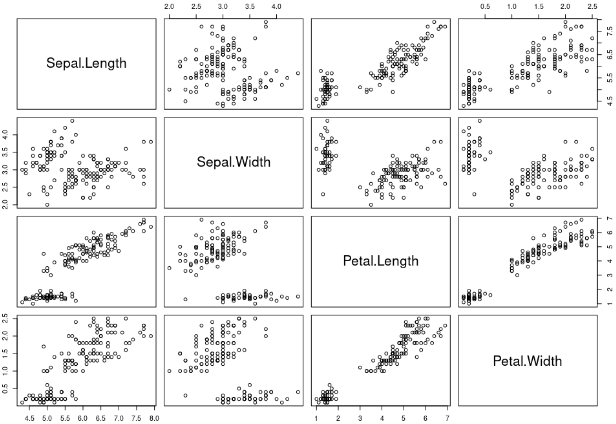
Mapas de calor



Mapas de calor – contorno



Datos multivariados



Correlación

Covarianza

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

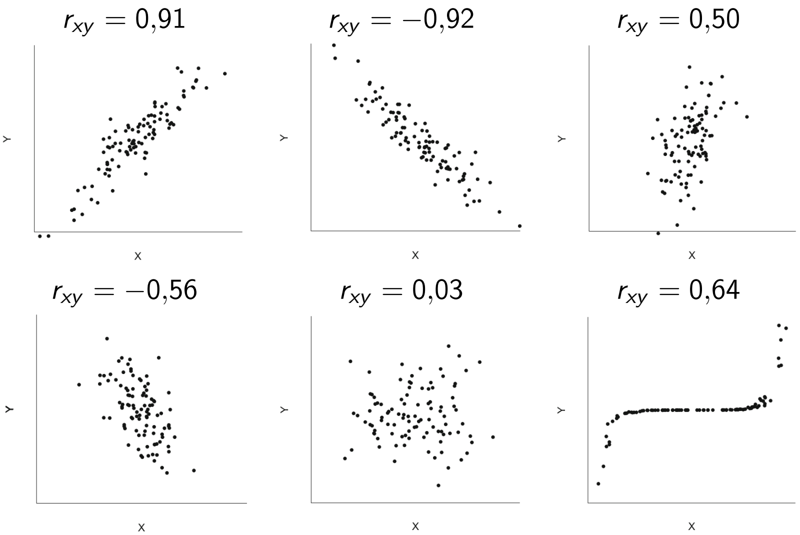
- Indica dependencia *lineal*.
- Notar $\mathrm{Cov}(x, x) = s_x^2$.

Correlación

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$

- Indica dependencia *lineal*.
- Se puede mostrar que $-1 \leq r_{xy} \leq 1$.

Correlación



Matriz de varianzas y covarianzas

$$S = \begin{bmatrix} \mathrm{Cov}(x^1, x^1) & \cdots & \mathrm{Cov}(x^1, x^j) & \cdots & \mathrm{Cov}(x^1, x^J) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x^j, x^1) & \cdots & \mathrm{Cov}(x^j, x^j) & \cdots & \mathrm{Cov}(x^j, x^J) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x^J, x^1) & \cdots & \mathrm{Cov}(x^J, x^j) & \cdots & \mathrm{Cov}(x^J, x^J) \end{bmatrix}$$

Propiedades

- Diagonal contiene las varianzas.
- Simétrica: $S = S^T$.
- Semidefinida positiva:
  - $\forall a \in \mathbb{R}^J, a^T S a \geq 0$.
  - Valores propios son todos no negativos.

Matriz de varianzas y covarianzas

Matriz de datos

$$\mathrm{X} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^J \end{bmatrix} = \begin{bmatrix} \mathrm{x}_1^T \\ \vdots \\ \mathrm{x}_i^T \\ \vdots \\ \mathrm{x}_n^T \end{bmatrix}$$

Construcción

$$\bar{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i,$$

$$\mathrm{S} = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{x}_i - \bar{\mathrm{x}})(\mathrm{x}_i - \bar{\mathrm{x}})^T.$$

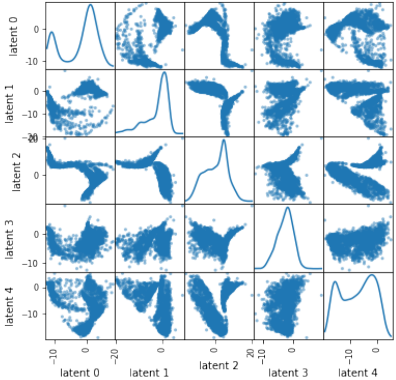
Matriz de correlación

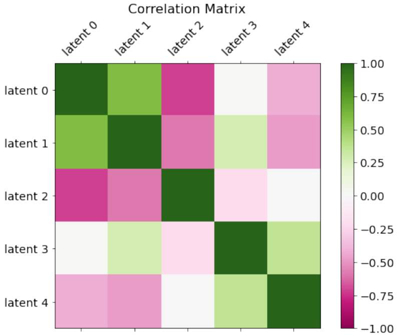
$$
R = \begin{bmatrix}
1 & \cdots & r_{1j} & \cdots & r_{1J} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
r_{j1} & \cdots & 1 & \cdots & r_{jJ} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
r_{J1} & \cdots & r_{Jj} & \cdots & 1
\end{bmatrix}
$$

Propiedades

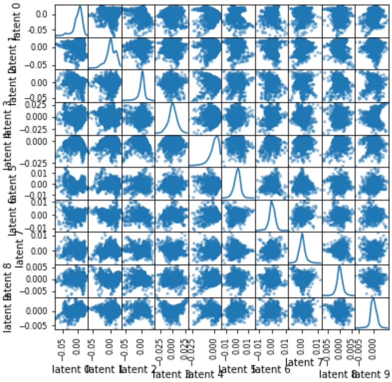
- No es igual a la matriz de varianzas y covarianzas escalada.
- Se puede construir como S estandarizando los atributos primero.

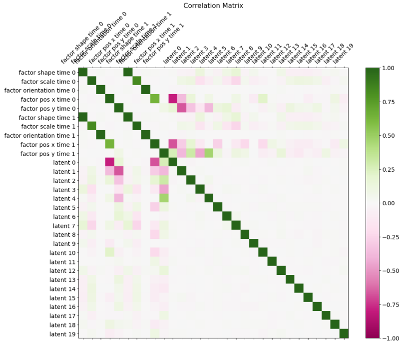
Matriz de correlación





Matriz de correlación





Matrices de correlación y de varianzas y covarianzas