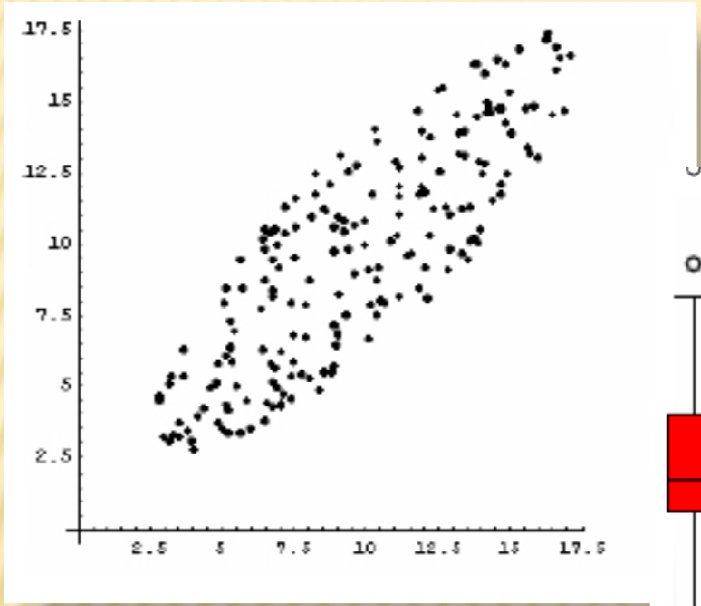
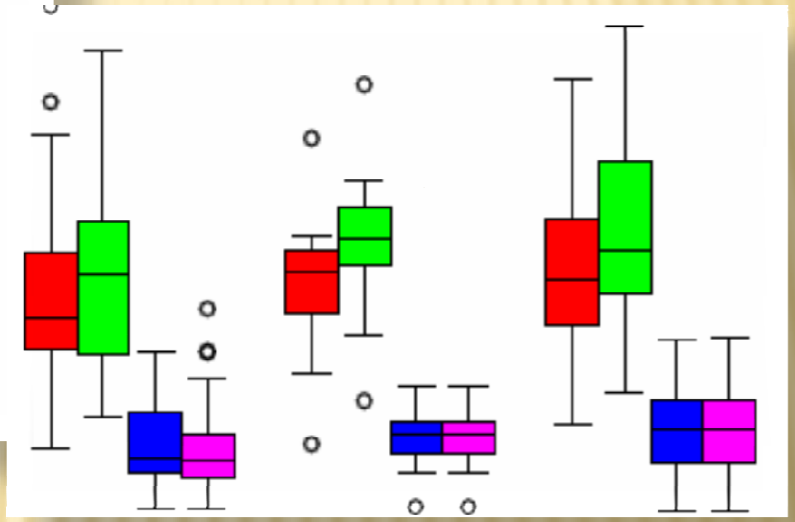
ANÁLISIS MULTIVARIANTE





Wenceslao González Manteiga

Wenceslao.gonzalez@usc.es

ÍNDICE

ÍNDICE

Bibliografía

- ANDERSON, T.W. (2003), *An introduction to multivariate statistical analysis.* Wiley
- EVERITT, B. (2005), *An R and S-Plus companion to multivariate analysis.* Springer.
- HASTIE, T., TIBSHIRANI, R. y FRIEDMAN, J. (2009) *The elements of statistical learning.* Springer
- JOHNSON, R.A. y Wichern, D.W. (2007) *Applied multivariate statistical analysis.* Pearson Education.
- MARDIA, K.V., KENT, J.T. y BIBBY, J.M. (1979). *Multivariate analysis*. Academic Press
- PEÑA, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill.
- PÉREZ, C. (2004). *Técnicas de análisis multivariante de datos.* Pearson Education, S.A.
- SEBER, G.A.F. (1984). *Multivariate observations.* Wiley.

TEMA 1:
ANALISIS EXPLORATORIO DE DATOS MULTIVARIANTES

TEMA 1. ANÁLISIS EXPLORATORIO DE DATOS

- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

TEMA 1. ANÁLISIS EXPLORATORIO DE DATOS

Los **DATOS** consisten en observaciones de **n individuos** en los que se miden **p características o variables**, las mismas en todos.
Los datos se disponen ordenadamente en la MATRIZ DE DATOS:
- *Los individuos en filas* - *Las variables en columnas*

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

x_{ij} es el valor de la variable j para el individuo i

X_i es un vector columna que contiene los valores de las p variables en el individuo i (lo que sabemos del individuo)
$X'_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ es el vector traspuesto de X_i

X es una matriz $n \times p$

Variables

Individuos

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

- MATRIZ DE DISTANCIAS
- MATRIZ DE FRECUENCIAS
- MATRIZ DE CORRELACIONES

TEMA 1. ANÁLISIS EXPLORATORIO DE DATOS

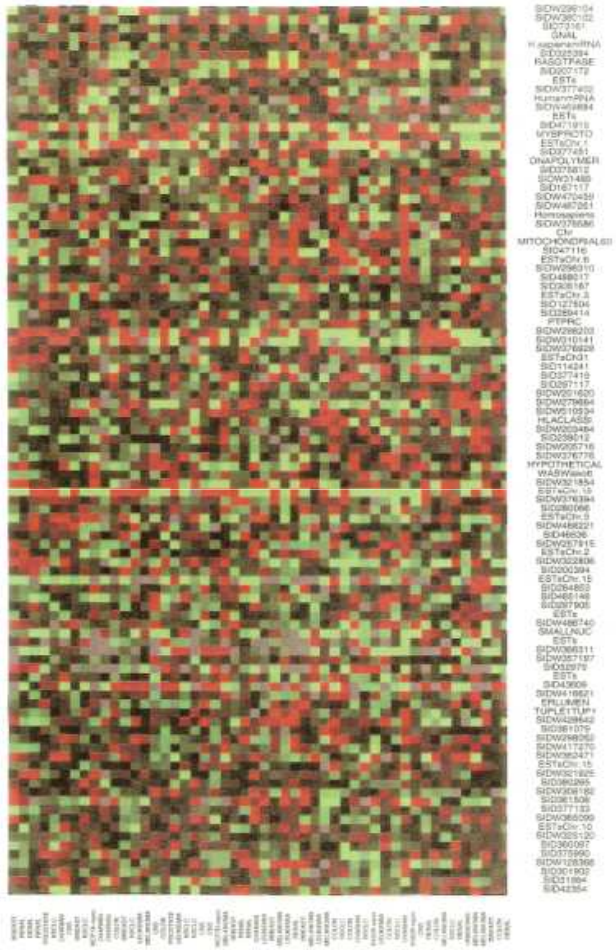
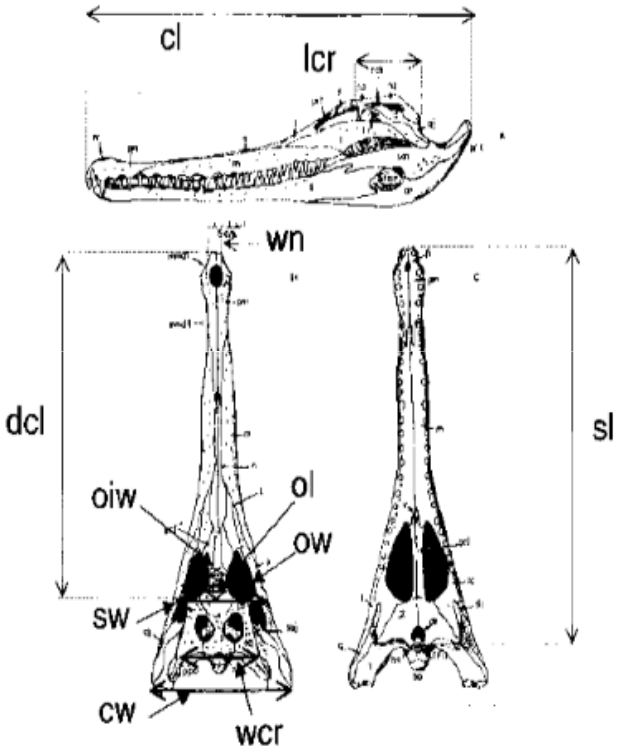


FIGURE 1.3. *DNA microarray data: expression matrix of 6830 genes (rows) and 64 samples (columns), for the human tumor data. Only a random sample of 100 rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*

Ejemplo: **Medidas de cráneos de cocodrilos** (alligator.txt)

Código	Descripción
cl	Longitud del cráneo
cw	Ancho del cráneo
sw	Ancho del hocico
sl	Longitud del hocico
dcl	Longitud dorsal del cráneo
ow	Ancho máximo orbital
oiw	Ancho mínimo inter–orbital
ol	Longitud máxima orbital
lcr	Longitud del paladar post–orbital
wcr	Ancho posterior del paladar craneal
wn	Ancho máximo entre orificios nasales



Tenemos todas estas medidas registradas para 44 cocodrilos de cuatro especies distintas

Ejemplo: **Medidas de cráneos de cocodrilos** (alligator.txt)

Valores de todas las variables en un ejemplar de cada especie:

Especie	cl	cw	sw	sl	dcl	ow	iow	ol	lcr	wcr	wn
Crocodylus niloticus	160	64	46	100	153	20	9	22	30	39	9
Crocodylus porosus	76	30	22	41	73	13	3.5	17	16	20	4
Osteolaemus tetraspis	164	90	70	90	160	36	16	42	32	57	20
Alligator mississippiensis	72.3	40.0	37.3	35	70.5	16.7	5.2	20	15	24.6	10.5

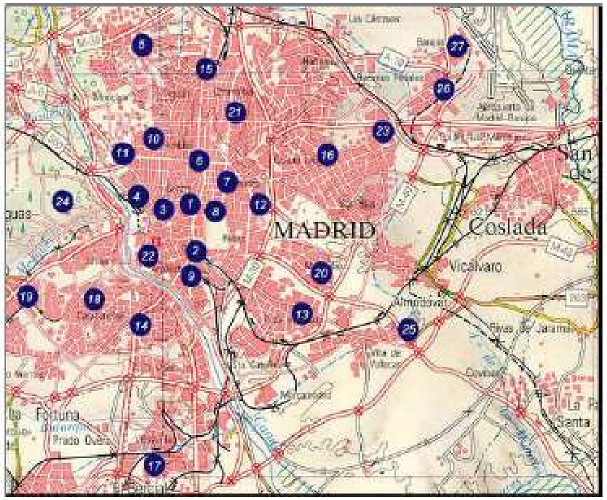
La matriz de datos, considerando sólo las variables cuantitativas, es:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \mathbf{x}_3' \\ \vdots \\ \mathbf{x}_{44}' \end{bmatrix} = \begin{bmatrix} 160 & 64 & 46 & 100 & 153 & 20 & 9 & 22 & 30 & 39 & 9 \\ 76 & 30 & 22 & 41 & 73 & 13 & 3{,}5 & 17 & 16 & 20 & 4 \\ 164 & 90 & 70 & 90 & 160 & 36 & 16 & 42 & 32 & 57 & 20 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 380 & 236 & 210 & 238 & 358 & 52 & 27 & 63 & 63 & 120 & 64 \end{bmatrix}$$

donde $\mathbf{x}_i'$ (la fila i de $\mathbf{X}$) corresponde a las mediciones en el cocodrilo i.

Ejemplo: **Calidad del aire en la ciudad de Madrid**

Para establecer un "Ranking de calidad del aire" por distritos en la ciudad de Madrid disponemos de la información registrada en 19 estaciones de medición atmosférica, que proporcionan datos de C0, SO2, NOX, P10 y O3.



SO2	Dióxido de Azufre
CO	Monóxido de Carbono
NOX	Óxidos de Nitrógeno
P10	Partículas PM10
O3	Ozono

12-5-09 9:00	CO (mg/m3)	SO2 (µg/m3)	NOX (µg/m3)	P10 (µg/m3)	O3 (µg/m3)
PLAZA DEL CARMEN	0,52	10,67	91,43	36,66	15,92
PLAZA ESPAÑA	0,87	13,37	157,76	72,75	17,08
BARRIO DEL PILAR	0,35	7,58	37,47	35,03	19,55
MARAÑON	1,14	13,4	166,81	55,45	14,96
MARQUES DE SALAMANCA	0,79	12,66	135,1	51,55	13,5
ESCUELAS AGUIRRE	0,65	10,99	96,57	51,72	9,43
LUCA DE TENA	0,86	6,65	186,72	50,85	10,97
CUATRO CAMINOS	0,49	7,66	63,35	35,33	18,18
AVDA. RAMON Y CAJAL	0,36	11,03	66,87	35,63	9,54
MANUEL BECERRA	0,76	9,71	161,1	43,5	9,73
VALLECAS	0,48	8,46	72,42	38,85	7,6
PLAZA FERNANDEZ LADREDA	0,49	9,59	106,28	36,85	1,64
ARTURO SORIA	0,57	11,2	143,47	44,42	10,97
GRAL.RICARDOS	0,50	13,55	49,89	64,8	9,47
Pº EXTREMADURA	0,69	10,95	114,09	61,54	8,93
MORATALAZ	0,59	11,55	72,49	38,75	11,46
ISAAC PERAL	0,50	14,71	117,57	30,7	4,65
Pº PONTONES	0,60	10,91	170,42	83,31	7,37
SANTA EUGENIA	0,35	0,7	41,31	29,16	17,18

Ejemplo: **Esclerosis múltiple**

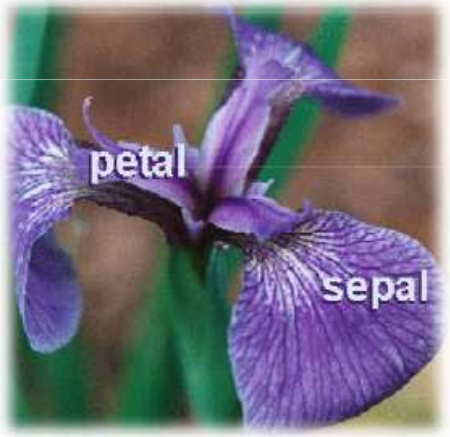
En un estudio sobre esclerosis múltiple se registran las respuestas del ojo izquierdo (I) y del ojo derecho (D) a dos estímulos visuales diferentes. Se consideran dos grupos, 29 individuos que padecen esclerosis múltiple y un grupo control de 69 individuos que no la padecen. Se registran las siguientes variables: X_1: Edad, $X_2 = R1L+R1D$, $X_3 = |R1L-R1D|$, $X_4 = R2L+R2D$, $X_5 = |R2L - R2D|$.

X_1	X_2	X_3	X_4	X_5	Paciente/Control
23	148.0	0.8	205.4	0.6	1
25	195.2	3.2	262.8	0.4	1
25	158.0	8.0	209.8	12.2	1
28	134.4	0.0	198.4	3.2	1
29	190.2	14.2	243.8	10.6	1
18	152.0	1.6	198.4	0.0	0
19	138.0	0.4	180.8	1.6	0
20	144.0	0.0	186.4	0.8	0
20	143.6	3.2	194.8	0.0	0
20	148.8	0.0	217.6	0.0	0

Ejemplo: **Lirios**

En un estudio del estadístico y genetista Sir Ronald A. Fisher se utilizaron cuatro características de los sépalos y pétalos para identificar los lirios de las especies *iris setosa*, *iris versicolor* e *iris virginica*.

Código	Descripción
CLASS	Especie
SL	Longitud del sépalo
SW	Anchura del sépalo
PL	Longitud del pétalo
PW	Anchura del pétalo



Ejemplo: **Lirios**

CLASS	PL	PW	SL	SW
setosa	5.1	3.5	1.4	0.2
versicolor	7	3.2	4.7	1.4
virginica	6.3	3.3	6	2.5

En total hay 50 lirios de cada especie (es decir, la matriz de datos es 150×4, si no tenemos en cuenta la variable que indica el nombre de la especie)

Ejemplo: **Lirios**

CLASS	PL	PW	SL	SW
setosa	5.1	3.5	1.4	0.2
versicolor	7	3.2	4.7	1.4
virginica	6.3	3.3	6	2.5

En total hay 50 lirios de cada especie (es decir, la matriz de datos es 150×4, si no tenemos en cuenta la variable que indica el nombre de la especie)

Ejemplo: **Rectángulos**

Ejemplo 5.9 del libro *Análisis de Datos Multivariantes* de Daniel Peña. Se tienen 6 observaciones bivariantes, cada observación corresponde con un rectángulo y las variables univariantes son la longitud de la base y la altura del rectángulo. La matriz de datos es:

$$\mathbf{X} = \begin{bmatrix} 2{,}0 & 2{,}0 \\ 1{,}5 & 0{,}5 \\ 0{,}7 & 0{,}5 \\ 0{,}5 & 1{,}5 \\ 0{,}5 & 0{,}7 \\ 0{,}7 & 0{,}7 \end{bmatrix}.$$

- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

Resumen numérico de los datos variable por variable (UNIVARIANTE)

Media muestral de la variable x_j:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Varianza muestral de la variable x_j:

$$s_j^2 = s_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

… o de dos en dos variables (BIVARIANTE)

Covarianza muestral entre las variables x_j y x_k:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

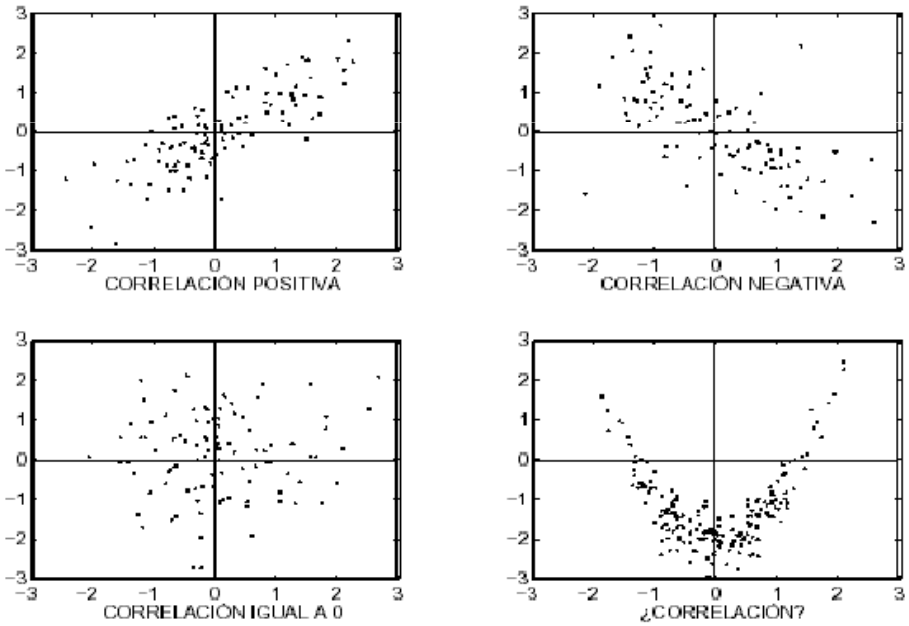
como depende de las unidades de medida, se utiliza el

Coeficiente de correlación muestral entre las variables x_j y x_k:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{s_{jk}}{s_j s_k}$$

El coeficiente de correlación (de Pearson) mide el grado de asociación lineal entre dos variables.

Toma valores entre -1 y 1. El signo indica si la relación es positiva o negativa

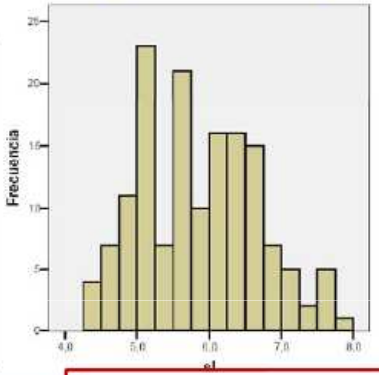


Ejemplo: **Lirios**

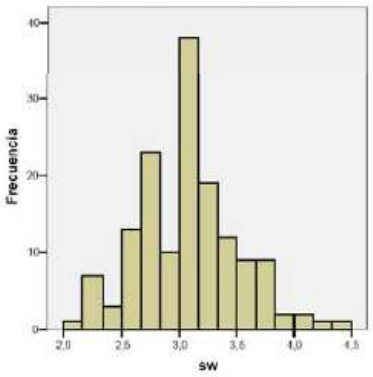
Descripción univariante: longitud del sépalo

SL

N	Válidos	150
	Perdidos	0
Media		5,843
Mediana		5,800
Desv. típ.		,8281
Varianza		,6857
Mínimo		4,3
Máximo		7,9
Percentiles	25	5,100
	50	5,800
	75	6,400



Descripción univariante: anchura del sépalo



SW

N	Válidos	150
	Perdidos	0
Media		3,054
Mediana		3,000
Desv. típ.		,4336
Varianza		,1880
Mínimo		2,0
Máximo		4,4
Percentiles	25	2,800
	50	3,000
	75	3,300

Ejemplo: **Lirios**

Dimensiones del sépalo: covarianza y correlación

Covarianzas

	Longitud del sepalo	Anchura del sepalo
Longitud del sepalo	0.68569351	-0.04243400
Anchura del sepalo	-0.04243400	0.18997942

Correlaciones

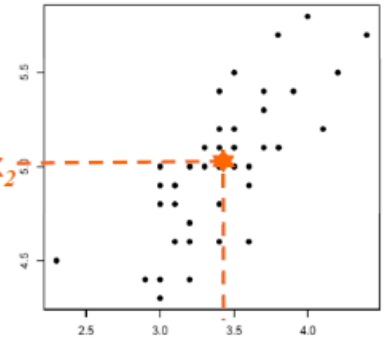
	Longitud del sepalo	Anchura del sepalo
Longitud del sepalo	1.0000000	-0.1175698
Anchura del sepalo	-0.1175698	1.0000000

Resumen numérico de DATOS MULTIVARIANTES

Vector de medias muestrales:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Es un vector de dimensión $p \times 1$ y se interpreta como el centro de los puntos en dimensión p



La MATRIZ DE DATOS CENTRADOS tiene como elementos los datos originales menos la media de la variable que les corresponde. Su vector de medias es siempre un vector de ceros.

Matriz de varianzas-covarianzas:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

S contiene las varianzas de todas las variables en la diagonal, y en el resto de elementos la información de todas las relaciones lineales entre cada dos variables

S siempre es una matriz simétrica y cuadrada ***(p x p)***

La matriz que tiene como elementos las correlaciones en lugar de las covarianzas, y unos en la diagonal, se llama MATRIZ DE CORRELACIONES ***R***

Ejemplo: **Lirios**

Lirios: matrices de covarianzas y de correlaciones

	Longitud.Sepalo	Ancho.Sepalo	Longitud.Petalo	Ancho.Petalo
Longitud.Sepalo	0.68569351	-0.04243400	1.2743154	0.5162707
Ancho.Sepalo	-0.04243400	0.18997942	-0.3296564	-0.1216394
Longitud.Petalo	1.27431544	-0.32965638	3.1162779	1.2956094
Ancho.Petalo	0.51627069	-0.12163937	1.2956094	0.5810063

	Longitud.Sepalo	Ancho.Sepalo	Longitud.Petalo	Ancho.Petalo
Longitud.Sepalo	1.0000000	-0.1175698	0.8717538	0.8179411
Ancho.Sepalo	-0.1175698	1.0000000	-0.4284401	-0.3661259
Longitud.Petalo	0.8717538	-0.4284401	1.0000000	0.9628654
Ancho.Petalo	0.8179411	-0.3661259	0.9628654	1.0000000

Ejemplo: **De salida de SPSS para descriptivos multivariantes**

Estadísticos descriptivos

	N	Media	Desv. típ.	Varianza
BASE	6	,9833	,62102	,386
ALTURA	6	,9833	,62102	,386
N válido (según lista)	6			

$$\bar{\mathbf{x}} = \begin{bmatrix} 0{,}9833 \\ 0{,}9833 \end{bmatrix}$$

Correlaciones

		BASE	ALTURA
BASE	Correlación de Pearson	1	,461
	Covarianza	,386	,178
	N	6	6
ALTURA	Correlación de Pearson	,461	1
	Covarianza	,178	,386
	N	6	6

$$\mathbf{S} = \begin{bmatrix} 0{,}386 & 0{,}178 \\ 0{,}178 & 0{,}386 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}461 \\ 0{,}461 & 1{,}000 \end{bmatrix}$$

Ejemplo: **Gases contaminantes**

Correlaciones

		VIENTO	RADIACIO	CO	NO	NO2	O3	HC
VIENTO	Correlación de Pearson	1	-,101	-,194	-,270	-,110	-,254	,156
	Sig. (bilateral)	.	,523	,219	,084	,489	,105	,324
	N	42	42	42	42	42	42	42
RADIACIO	Correlación de Pearson	-,101	1	,183	-,074	,116	,319*	,052
	Sig. (bilateral)	,523	.	,247	,643	,465	,039	,744
	N	42	42	42	42	42	42	42
CO	Correlación de Pearson	-,194	,183	1	,502**	,557**	,411**	,166
	Sig. (bilateral)	,219	,247	.	,001	,000	,007	,293
	N	42	42	42	42	42	42	42
NO	Correlación de Pearson	-,270	-,074	,502**	1	,297	-,134	,235
	Sig. (bilateral)	,084	,643	,001	.	,056	,398	,135
	N	42	42	42	42	42	42	42
NO2	Correlación de Pearson	-,110	,116	,557**	,297	1	,167	,448**
	Sig. (bilateral)	,489	,465	,000	,056	.	,292	,003
	N	42	42	42	42	42	42	42
O3	Correlación de Pearson	-,254	,319*	,411**	-,134	,167	1	,154
	Sig. (bilateral)	,105	,039	,007	,398	,292	.	,329
	N	42	42	42	42	42	42	42
HC	Correlación de Pearson	,156	,052	,166	,235	,448**	,154	1
	Sig. (bilateral)	,324	,744	,293	,135	,003	,329	.
	N	42	42	42	42	42	42	42

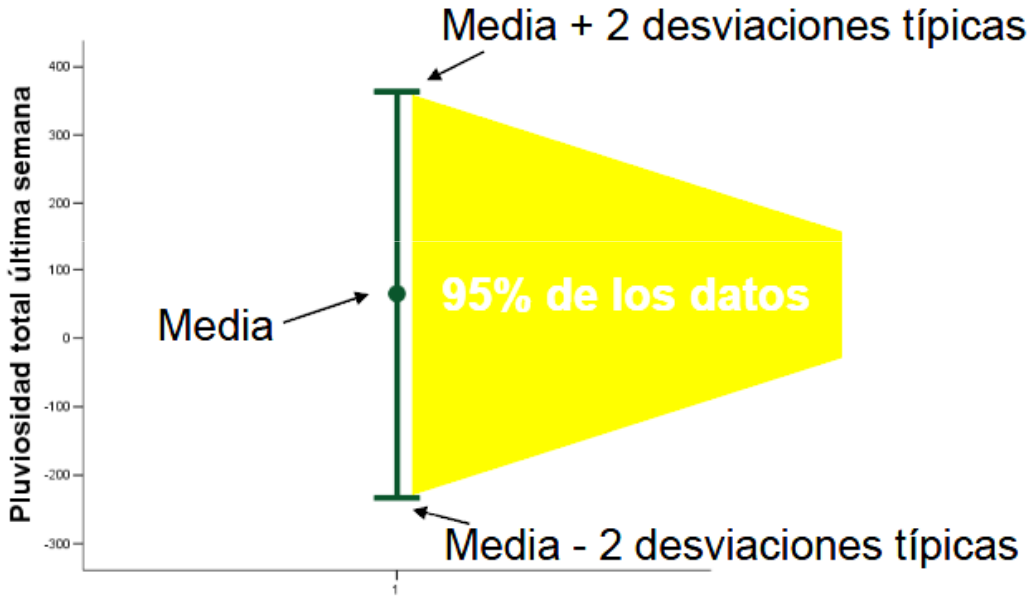
*. La correlación es significante al nivel 0,05 (bilateral).

**. La correlación es significativa al nivel 0,01 (bilateral).

- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

Resumen gráfico de DATOS UNIVARIANTES

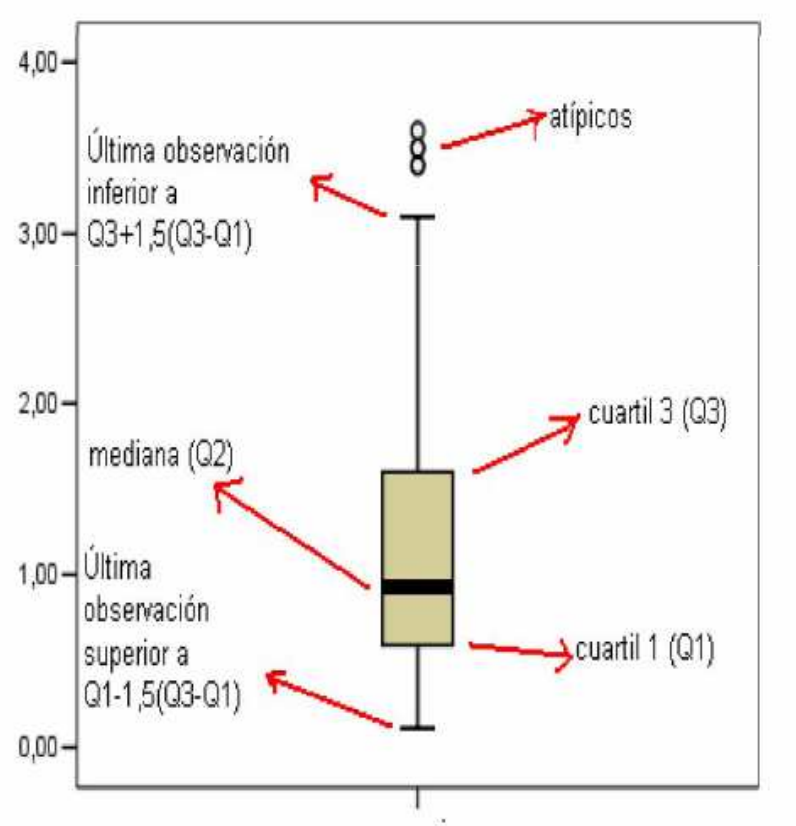
Barras de error



La desviación típica es muy sensible a los datos atípicos
Siempre es un gráfico simétrico
Es una buena herramienta cuando los datos son normales

Resumen gráfico de DATOS UNIVARIANTES

Diagrama de cajas o Boxplot



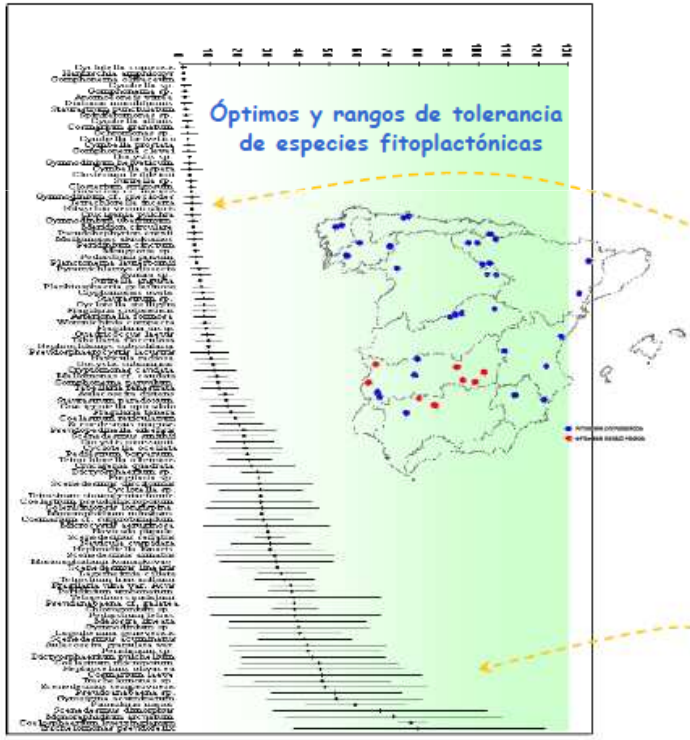
1. Ordenar la muestra
2. Calcular la mediana, el primer y el tercer cuartil
3. Calcular el rango intercuartílico

Resumen gráfico de DATOS UNIVARIANTES

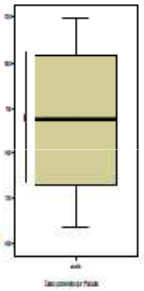
¿ Box-plot o barra de error ?

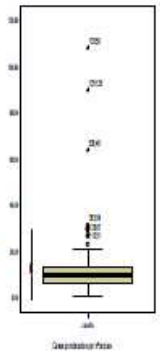
UTILIZACIÓN DEL FITOPLACTON COMO INDICADOR BIOLÓGICO PARA LA EVALUACIÓN DE LA EUTROFIZACIÓN EN LOS EMBALSES ESPAÑOLES

C. NUÑO, C. DE HOYOS, A. JUSTEL



En la mayor parte de los casos la barra de error y el box-plot dan el mismo resultado





Algunas especies no son consideradas indicadoras debido al amplio rango de tolerancia (barra de error)

Si utilizamos la mediana y el rango intercuartílico pasan a ser especies indicadoras

Resumen gráfico de DATOS UNIVARIANTES

Queremos herramientas gráficas que nos ayuden en el estudio de las **relaciones entre las variables** (forma, fuerza, etc.), a **identificar grupos**, y a detectar posibles **datos atípicos**

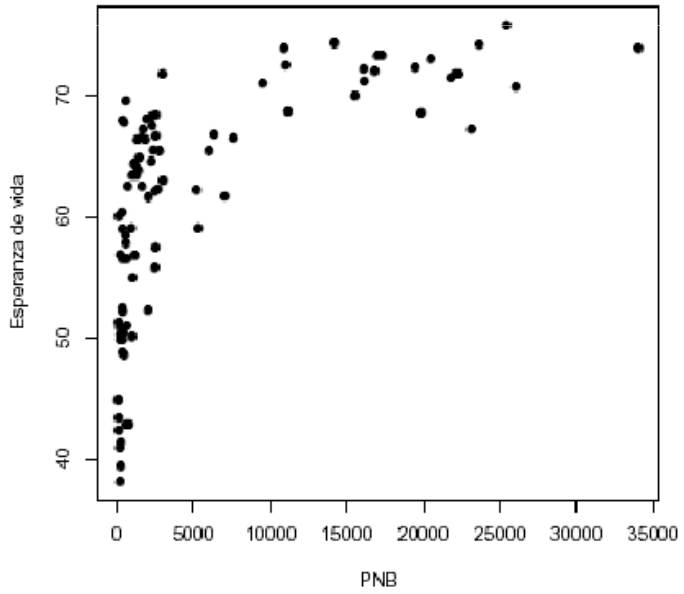
- Matriz de diagramas de dispersión
- Diagramas de cajas múltiples (box-plot múltiples)
- Gráficos de estrellas (o de caras)

Diagramas de dispersión

Relación entre conjuntos de datos

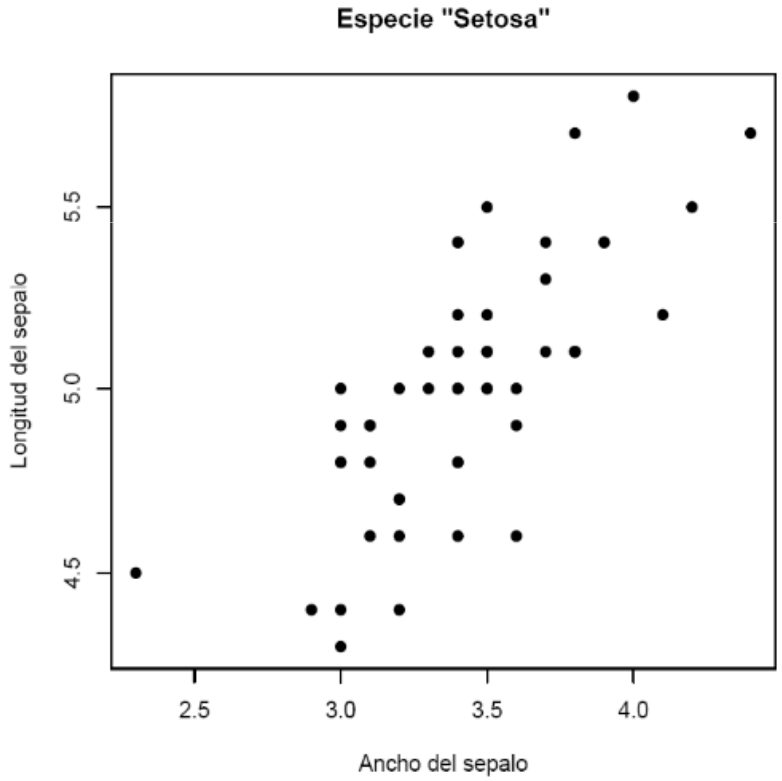
Los datos corresponden a distintas mediciones en los mismos individuos

¿Influyen las medidas $(x_1, \ldots, x_n)$ en las medidas $(y_1, \ldots, y_n)$?



Ejemplo: **Lirios**

Dimensiones del sépalo de la especie *setosa*



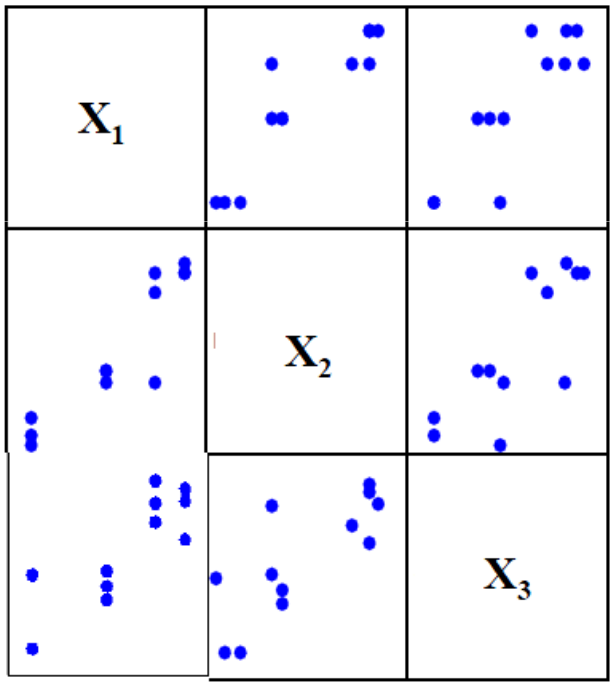
Matriz de diagramas de dispersión

Se construye una cuadrícula con tantas filas y columnas como variables.

En la diagonal se da información de cada una de las variables.

En el resto de casillas se construyen los gráficos de dispersión entre todos los pares de variables.

- Todos los gráficos de la misma **FILA** comparten la misma variable en el **EJE VERTICAL** (la que se indique en la diagonal)
- Todos los gráficos de la misma **COLUMNA** comparten la misma variable en el **EJE HORIZONTAL** (la que se indique en la diagonal)

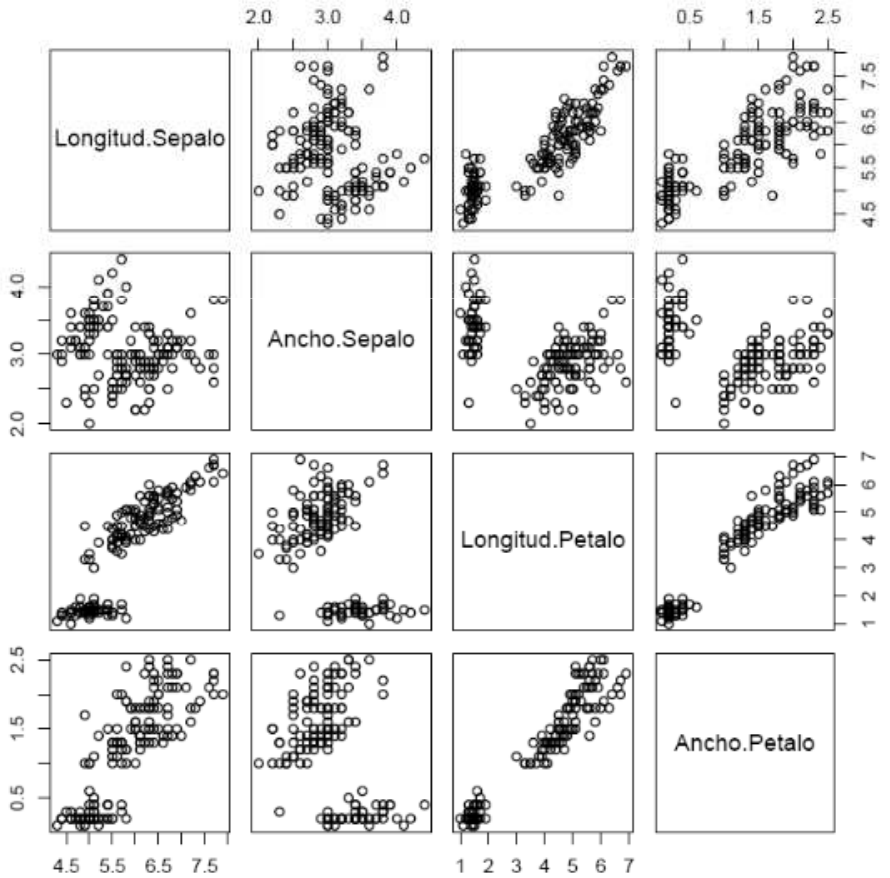


Informa de cómo son las relaciones entre variables, pero sólo dos a dos, no se puede saber como son todas las relaciones.

La matriz es simétrica, la diagonal es como un espejo.

Matriz de diagramas de dispersión

Ejemplo: Lirios



Matriz de diagramas de dispersión

Ejemplo: Lirios

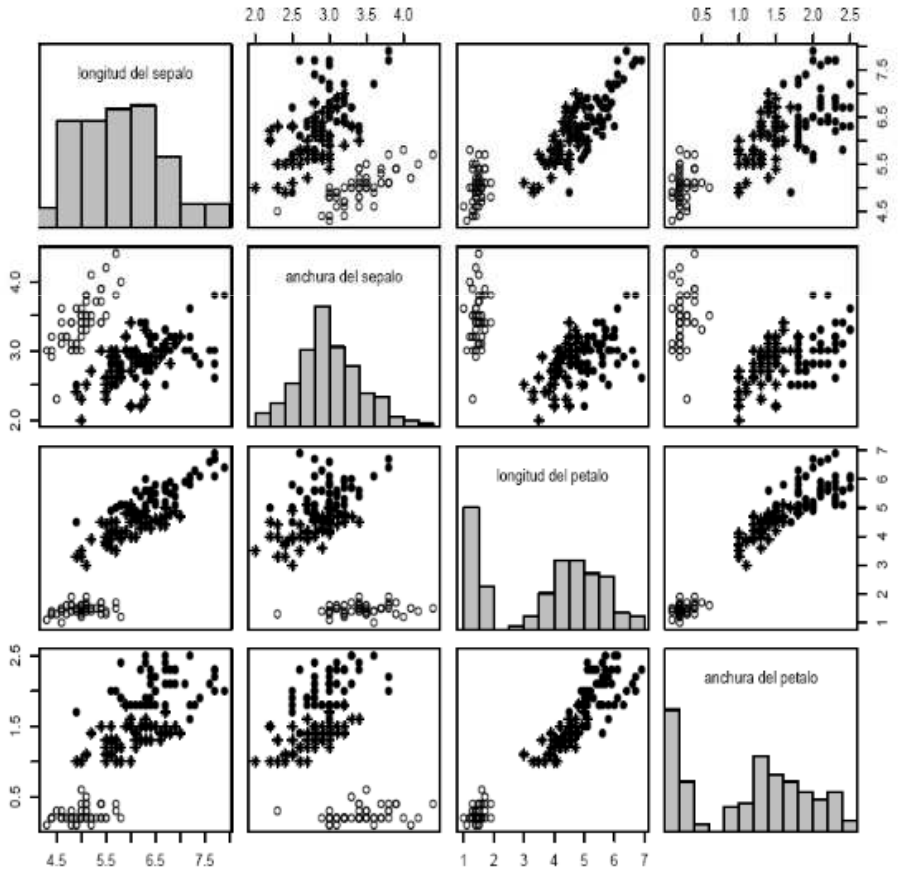


Diagrama de cajas múltiple

Se emplean para comparar:

- Variables en distintos grupos.
- Variables sólo cuando las unidades de medida sean "compatibles".

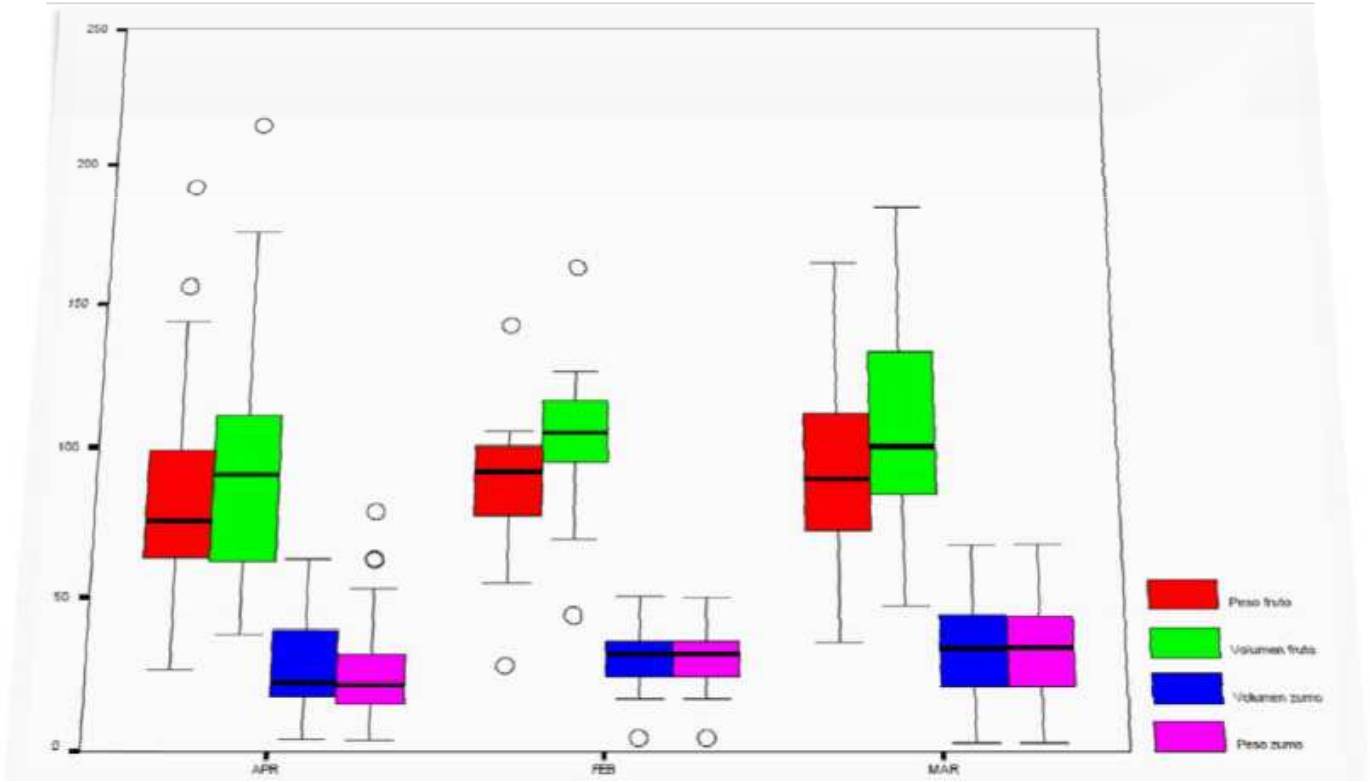
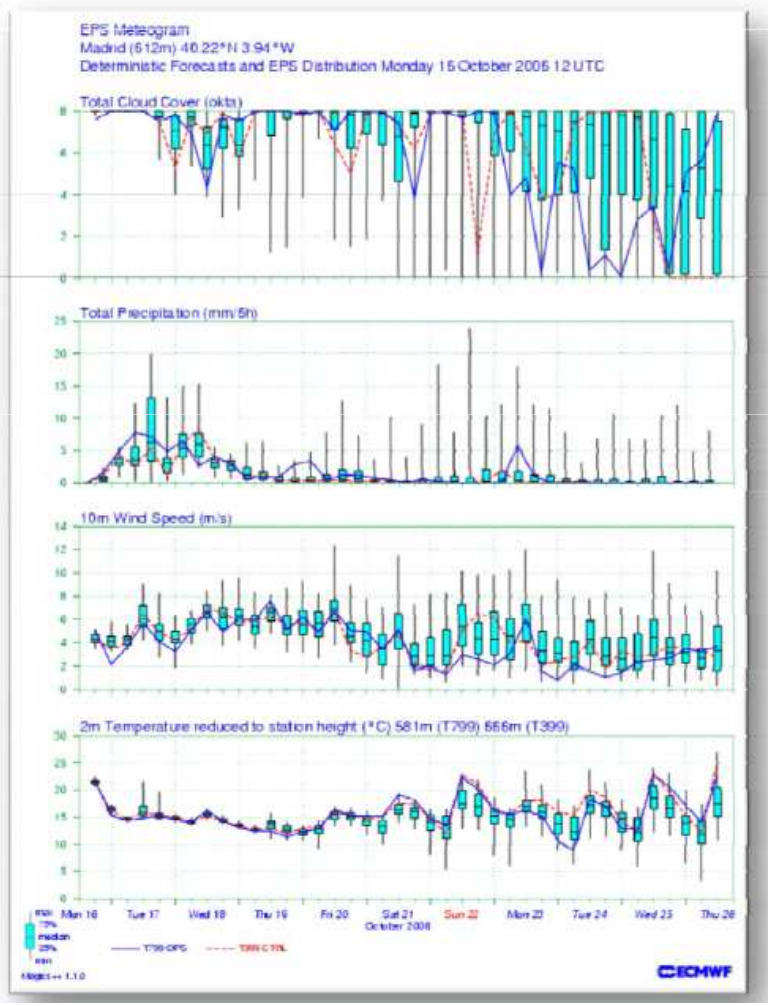


Diagrama de cajas múltiple

Predicción meteorológica:

En cada diagrama de cajas múltiple se muestra para una característica meteorológica, las predicciones con distintos modelos (individuos) en distintos días (variables)



Ejemplo: **Lirios**

Dimensiones del sépalo: diagrama de cajas

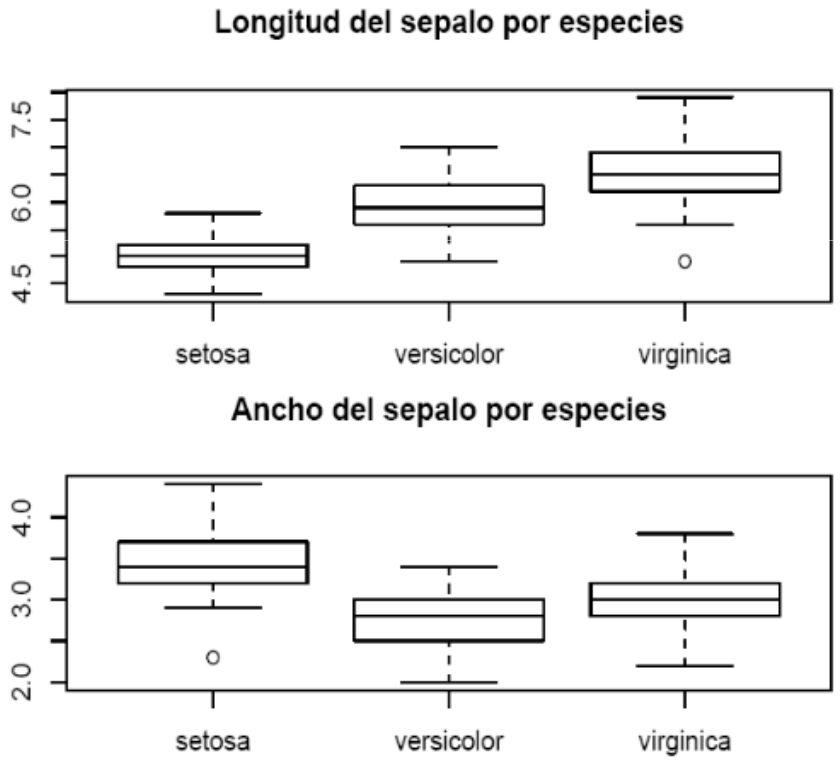
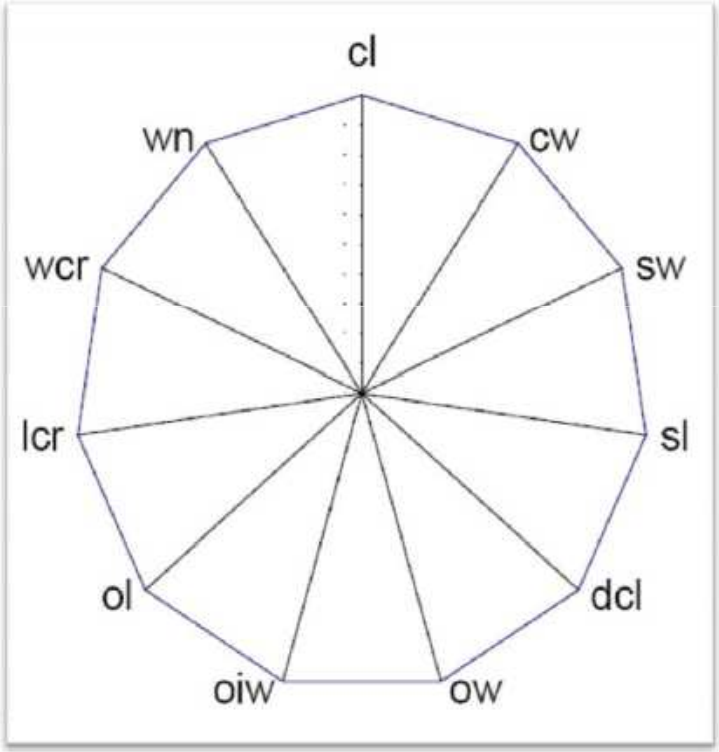


Gráfico de estrellas



Cada dato se representa mediante una estrella, que tendrá tantos rayos o ejes como variables queramos representar.

En cada rayo se representa el valor de una variable. En todas las estrellas se usa el mismo rayo para la misma variable.

El rayo j en la estrella del dato i dependerá del valor (absoluto o relativo) de x_{ij}.

Gráfico de estrellas

Ejemplo: **Medidas de cráneos de cocodrilos** (alligator.txt)

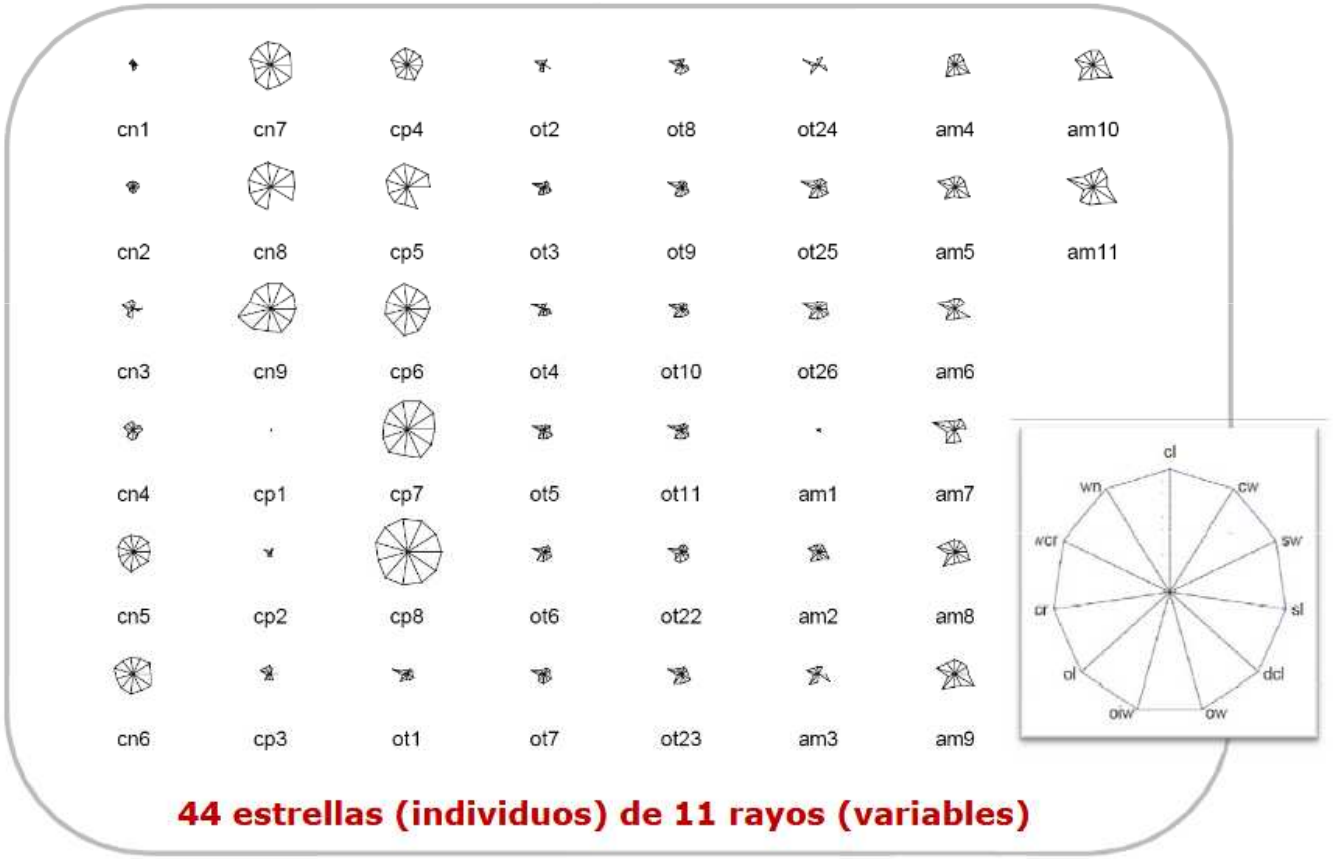
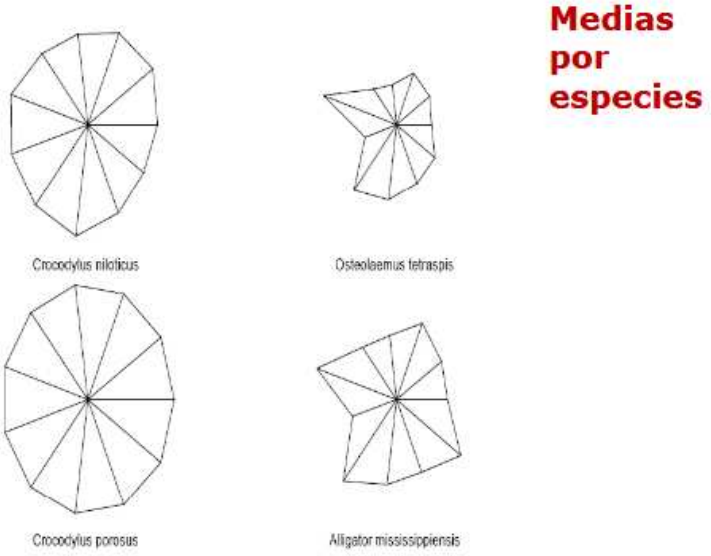


Gráfico de estrellas

Ejemplo: **Medidas de cráneos de cocodrilos** (alligator.txt)



Conclusión: Hay cocodrilos grandes y pequeños de todas las especies, así que el tamaño no sirve para distinguir unas especies de otras. Usando todas las medidas de los cráneos a la vez parece que podremos distinguir bastante bien si un cocodrilo es de la especies cn y cp o de las op y am, pero no podremos distinguir bien entre las cuatro.

Gráfico de estrellas

Ejemplo: **Lirios**

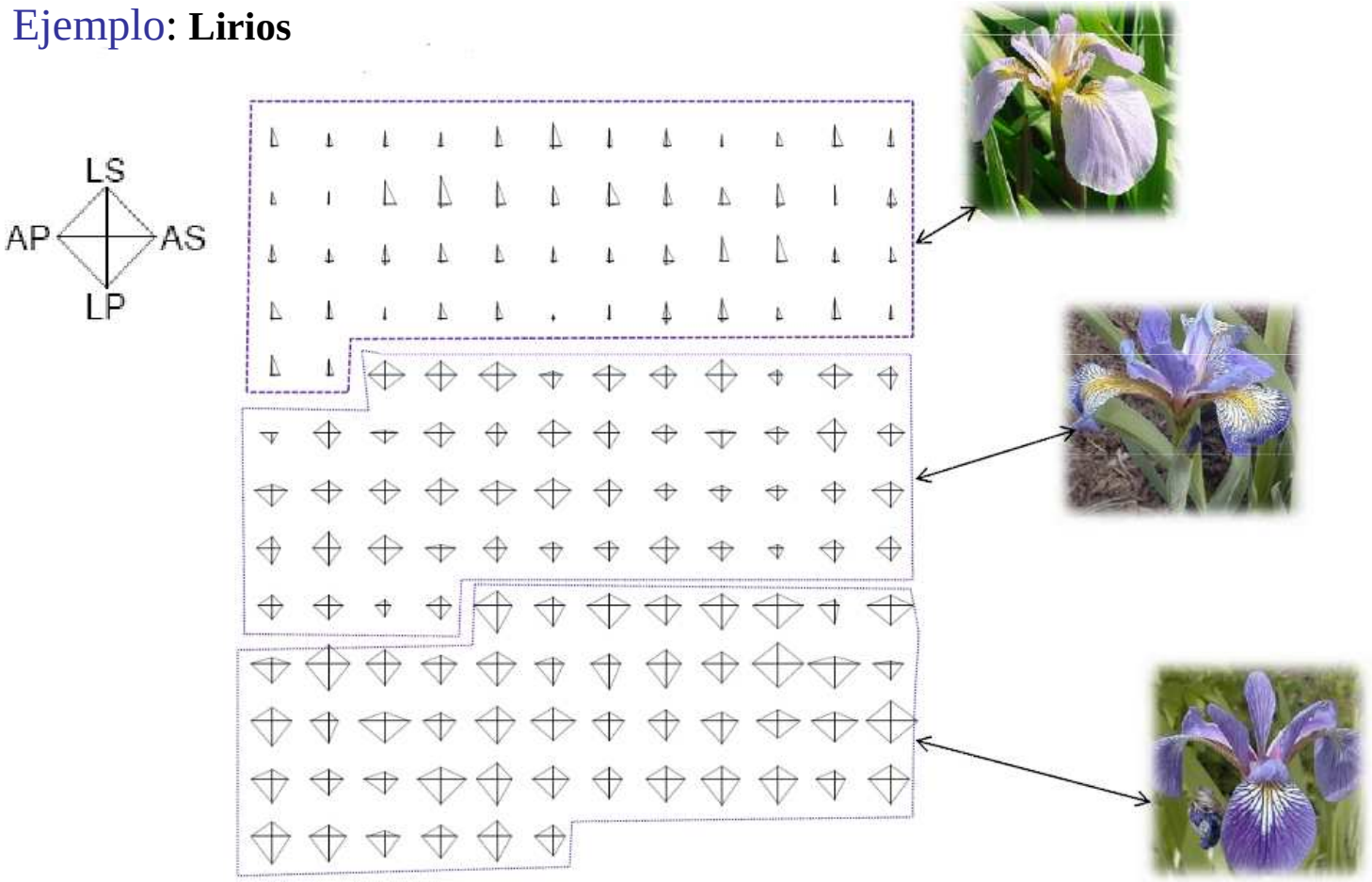
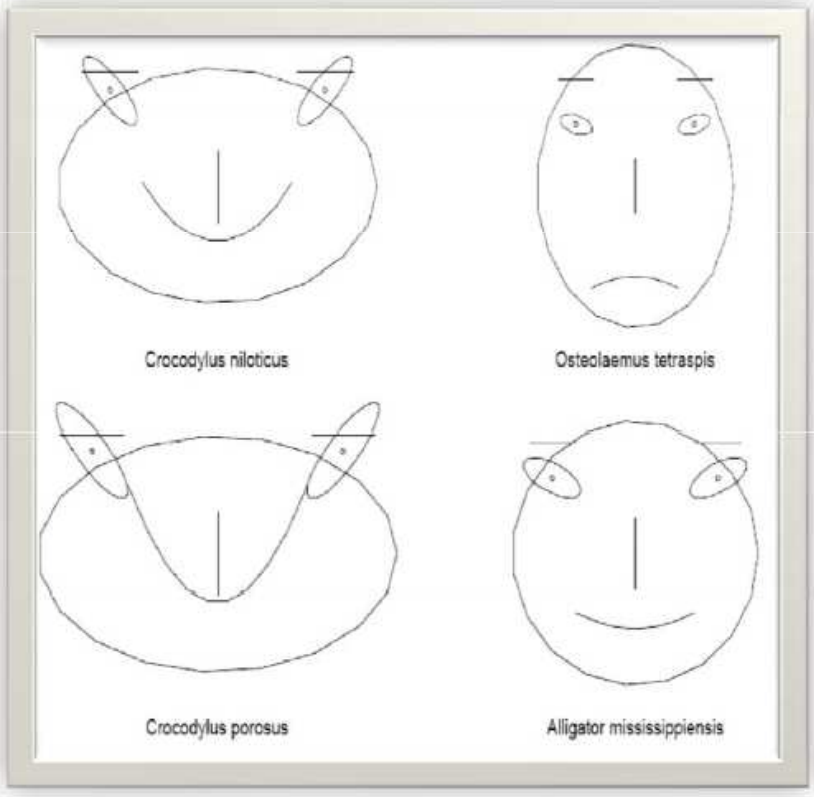


Gráfico de caras

Es como un gráfico de estrellas, pero cada individuo ahora se representa en una CARA y las variables en los rasgos físicos.

Variables en

1.- Tamaño de la cara,
2.- Forma de la cara,
3.- Tamaño de la nariz,
4.- Posición de la boca,
5.- Tamaño de la sonrisa
6.- Grosor de la boca,
7.- Posición de los ojos,
8.- Separación de los ojos,
9.- Inclinación de los ojos,
10.- Tamaño de los ojos
11.- Forma de los ojos



- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

Estandarización de los datos

Cuando queremos comparar conjuntos de datos buscamos una manera de transformar las observaciones que no dependa de la magnitud de los datos, ni de las unidades de medida, y que tenga en cuenta cómo de dispersos están.

Estandarización univariante

$$z_i = D^{-1/2}(x_i - \bar{x})$$

Vector de dimensión ***p*** con los valores estandarizados (o tipificados) del individuo ***i***

$$\mathbf{D}^{-1/2} = \begin{bmatrix} s_1^{-1} & 0 & \cdots & 0 \\ 0 & s_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p^{-1} \end{bmatrix}$$

Propiedades:

- El vector de medias de los datos estandarizados $z_1, \ldots, z_n$ es un vector de ceros.
- La matriz de covarianzas de los datos estandarizados $z_1, \ldots, z_n$ es la matriz de correlaciones de los datos $x_1, \ldots, x_n$

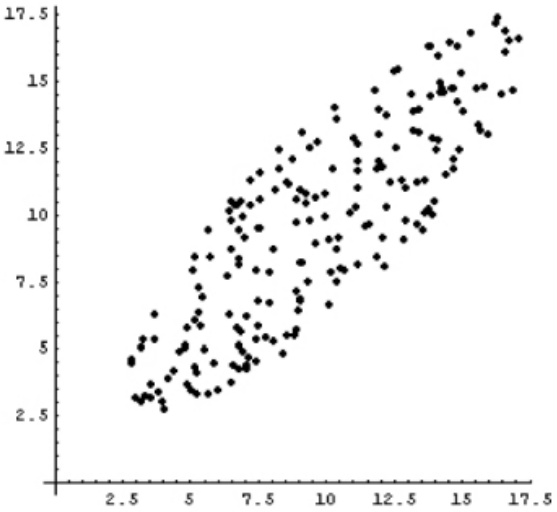
Estandarización de los datos

Ejemplo:

Datos: x

$$\overline{\mathbf{x}} = \begin{bmatrix} 10 \\ 10 \end{bmatrix} \quad \mathbf{S}_{\mathrm{x}} = \begin{bmatrix} 14{,}8 & 12{,}8 \\ 12{,}8 & 15{,}7 \end{bmatrix}$$

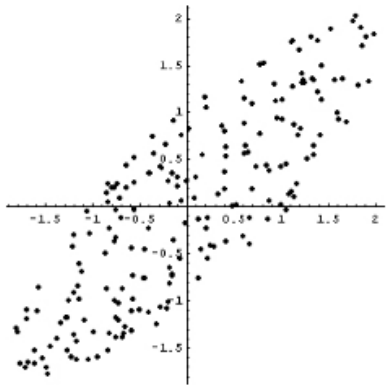
$$\mathbf{R}_{\mathrm{x}} = \begin{bmatrix} 1 & 0{,}84 \\ 0{,}84 & 1 \end{bmatrix}$$



Estandarización univariante: $\mathbf{y}_u$

$$\overline{\mathbf{y}}_u = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{S}_{\mathrm{y}_u} = \mathbf{R}_{\mathrm{x}} = \begin{bmatrix} 1 & 0{,}84 \\ 0{,}84 & 1 \end{bmatrix}$$



- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

Distancias ESTADÍSTICAS

Una distancia en d en $\mathbb{R}^p$ es una aplicación

$$d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+,$$

con las propiedades siguientes:

(a) $d(\mathbf{x}, \mathbf{y}) \geq 0,\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$

(b) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$

(c) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

(d) $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ **(Desigualdad triangular)**

Distancias ESTADÍSTICAS

Distancia euclídea

$$d_E(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{p}(x_i - y_i)^2\right)^{1/2}$$

Distancia rectangular o de Manhattan

$$d_R(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}|x_i - y_i|$$

Distancia de Minkowski

$$d_r(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{p}|x_i - y_i|^r\right)^{1/r}$$

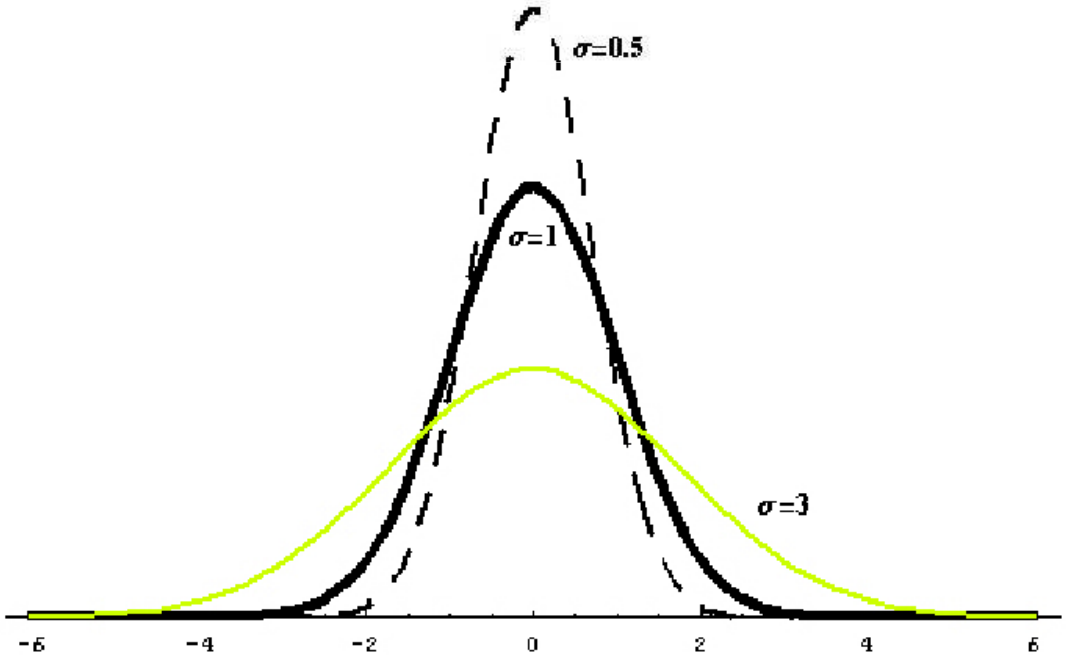
Distancia de Mahalanobis

$$d_M(x,\overline{x}) = (x-\overline{x})' S^{-1}(x-\overline{x})$$

Distribución normal univariante

Una variable aleatoria X se dice que tiene distribución normal de media μ y varianza σ^2, y lo escribimos $X \in \mathcal{N}_1(\mu; \sigma^2)$, si tiene densidad dada por

$$f(x) = (\sigma^2)^{-1/2}(2\pi)^{-1/2} \exp\left((-1/2)(\sigma^2)^{-1}(x-\mu)^2\right)$$



Distribución normal multivariante

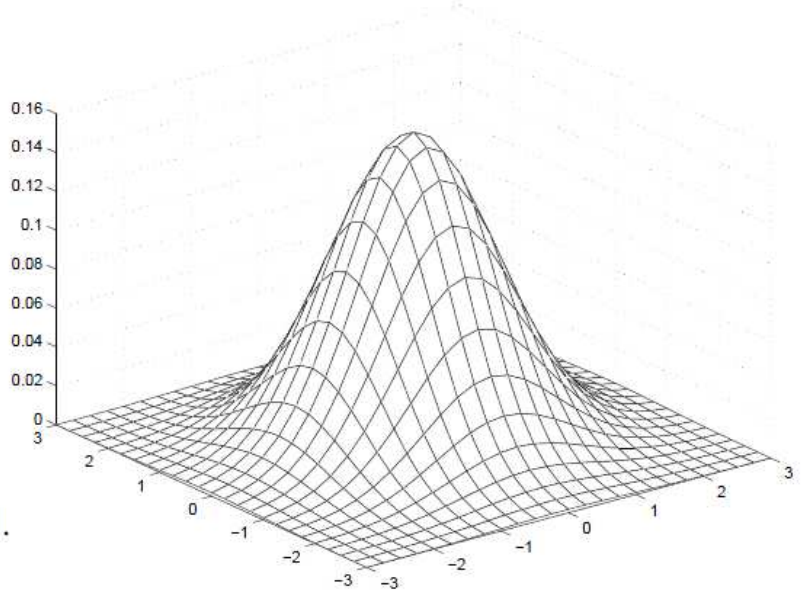
Sea $\mathbf{x}' = [x_1\, x_2 \cdots x_p]$ un vector aleatorio p-dimensional, diremos que el vector $\mathbf{x}$ sigue una distribución normal multivariante si su función de densidad es:

$$f(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\left(-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

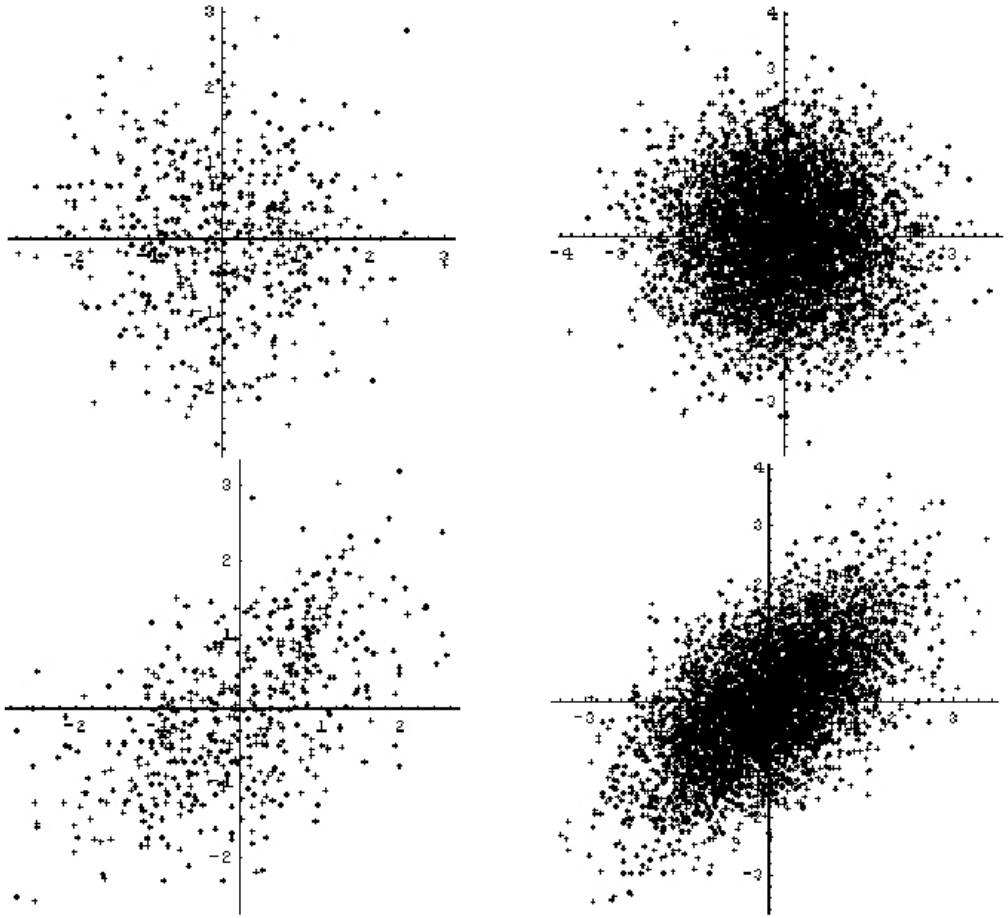
donde $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$,

y $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$.

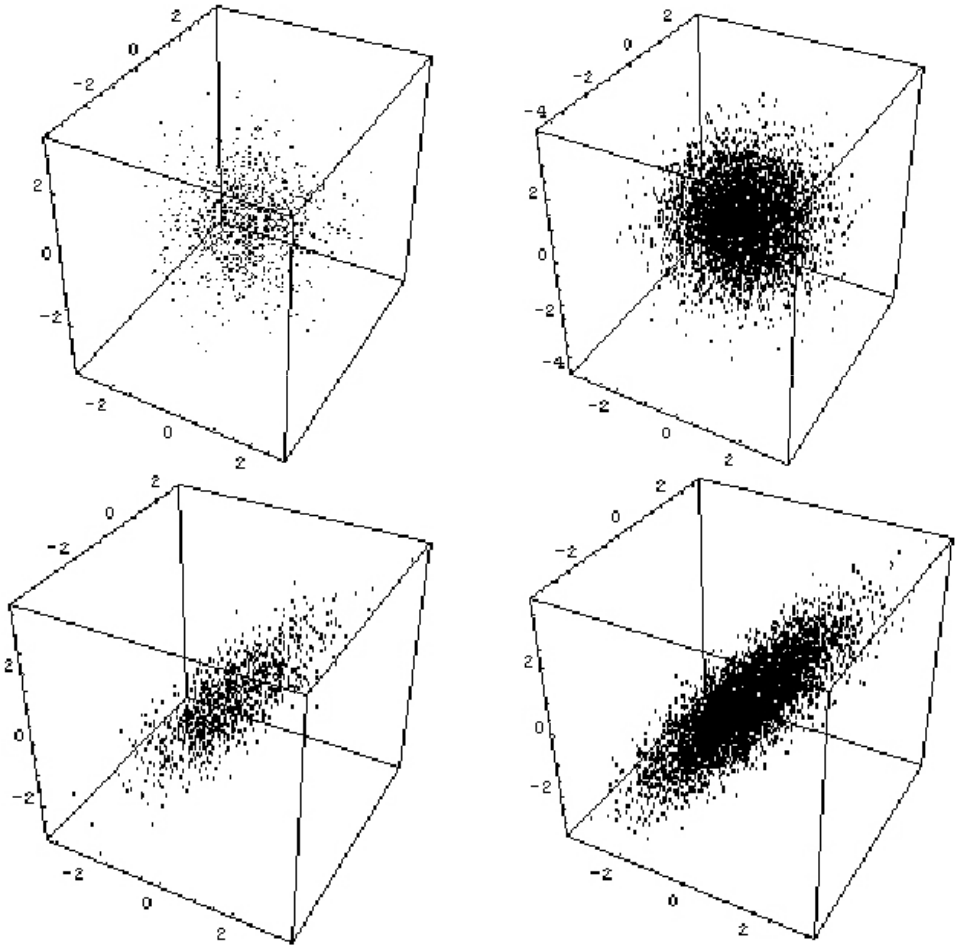
En tal caso escribiremos $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.



Datos normales $p = 2$



Datos normales $p = 3$



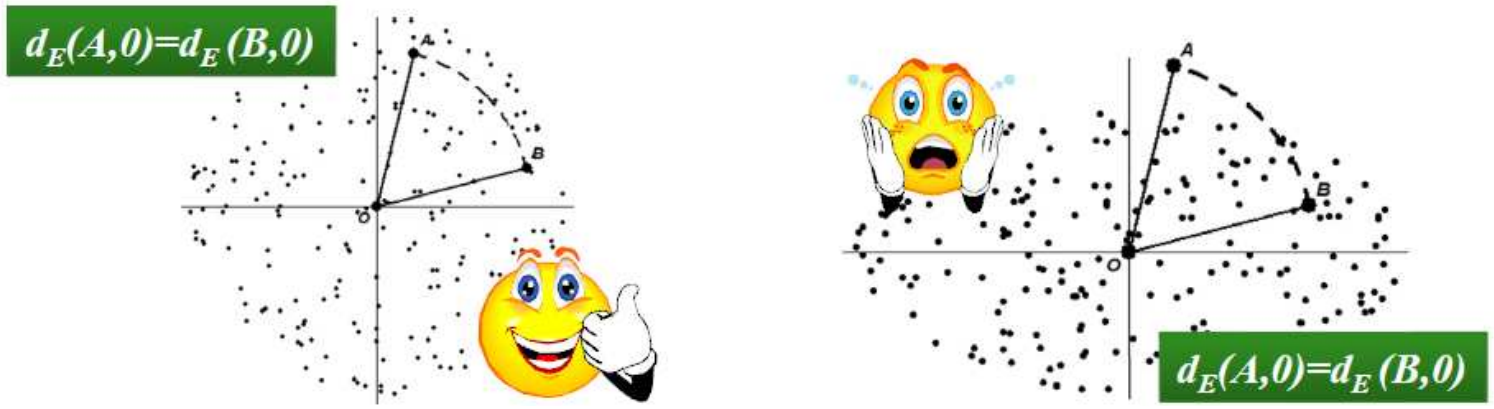
Distancias ESTADÍSTICAS

La distancia euclídea también se puede escribir como un producto de dos vectores. Por ejemplo, si calculamos la distancia de un individuo a la media de todos, la distancia euclídea es:

$$d_E(x_i,\bar{x})^2 = (x_i-\bar{x})'(x_i-\bar{x})$$

$$= (x_{i1}-\bar{x}_1,\ldots,x_{ip}-\bar{x}_p)\begin{pmatrix} x_{i1}-\bar{x}_1 \\ \ldots \\ x_{ip}-\bar{x}_p \end{pmatrix} = (x_{i1}-\bar{x}_1)^2+\ldots+(x_{ip}-\bar{x}_p)^2$$

Problema de la distancia euclídea:
No tiene en cuenta la variabilidad



Distancias ESTADÍSTICAS

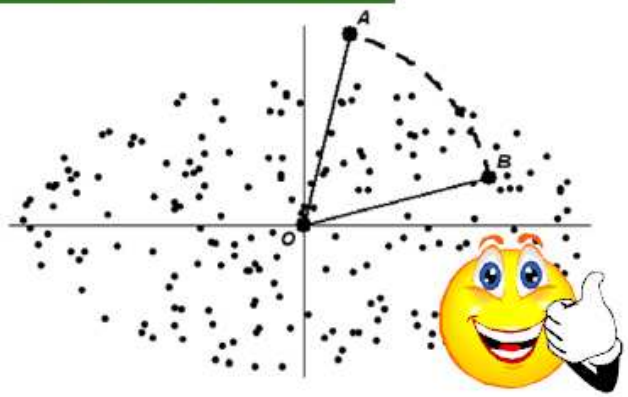
Para resolver este problema podemos estandarizar los datos y luego calcular la distancia euclídea.

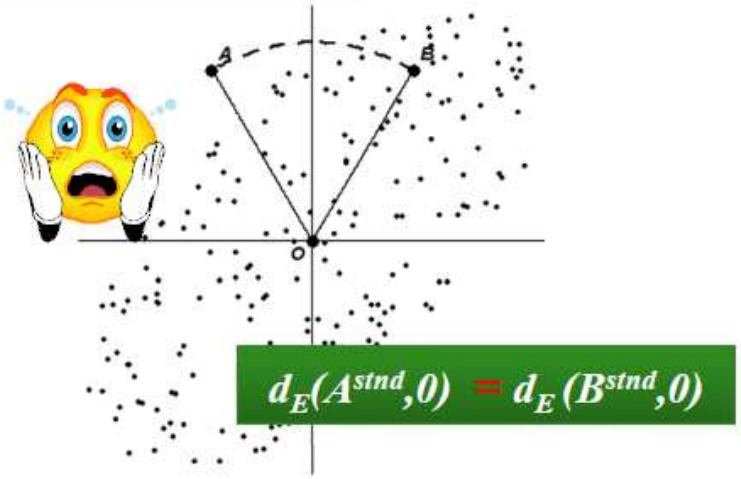
$$d_E(x_i^{stnd},0)^2 = (x_i - \bar{x})' D^{-1} (x_i - \bar{x})$$

$$= ((x_i - \bar{x})' D^{-1/2})(D^{-1/2}(x_i - \bar{x})) = \left(\frac{x_{i1} - \bar{x}_1}{s_1}\right)^2 + \ldots + \left(\frac{x_{ip} - \bar{x}_p}{s_p}\right)^2$$

Problema de la distancia euclídea:
No tiene en cuenta la correlación

$d_E(A^{stnd},0) > d_E(B^{stnd},0)$





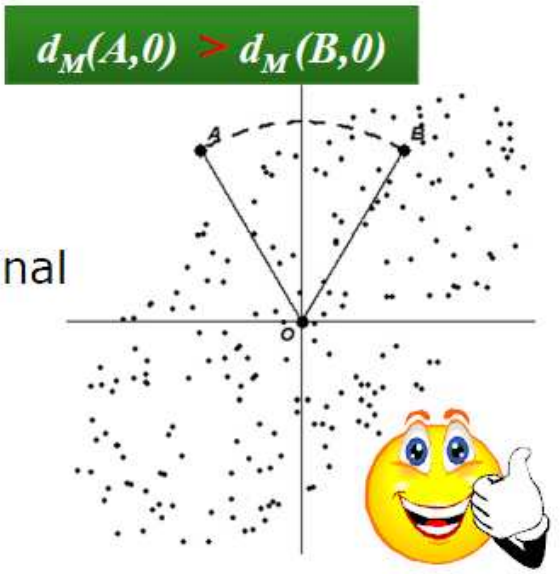
$d_E(A^{stnd},0) = d_E(B^{stnd},0)$

Distancias ESTADÍSTICAS

Distancia de Mahalanobis

$$d_M(x,\bar{x}) = (x-\bar{x})' S^{-1} (x-\bar{x})$$

- Consiste en sustituir la matriz D que sólo tiene información de las varianzas por la matriz **S** de varianzas-covarianzas
- Geométricamente equivale a girar la nube de puntos hasta eliminar las correlaciones y luego calcular la distancia para los datos estandarizados
- La distancia de Mahalanobis es adimensional
- Es la distancia más estadística, la que tiene en cuenta la relación estadística entre las variables.



- **Matriz de datos.**
- **Vector de medias y matriz de covarianzas.**
- **Representación gráfica de los datos: matriz de diagramas de dispersión, diagramas de estrellas y de caras.**
- **Estandarización de datos multivariantes.**
- **Distancias estadísticas**
- **Proyecciones y combinaciones lineales.**

COMBINACIÓN LINEAL DE VARIABLES y PROYECCIONES

Una forma sencilla de resumir varias variables para que la información sea mas *manejable* es construir "convenientemente" una nueva variable, que sea univariante, y una **COMBINACIÓN LINEAL** de todas ellas.

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$$

Combinación lineal de las variables $x_1,\ldots,x_p$

Para cada individuo toma un único valor, es univariante

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip}$$

Se puede escribir como el producto escalar entre los vectores

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} \text{ y } x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

$$\rightarrow \quad y = a' x$$

Cuando el vector $\boldsymbol{a}$ *tiene modulo uno* ($\|a\|^2 = a'a = 1$), *entonces* y_i *coincide con el valor de la proyección del vector* x_i *a lo largo de la dirección indicada por* $\boldsymbol{a}$

COMBINACIÓN LINEAL DE VARIABLES y PROYECCIONES

La manera más rápida de calcular la combinación lineal para todos los datos a la vez es multiplicar la matriz de datos X por el vector a que contiene los coeficientes de cada variable: $\boldsymbol{y=Xa}$

Ejemplo:

$$\mathbf{y} = \mathbf{Xa} = \begin{bmatrix} 2{,}0 & 2{,}0 \\ 1{,}5 & 0{,}5 \\ 0{,}7 & 0{,}5 \\ 0{,}5 & 1{,}5 \\ 0{,}5 & 0{,}7 \\ 0{,}7 & 0{,}7 \end{bmatrix} \begin{bmatrix} 2{,}0 \\ 2{,}0 \end{bmatrix} = \begin{bmatrix} 8{,}00 \\ 4{,}00 \\ 2{,}40 \\ 4{,}00 \\ 2{,}40 \\ 2{,}80 \end{bmatrix}$$