Tema 1
Análisis exploratorio de datos multivariantes

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Gorriones

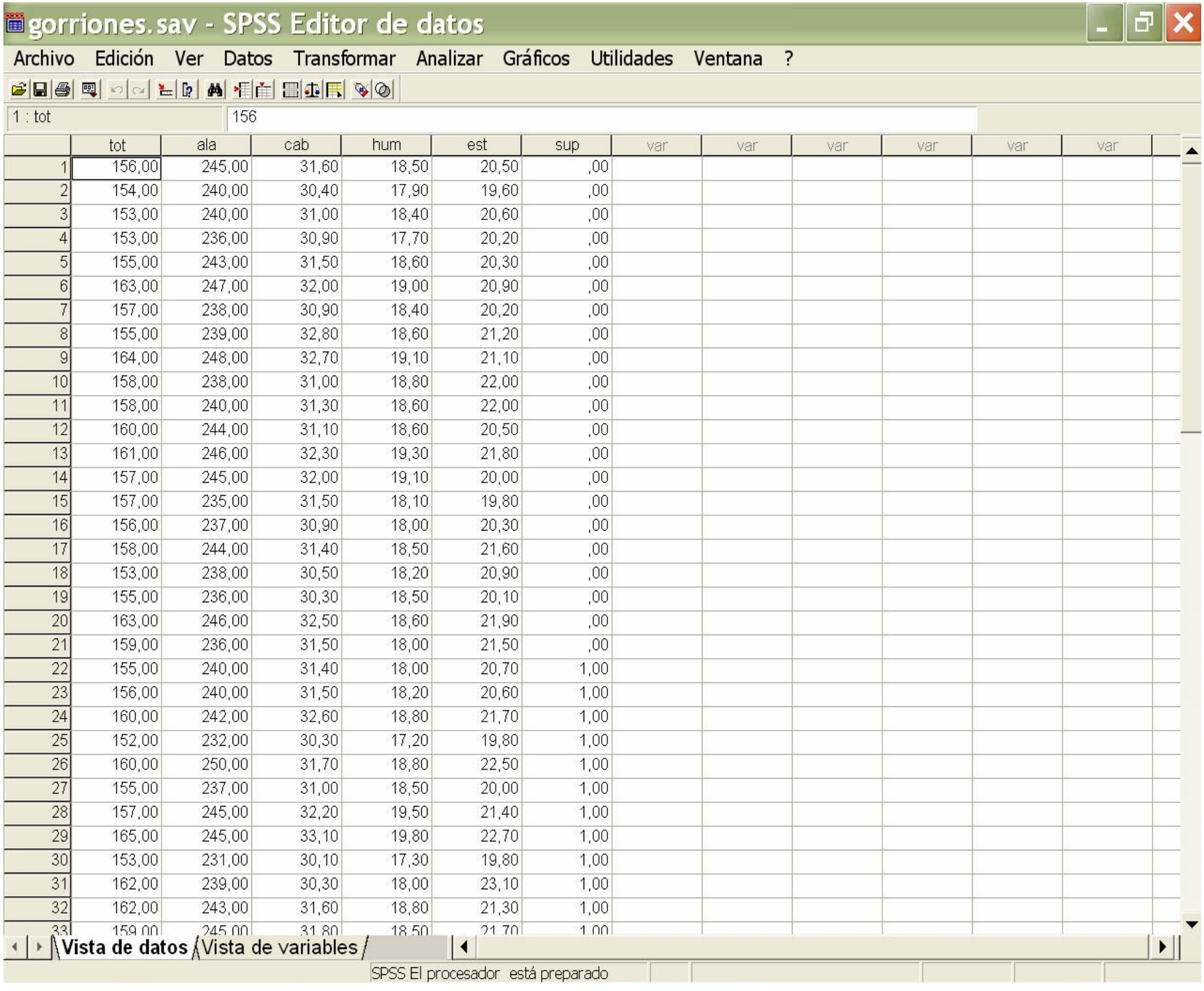
- Tras una fuerte tormenta en febrero de 1898, un grupo de gorriones moribundos fueron llevados a la Universidad Brown (Rhode Island).
- Alrededor de la mitad de los gorriones murieron. Se consideró la situación como una oportunidad de estudiar el efecto de la selección natural sobre los pájaros.
- Se tomaron diversas medidas morfológicas, de las que se incluyen 5 en el fichero.

Variables

Nombre variable	Descripción
`TOT`	Longitud total
`ALA`	Extensión de las alas
`CAB`	Longitud del pico y la cabeza
`HUM`	Longitud del húmero
`EST`	Longitud del esternón

Observaciones:

- Todas las variables se miden en mm.
- El fichero contiene datos de 49 gorriones.
- Los 21 primeros gorriones fueron los supervivientes.



	tot	ala	cab	hum	est	sup	var	var	var	var	var	var
1	156,00	245,00	31,60	18,50	20,50	,00						
2	154,00	240,00	30,40	17,90	19,60	,00						
3	153,00	240,00	31,00	18,40	20,60	,00						
4	153,00	236,00	30,90	17,70	20,20	,00						
5	155,00	243,00	31,50	18,60	20,30	,00						
6	163,00	247,00	32,00	19,00	20,90	,00						
7	157,00	238,00	30,90	18,40	20,20	,00						
8	155,00	239,00	32,80	18,60	21,20	,00						
9	164,00	248,00	32,70	19,10	21,10	,00						
10	158,00	238,00	31,00	18,80	22,00	,00						
11	158,00	240,00	31,30	18,60	22,00	,00						
12	160,00	244,00	31,10	18,60	20,50	,00						
13	161,00	246,00	32,30	19,30	21,80	,00						
14	157,00	245,00	32,00	19,10	20,00	,00						
15	157,00	235,00	31,50	18,10	19,80	,00						
16	156,00	237,00	30,90	18,00	20,30	,00						
17	158,00	244,00	31,40	18,50	21,60	,00						
18	153,00	238,00	30,50	18,20	20,90	,00						
19	155,00	236,00	30,30	18,50	20,10	,00						
20	163,00	246,00	32,50	18,60	21,90	,00						
21	159,00	236,00	31,50	18,00	21,50	,00						
22	155,00	240,00	31,40	18,00	20,70	1,00						
23	156,00	240,00	31,50	18,20	20,60	1,00						
24	160,00	242,00	32,60	18,80	21,70	1,00						
25	152,00	232,00	30,30	17,20	19,80	1,00						
26	160,00	250,00	31,70	18,80	22,50	1,00						
27	155,00	237,00	31,00	18,50	20,00	1,00						
28	157,00	245,00	32,20	19,50	21,40	1,00						
29	165,00	245,00	33,10	19,80	22,70	1,00						
30	153,00	231,00	30,10	17,30	19,80	1,00						
31	162,00	239,00	30,30	18,00	23,10	1,00						
32	162,00	243,00	31,60	18,80	21,30	1,00						
33	159,00	245,00	31,80	18,50	21,70	1,00						

Problemas de interés relacionados con estos datos

- ¿Están las variables relacionadas? Al aumentar una, ¿tienden a aumentar los valores de las otras?
- ¿Hay diferencias significativas entre las observaciones correspondientes a los supervivientes y a los que no sobrevivieron?
- Si la respuesta es afirmativa, ¿es posible construir una función de las variables que separe bien los dos grupos?
- ¿Es posible reducir la dimensión de los datos sin perder mucha información?

Temario

- Análisis exploratorio de datos multivariantes
  - Descripción numérica
  - Descripción gráfica
- Técnicas de análisis discriminante
  - Discriminación lineal de Fisher
- Técnicas de agrupación
  - Métodos jerárquicos
  - Métodos por división
- Técnicas de reducción de la dimensión
  - Análisis de componentes principales
  - Análisis factorial

Bibliografía básica

- Johnson, R.A. y Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* Prentice–Hall.
- Lattin, J.M., Carroll, J.D. y Green, P.E. (2003). *Analyzing multivariate data.* Thomson Brooks/Cole.
- Peña, D. (2002). *Análisis de datos multivariantes.* McGraw Hill.

Una bibliografía más amplia puede encontrarse en el programa de la página web de la asignatura

Lirios

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

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)

Problemas de interés relacionados con estos datos

- ¿Están las variables relacionadas? Al aumentar una, ¿tienden a aumentar los valores de las otras?
- ¿Hay diferencias significativas entre las observaciones correspondientes a cada una de las especies?
- Si la respuesta es afirmativa, ¿es posible construir una función de las variables que separe bien los tres grupos?
- ¿Es posible reducir la dimensión de los datos sin perder mucha información?

Liga española de fútbol 2005-2006

Equipo	G	P	GF	GC
Barcelona	25	6	80	35
RMadrid	20	8	70	40
Valencia	19	7	58	33
Osasuna	21	12	49	43
Sevilla	20	10	54	39
Celta	20	14	45	33
Villarreal	14	9	50	39
Deportivo	15	13	47	45
Getafe	15	14	54	49
AtMadrid	13	12	45	37
Zaragoza	10	12	46	51
AthBilbao	11	15	40	46
Mallorca	10	15	37	51
Betis	10	16	34	51
Espanyol	10	17	36	56
RSociedad	11	20	48	65
Racing	9	16	36	49
Alavés	9	17	35	54
Cádiz	8	18	36	52
Málaga	5	24	36	68

Variables

Nombre variable	Descripción
`Equipo`	Nombre del equipo
`G`	Número de partidos ganados
`P`	Número de partidos perdidos
`GF`	Goles a favor
`GC`	Goles en contra

Observaciones:

- El número de partidos empatados y el número de puntos se han omitido puesto que son variables redundantes.

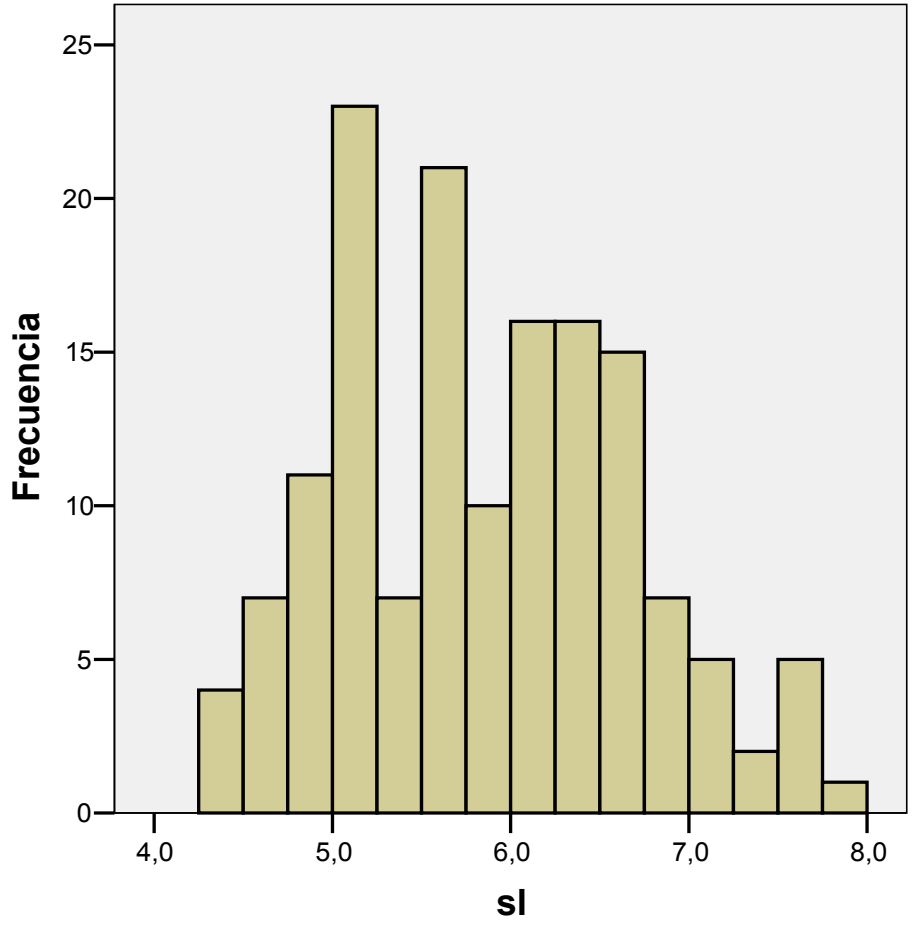
Problemas de interés relacionados con estos datos

- ¿Están las variables relacionadas? Al variar una, ¿cómo varían los valores de las otras?
- ¿Existen datos atípicos?
- ¿Es razonable suponer un modelo normal multivariante?
- ¿Es posible reducir la dimensión de los datos sin perder mucha información?
- ¿Se pueden establecer grupos homogéneos de equipos?

Descripción univariante: longitud del sépalo

SL

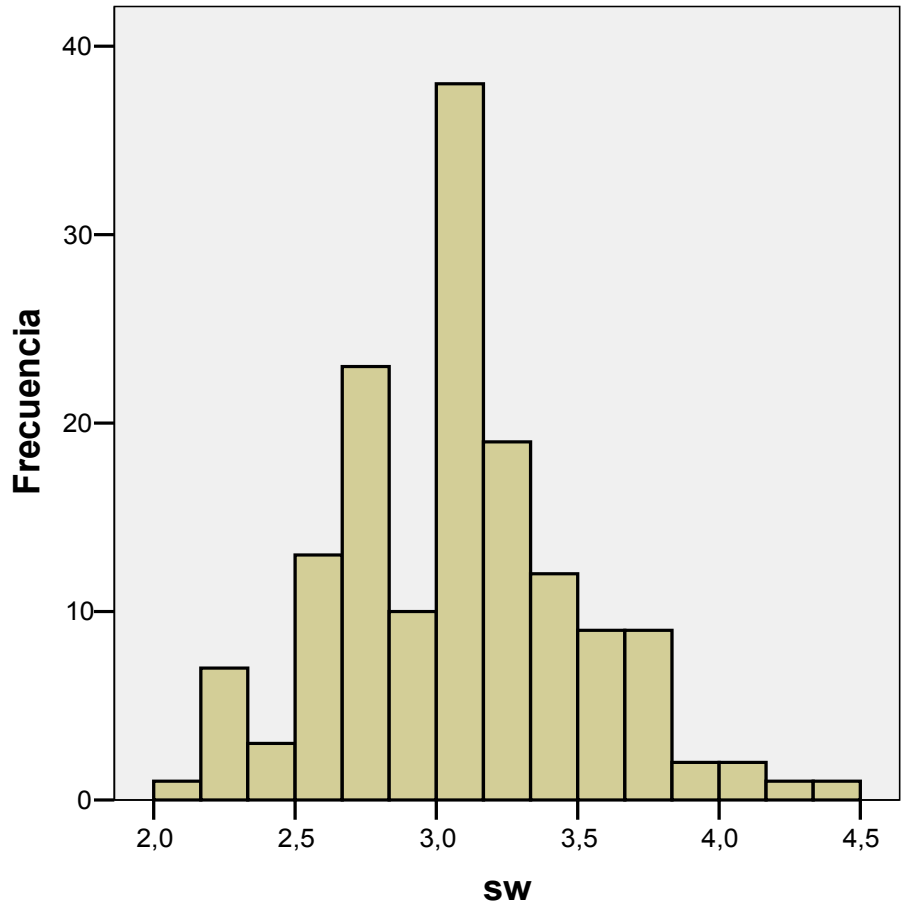
N	Válidos	150
	Perdidos	0
Media		5,843
Mediana		5,800
Desv. típ.		,8281
Varianza		,6857
Mínimo		4,3
Máximo		7,9
Percentiles	25	5,100
	50	5,800
	75	6,400



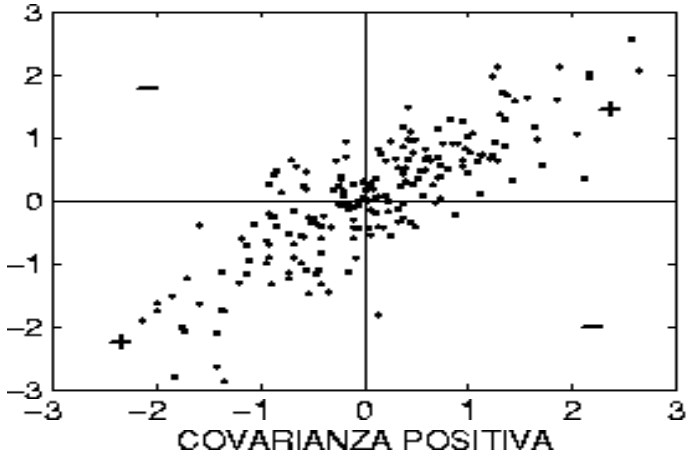
Descripción univariante: anchura del sépalo

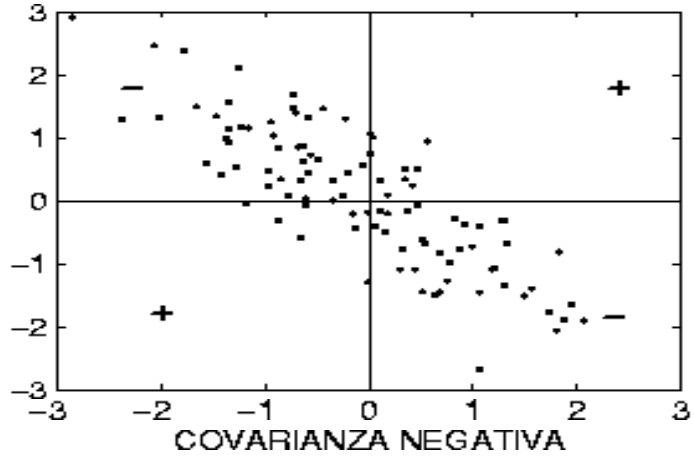
SW

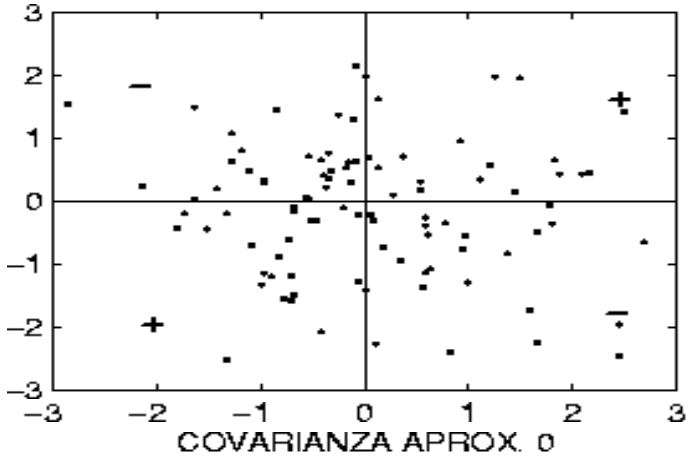
N	Válidos	150
	Perdidos	0
Media		3,054
Mediana		3,000
Desv. típ.		,4336
Varianza		,1880
Mínimo		2,0
Máximo		4,4
Percentiles	25	2,800
	50	3,000
	75	3,300

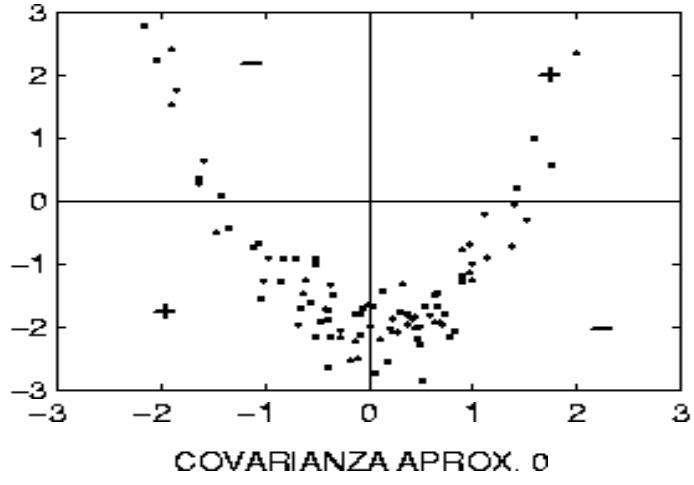


Interpretación de la covarianza









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

Dimensiones del sépalo: diagrama de dispersión

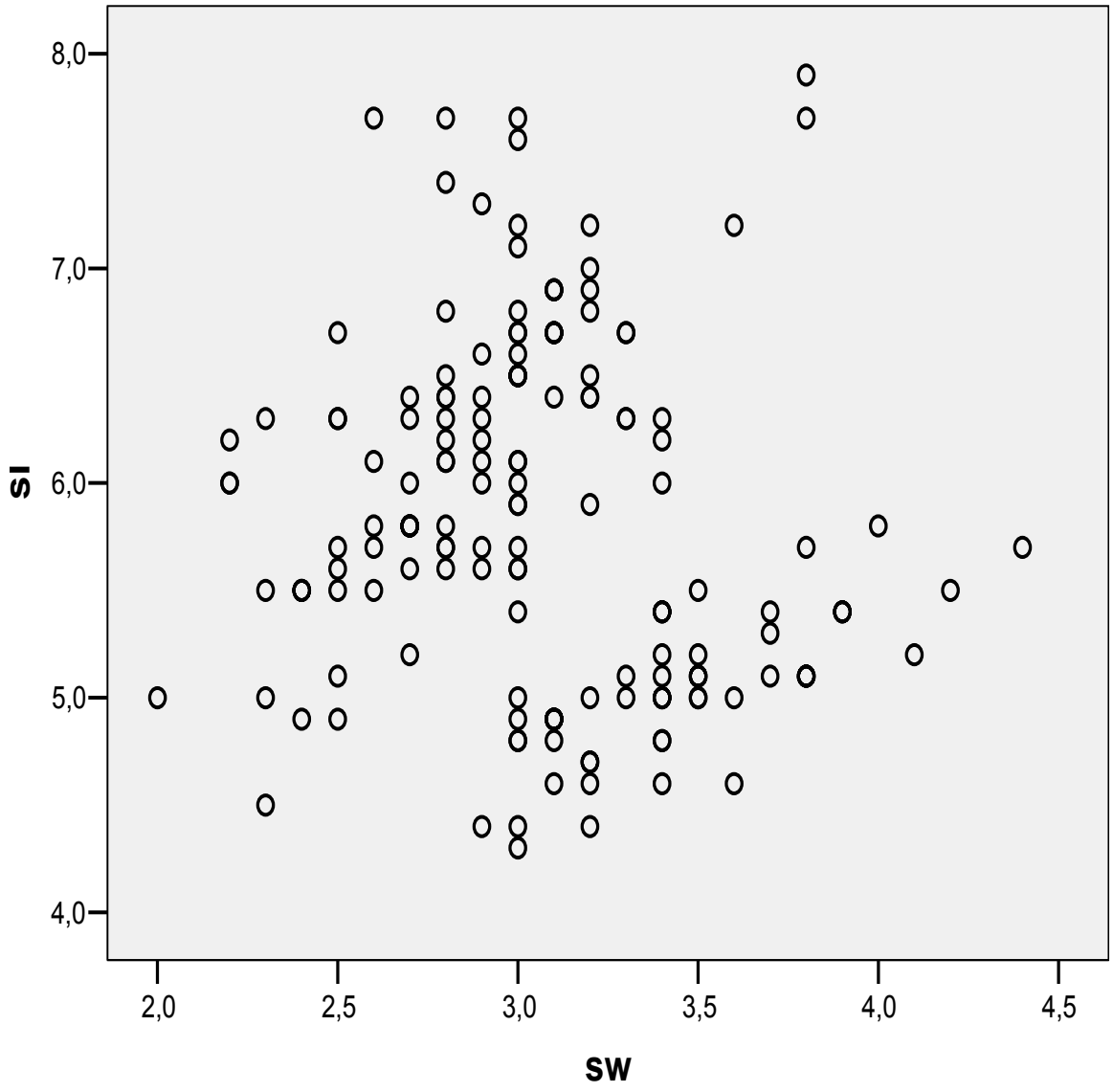
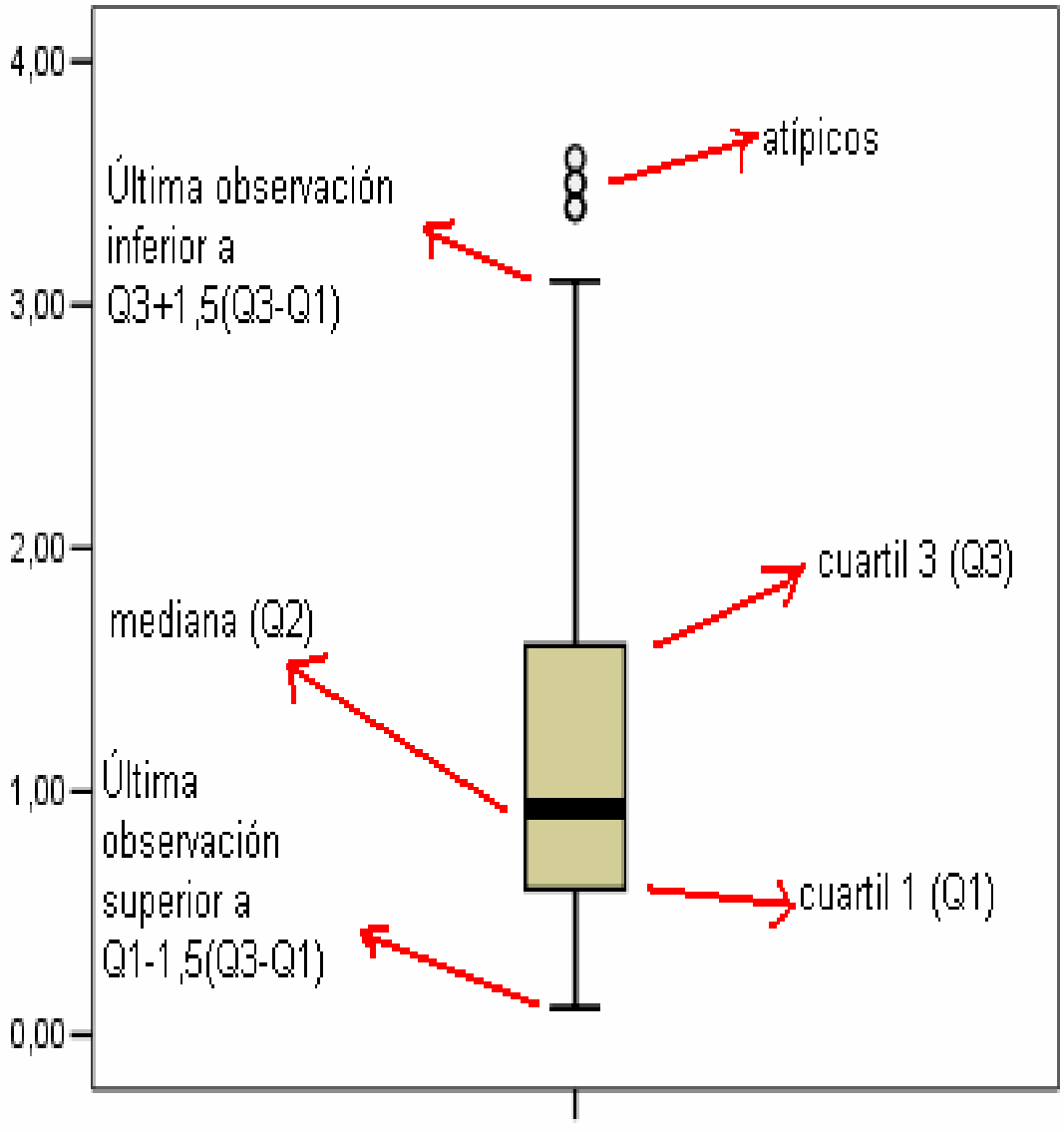
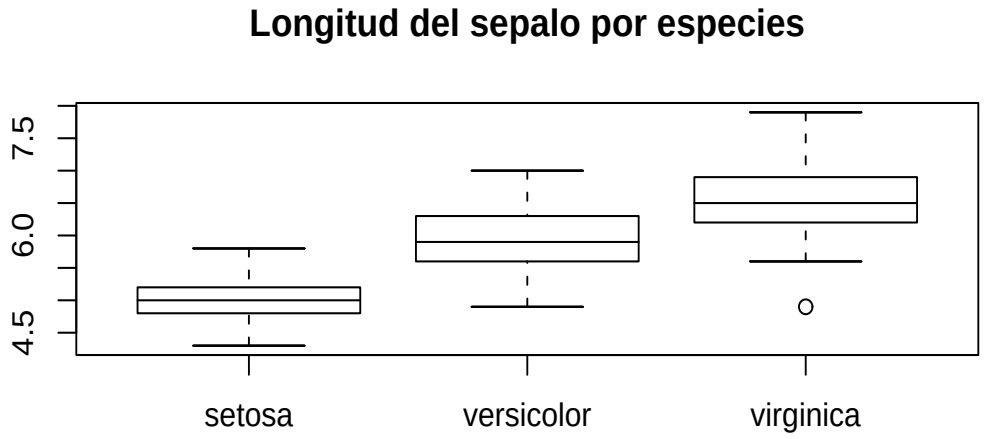
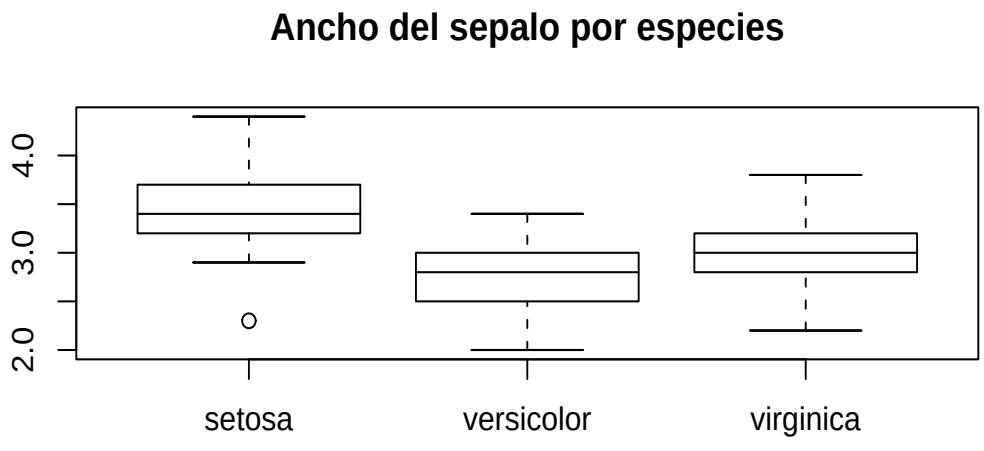


Diagrama de cajas

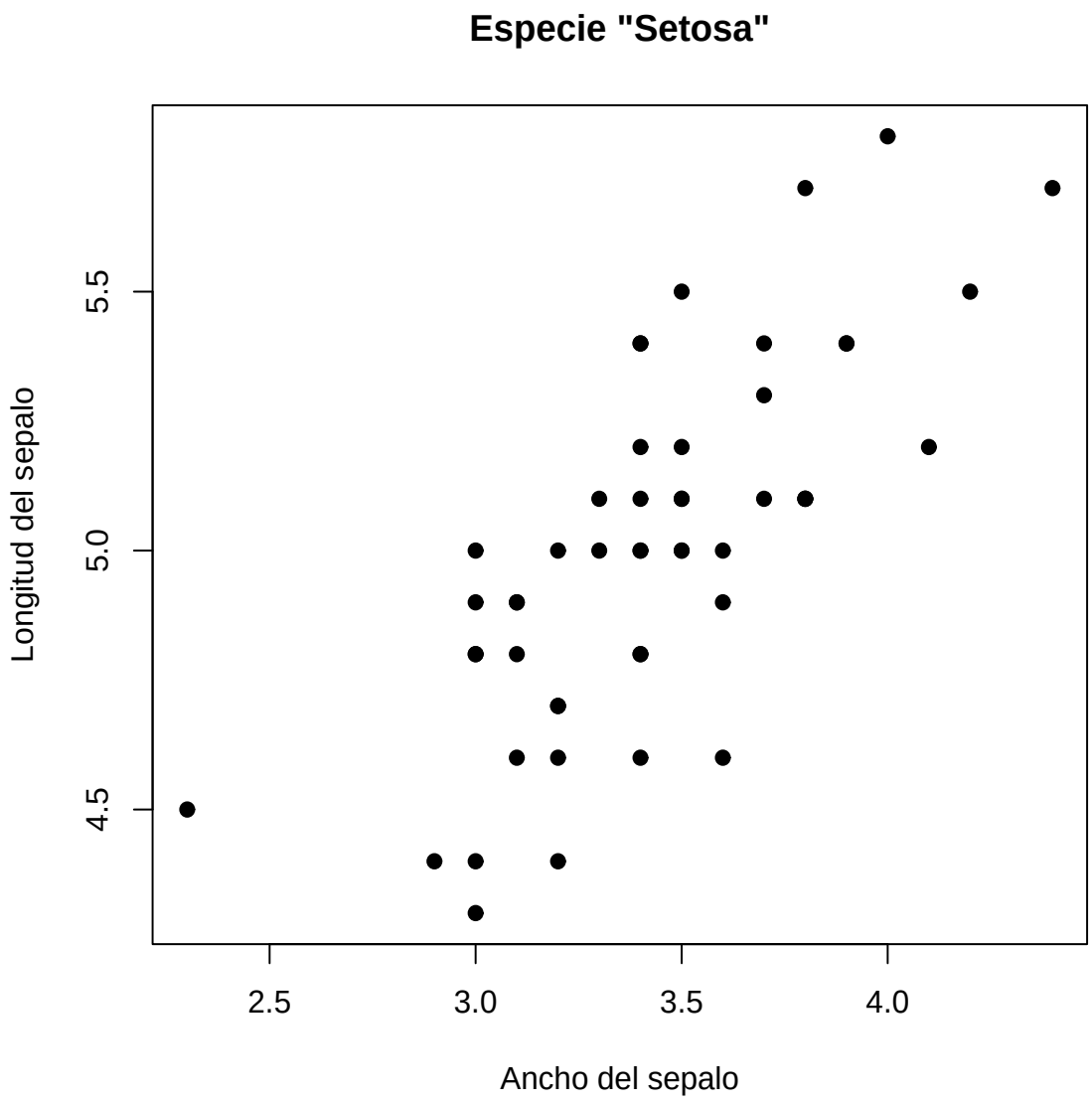


Dimensiones del sépalo: diagrama de cajas

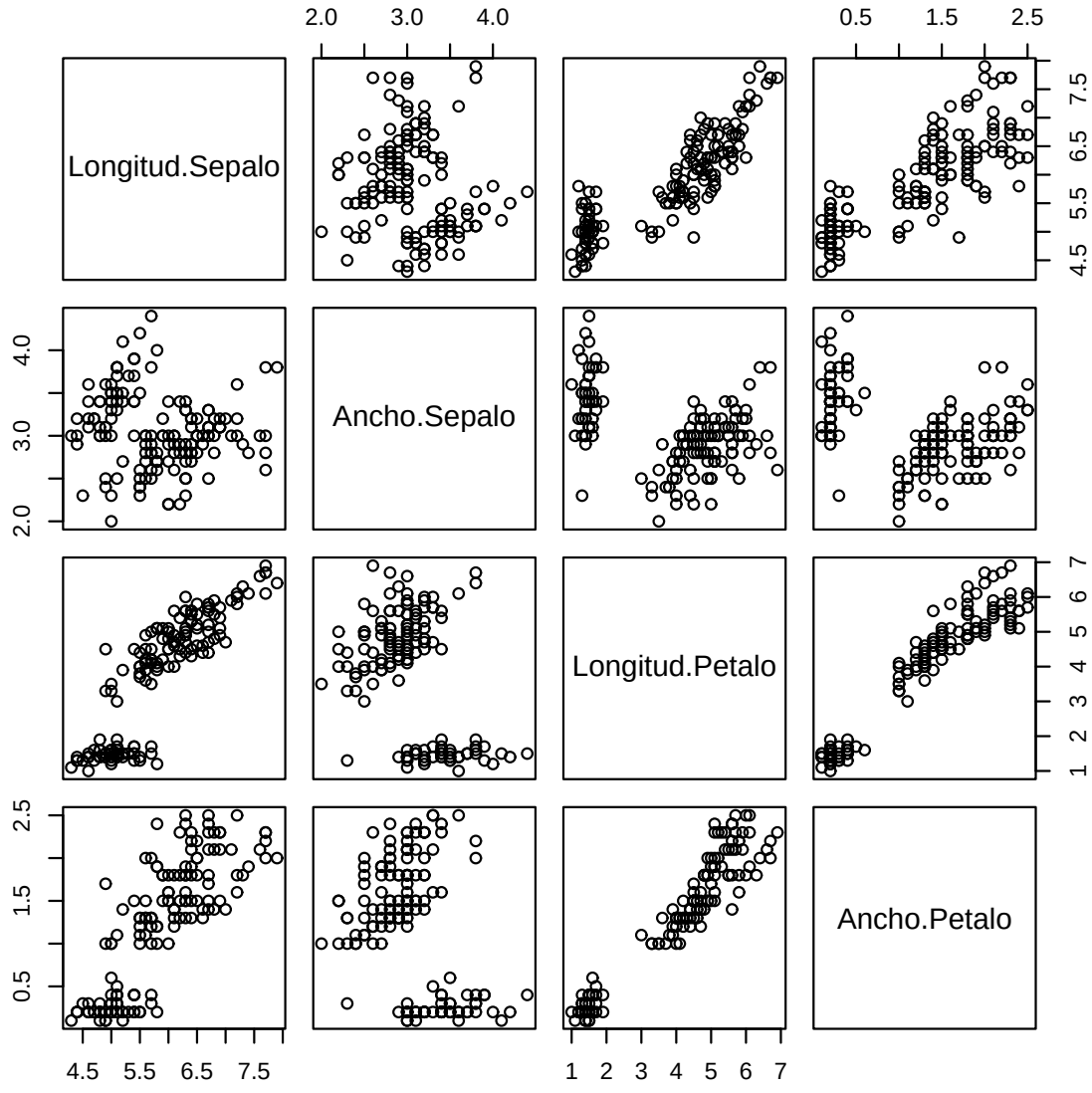




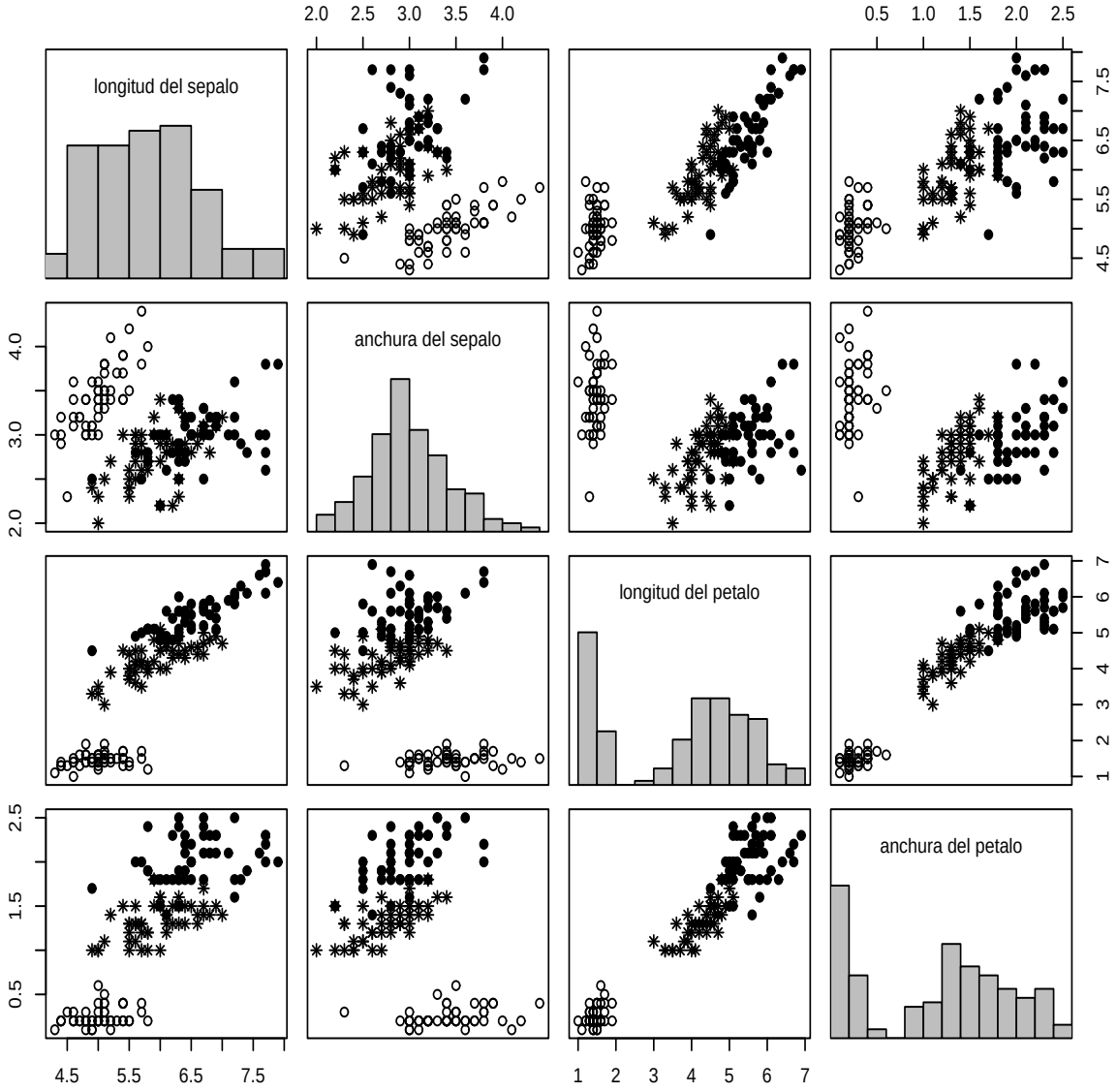
Dimensiones del sépalo de la especie *setosa*



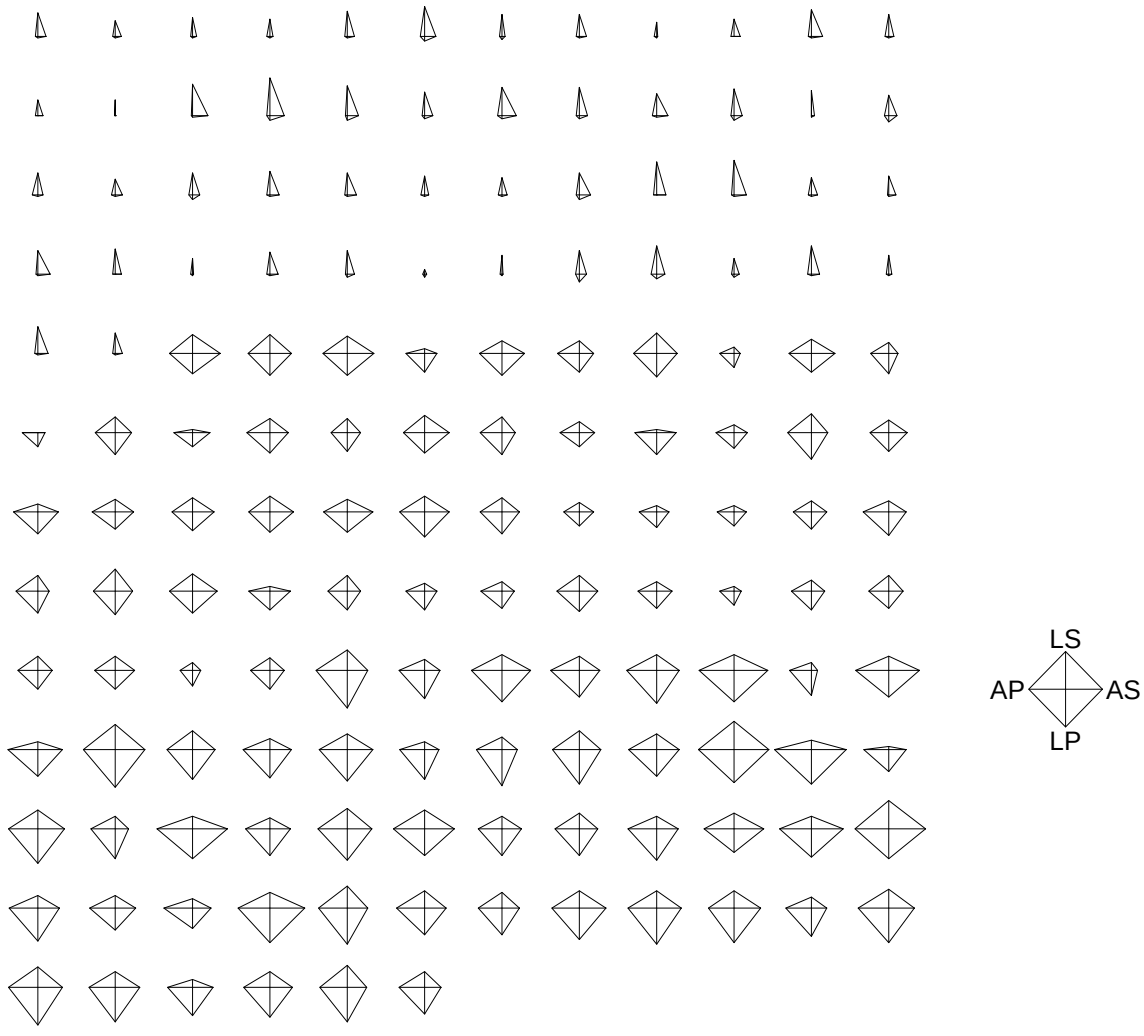
Lirios: matriz de diagramas de dispersión



Lirios: matriz de diagramas de dispersión



Lirios: gráfico de estrellas

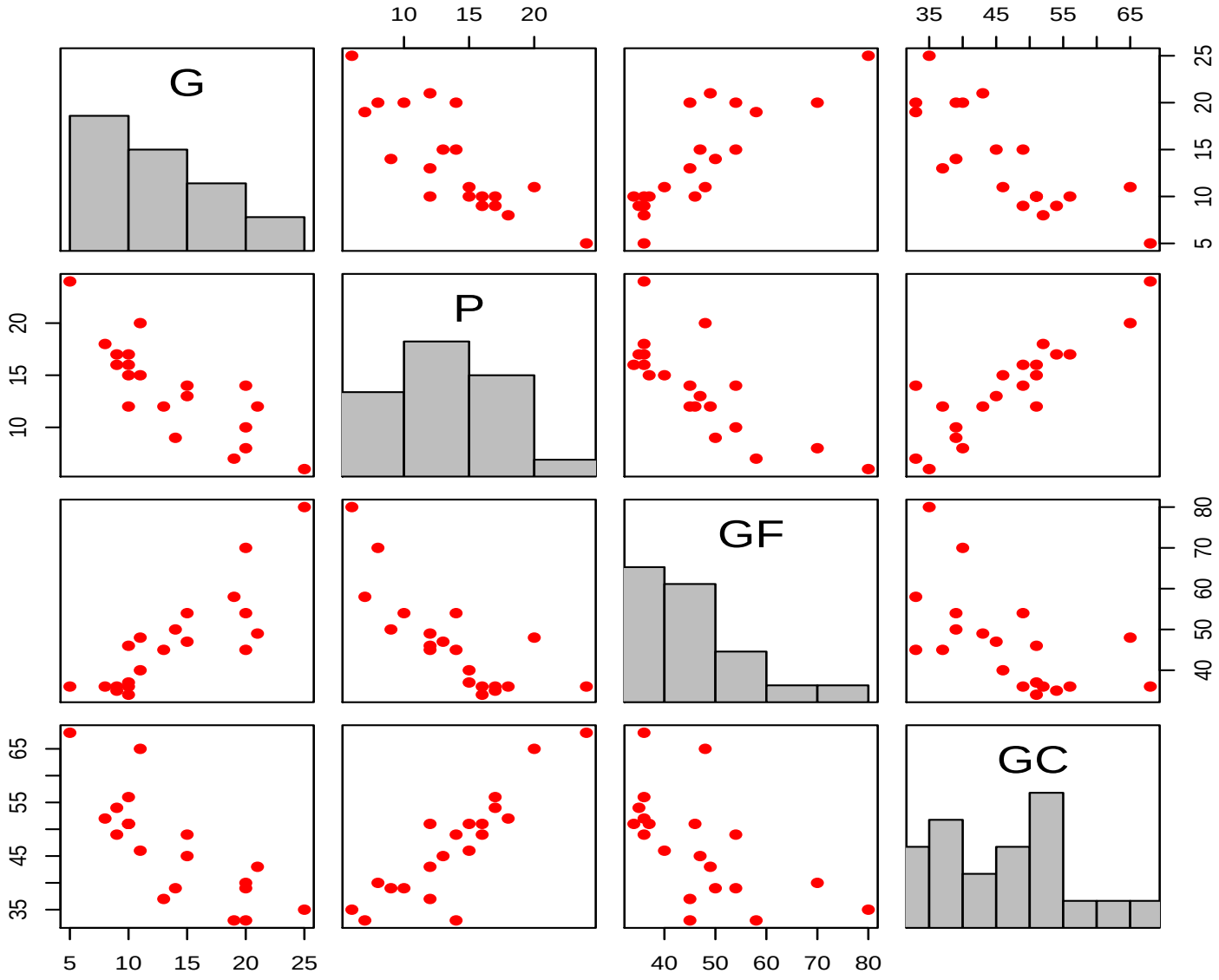


Lirios: matrices de covarianzas y de correlaciones

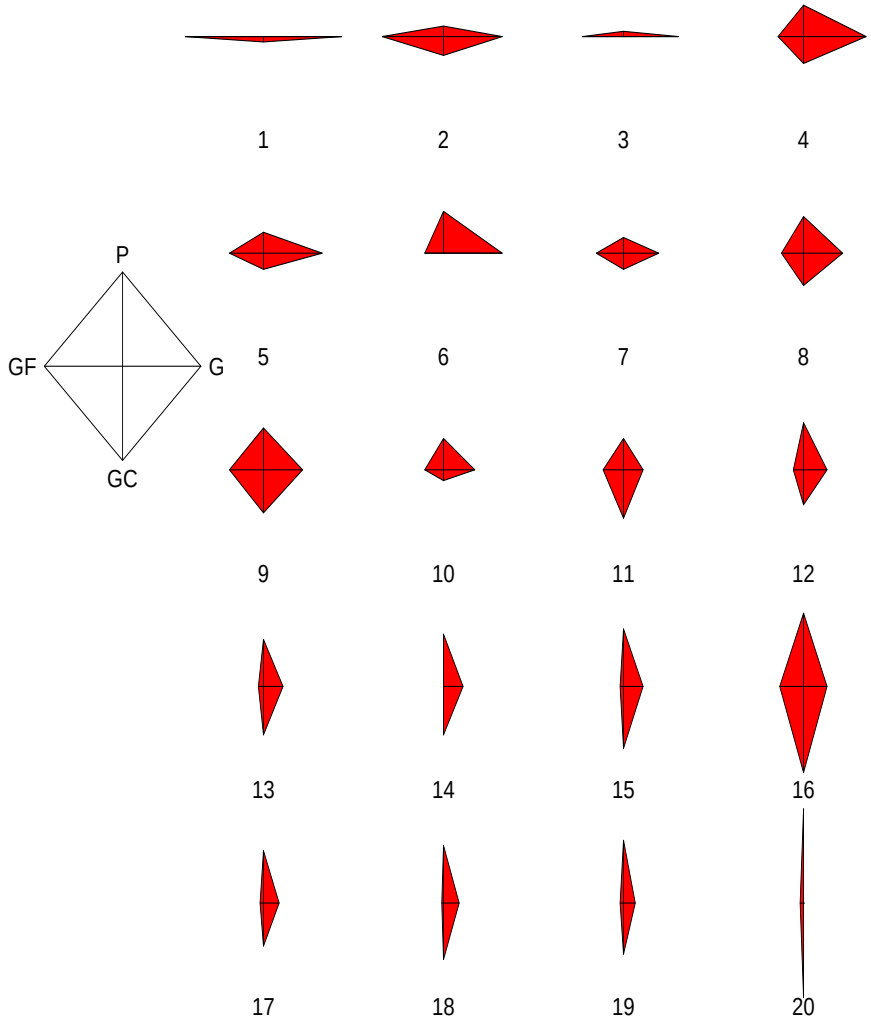
	Longitud.Sepalo	Ancho.Sepalo	Longitud.Petalo	Ancho.Petalo
Longitud.Sepalo	0.68569351	-0.04243400	1.2743154	0.5162707
Ancho.Sepalo	-0.04243400	0.18997942	-0.3296564	-0.1216394
Longitud.Petalo	1.27431544	-0.32965638	3.1162779	1.2956094
Ancho.Petalo	0.51627069	-0.12163937	1.2956094	0.5810063

	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

Fútbol: Matriz de diagramas de dispersión



Fútbol: Gráficos de estrellas



Fútbol: Medidas descriptivas numéricas

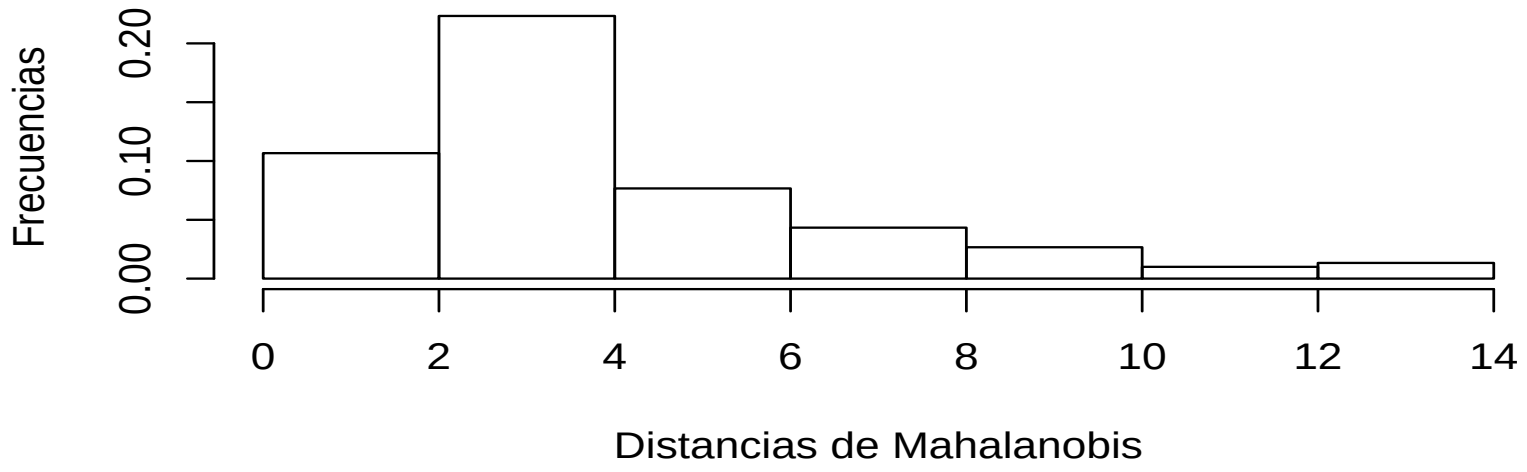
Estadísticos descriptivos

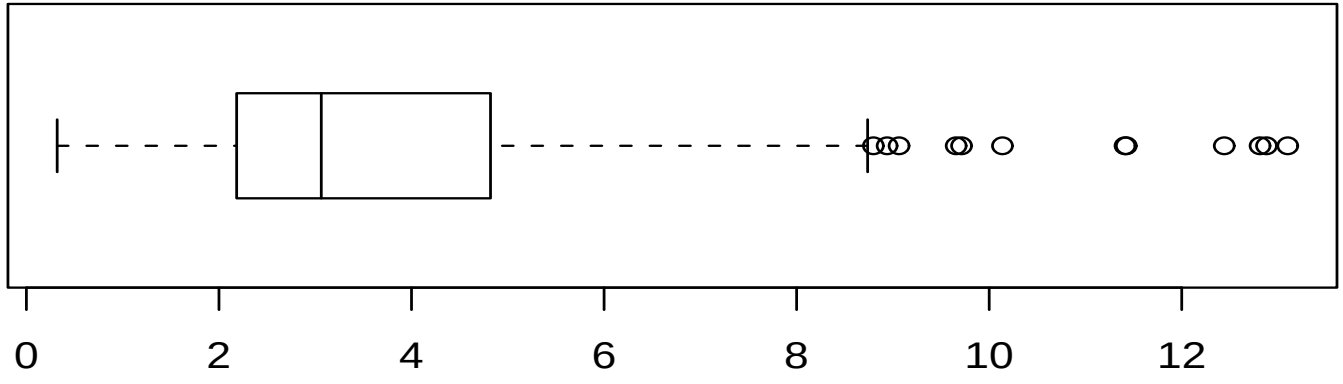
	Media	Desviación típica	N
G	13,750	5,3986	20
P	13,750	4,4824	20
GF	46,800	12,1508	20
GC	46,800	9,7581	20

Correlaciones

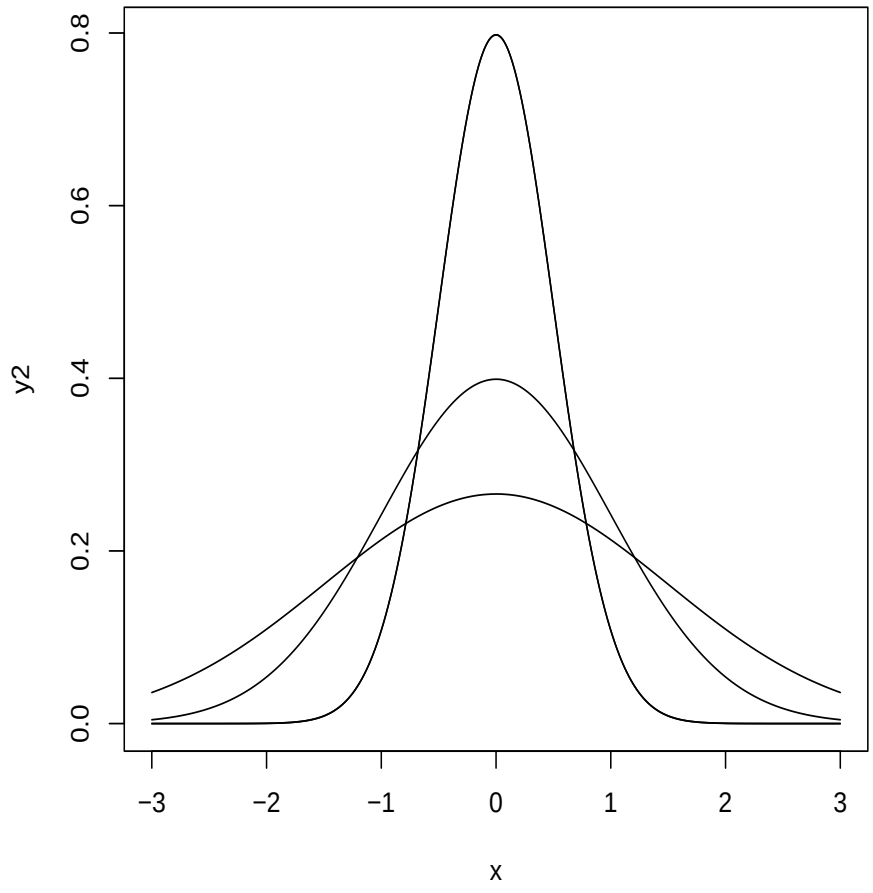
		G	P	GF	GC
G	Correlación de Pearson	1	-,812	,839	-,793
	Sig. (bilateral)	.	,000	,000	,000
	N	20	20	20	20
P	Correlación de Pearson	-,812	1	-,776	,878
	Sig. (bilateral)	,000	.	,000	,000
	N	20	20	20	20
GF	Correlación de Pearson	,839	-,776	1	-,577
	Sig. (bilateral)	,000	,000	.	,008
	N	20	20	20	20
GC	Correlación de Pearson	-,793	,878	-,577	1
	Sig. (bilateral)	,000	,000	,008	.
	N	20	20	20	20

Lirios: distancias de Mahalanobis



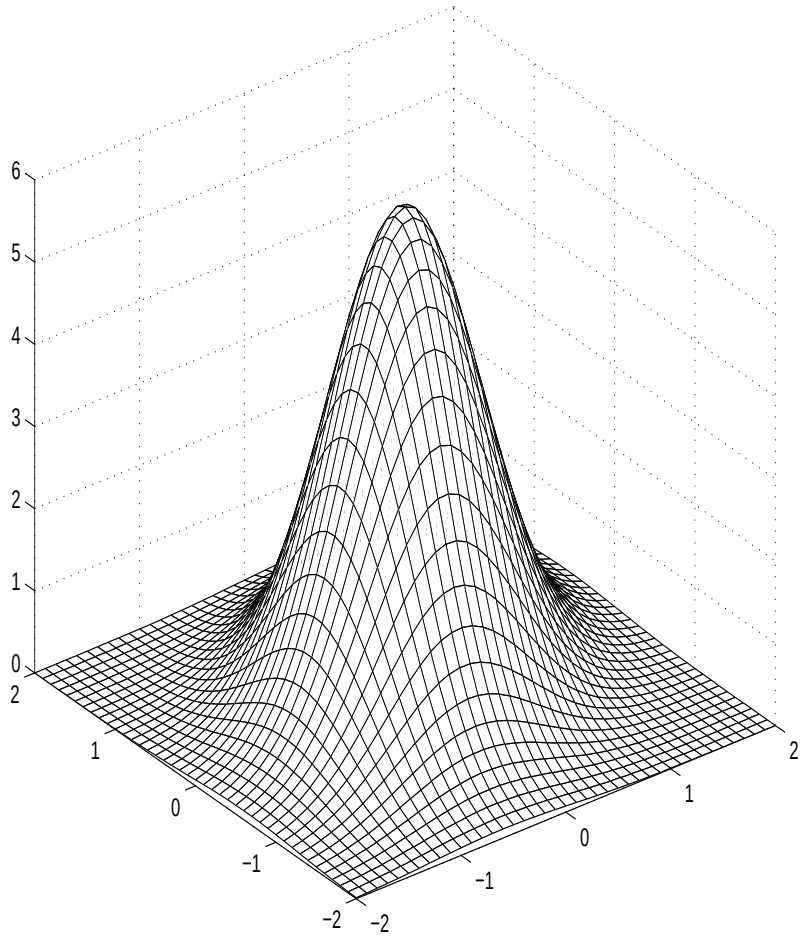


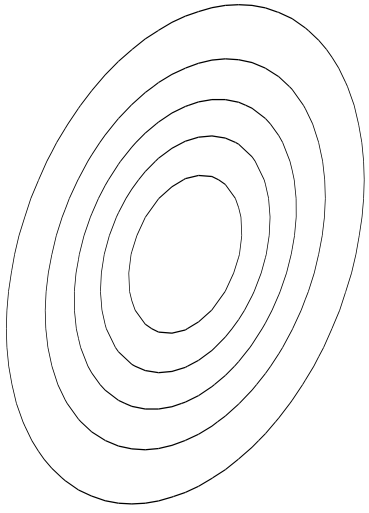
Normal univariante: densidad



¿Cuál de las tres corresponde a la normal estándar?

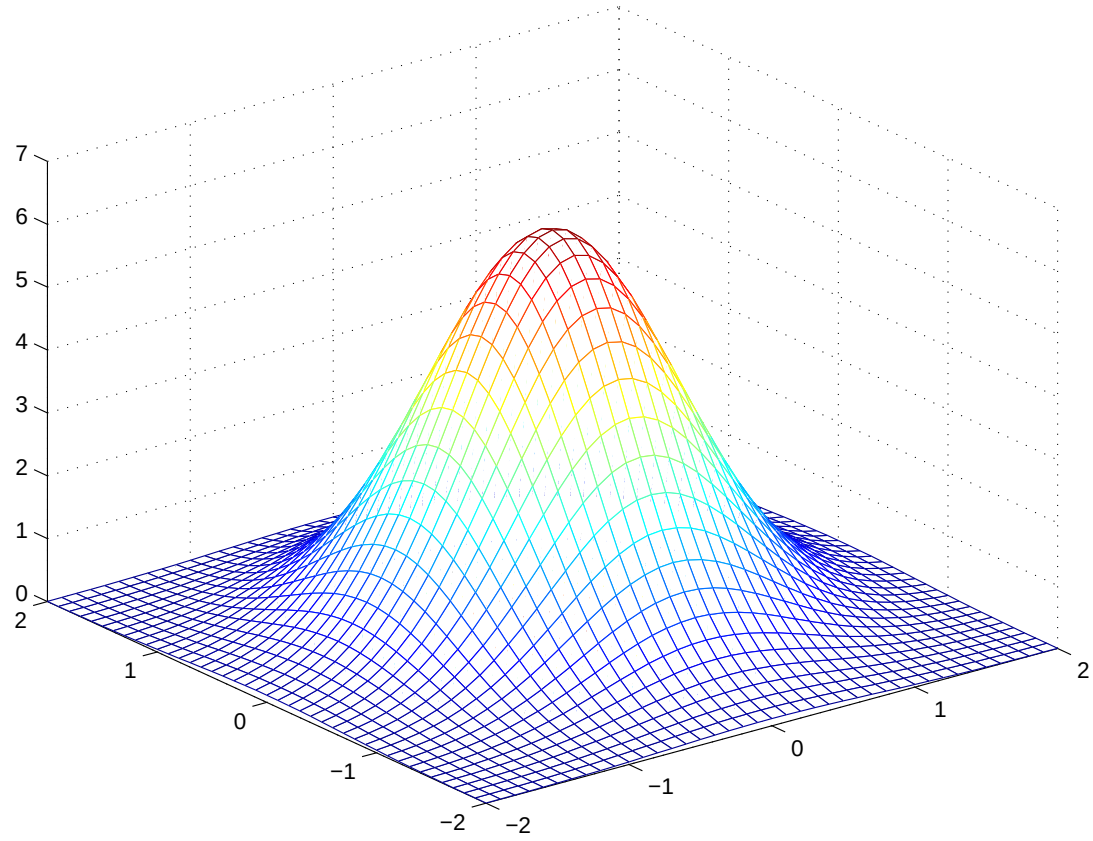
Normal multivariante: densidad



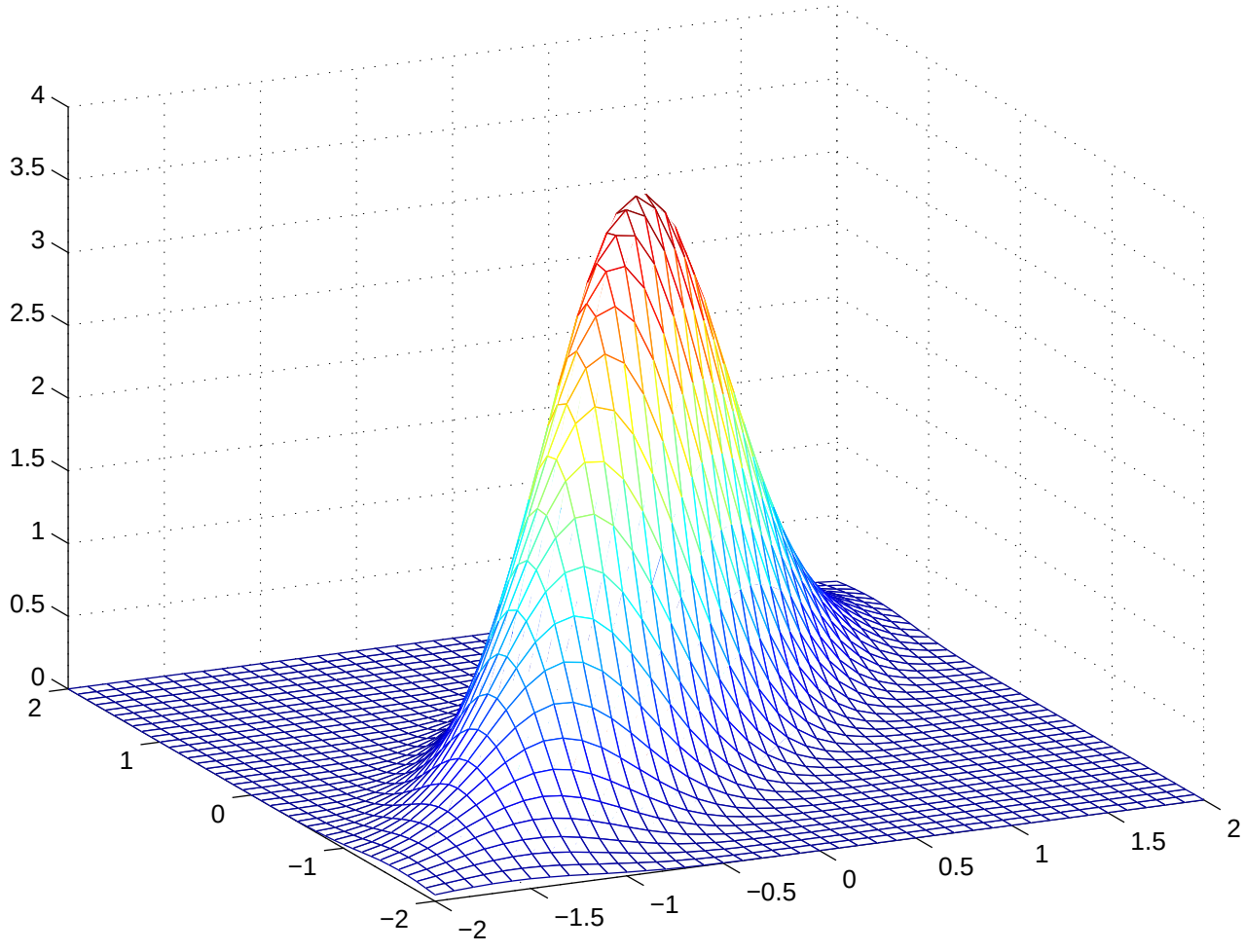


Densidad de la normal multivariante

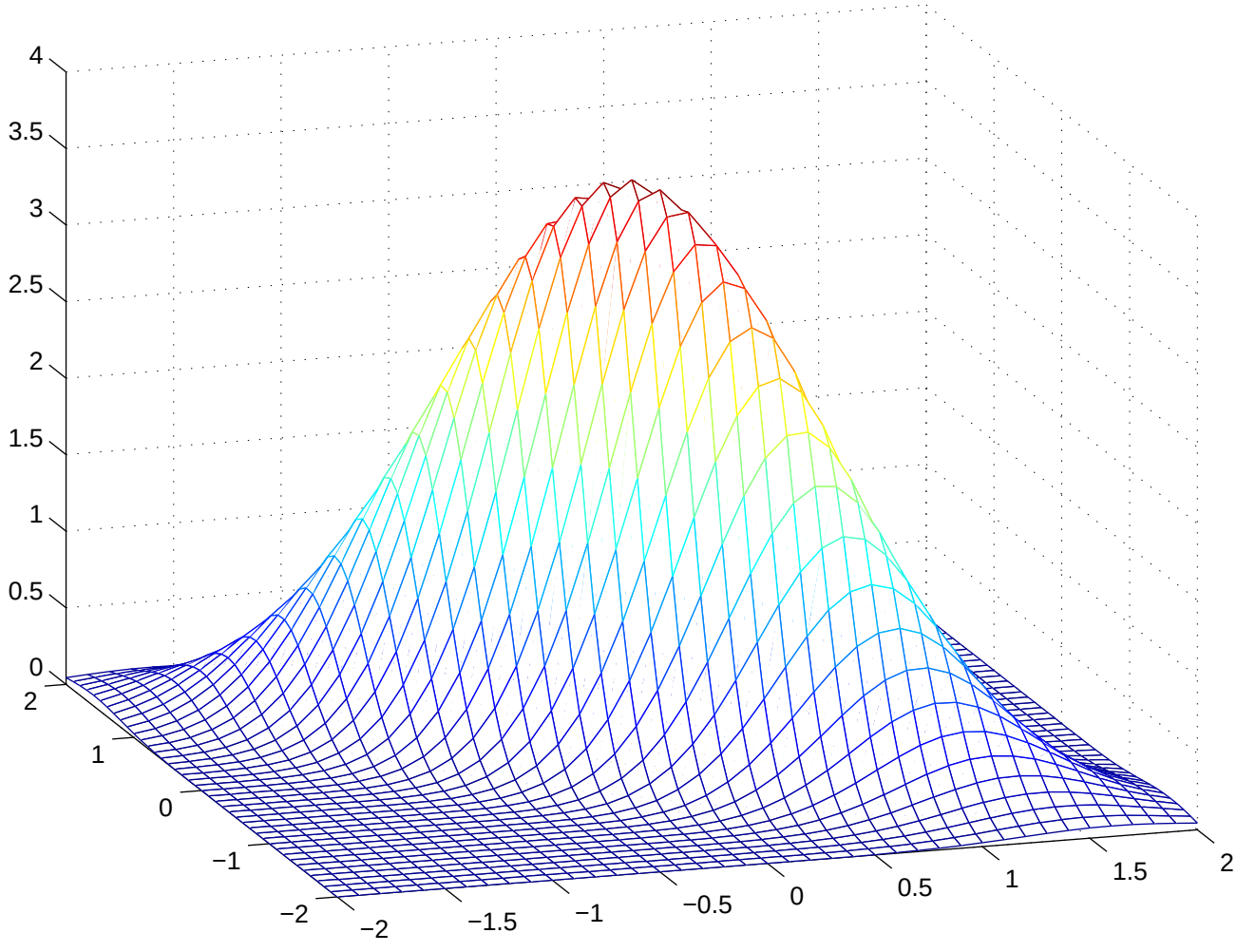
$$\mu = (0, 0)' \text{ y } \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



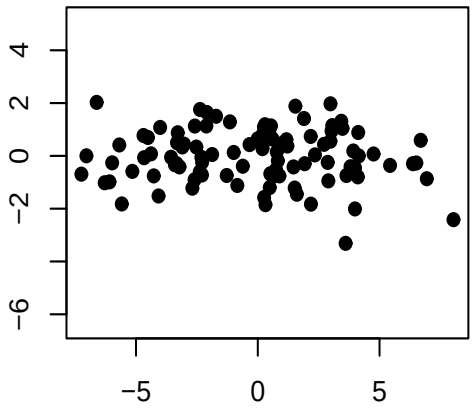
$$\mu = (0,0)' \text{ y } \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

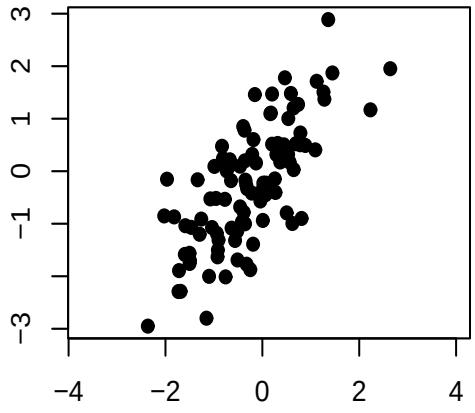


$$\mu = (0,0)' \text{ y } \Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$



Ejemplos de datos normales bidimensionales



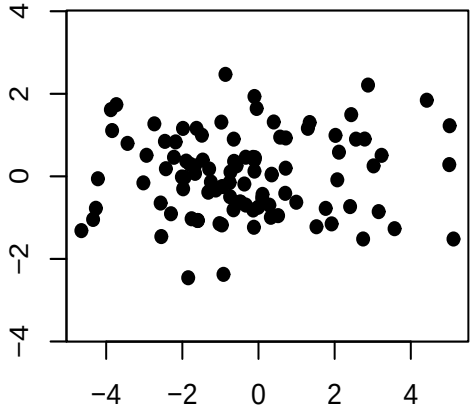


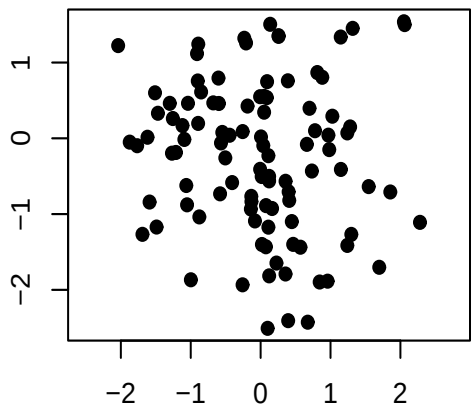
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

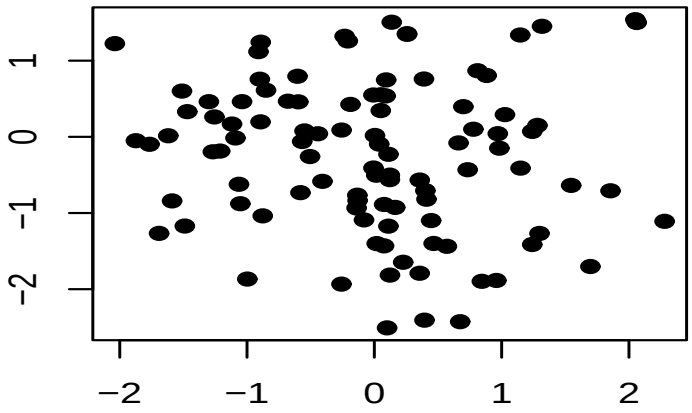
$$\Sigma = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$$

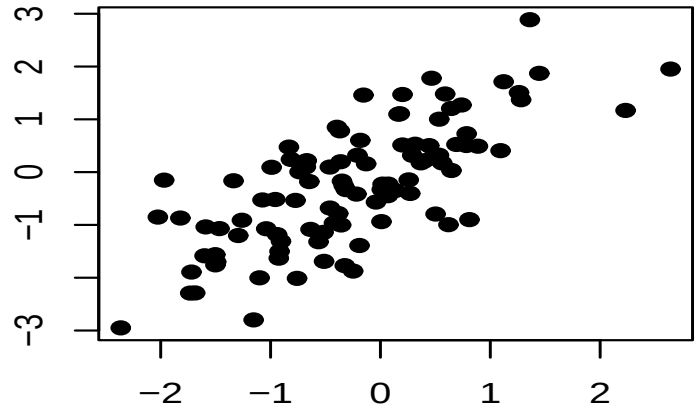
$$\Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

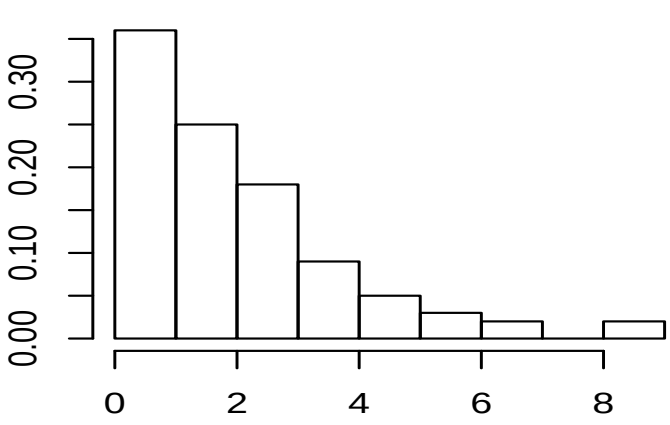


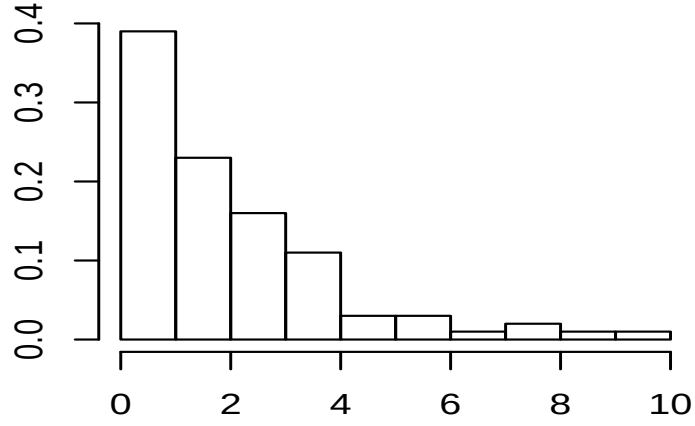


Distancias de Mahalanobis para datos normales







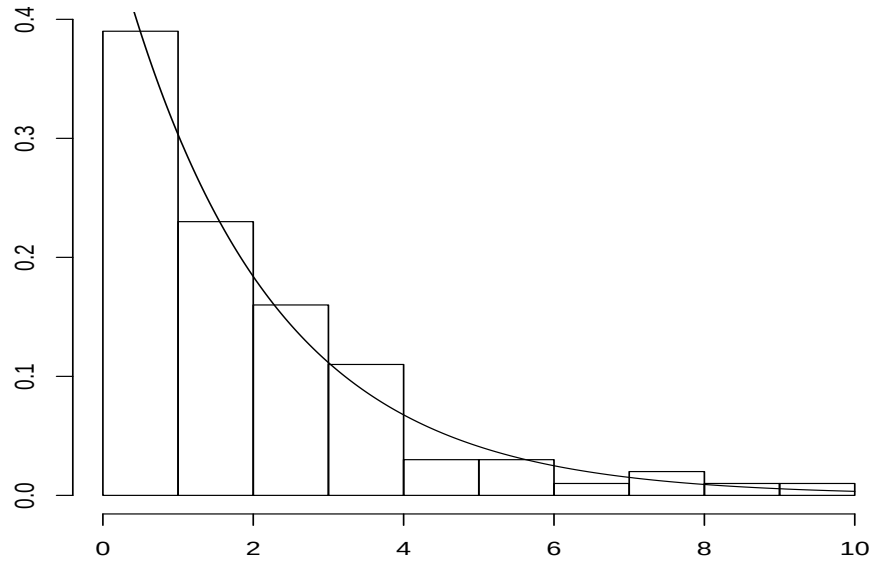


Distancias de Mahalanobis para datos normales

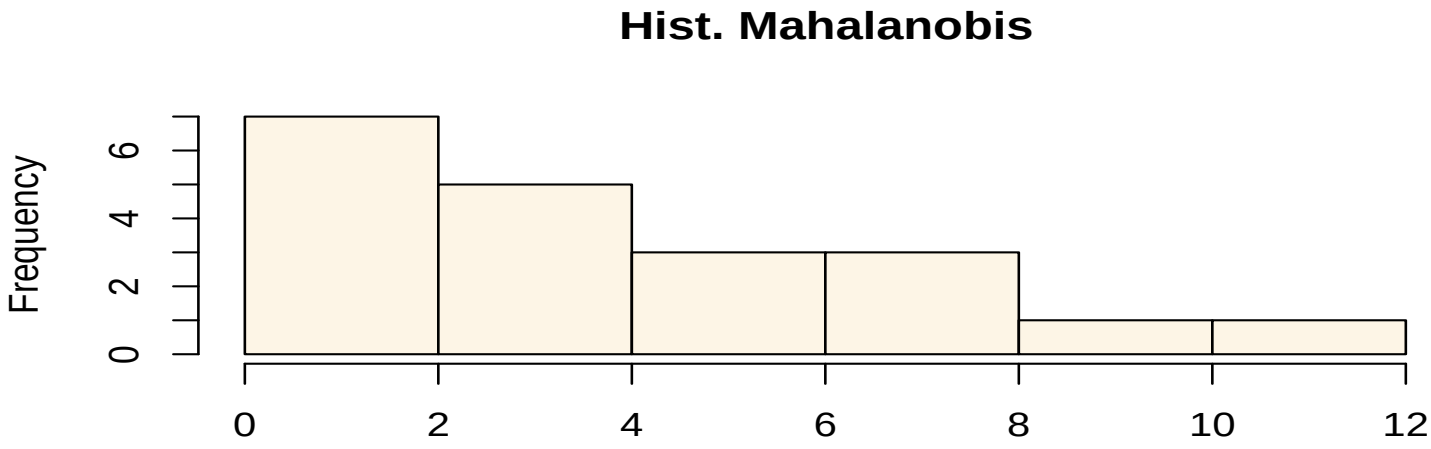
Estadísticos descriptivos para D_i^2 en el segundo ejemplo:

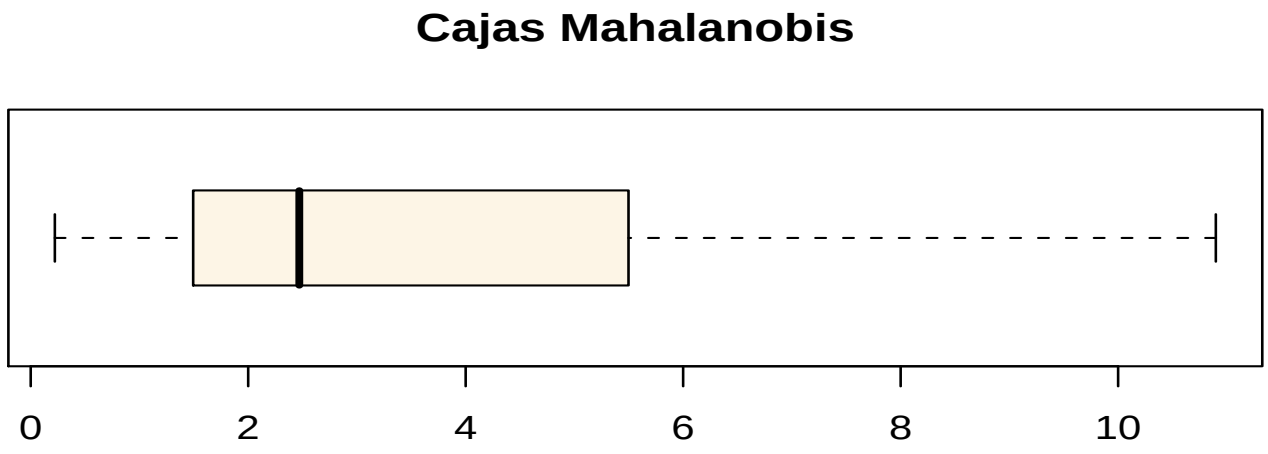
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.007255 0.565100 1.314000 1.980000 2.710000 9.735000

Desviacion tipica: 1.920563
```

Comparación con la densidad χ^2:



Fútbol: distancias de Mahalanobis





- La forma del histograma coincide con lo que se espera bajo normalidad (distribución χ^2)
- La distancia de Mahalanobis media es 3.8 y la varianza de las distancias es 9.03
- La mayor distancia es 10.90 y corresponde al `Celta`.
- La menor distancia es 0.22 y corresponde al `Deportivo`