Análisis exploratorio de datos multivariantes

1. Para el conjunto de datos que se presenta en la siguente tabla:

Individual	Var1	Var2	Var3	Var4
I1	1.96	1.31	3.49	3.99
I2	2.36	2.88	1.17	3.40
I3	2.95	1.44	3.21	2.16

(a) Escribir la matriz de datos $\mathbf{X}$.

(b) Calcular el vector de medias $\bar{\mathbf{x}}$.

(c) Calcular la matriz de covarianzas $\mathbf{S}$ y la matriz de correlaciones $\mathbf{R}$.

2. En la siguiente tabla presentamos un conjunto de datos (reducido) del famoso estudio sobre clasificación de tres especies de iris (*setosa*, *versicolor* y *virginica*) realizado por el estadístico y genetista Sir Ronald A. Fisher. Las variables son la longitud del sépalo (SL), el ancho del sépalo (SW), longitud del pétalo (PL), y el ancho del pétalo (PW).

Clase	SL	SW	PL	PW
setosa	5.1	3.5	1.4	0.2
setosa	4.9	3.0	1.4	0.2
setosa	4.7	3.2	1.3	0.2
versicolor	7.0	3.2	4.7	1.4
versicolor	6.4	3.2	4.5	1.5
versicolor	6.9	3.1	4.9	1.5
virginica	6.3	3.3	6.0	2.5
virginica	5.8	2.7	5.1	1.9
virginica	7.1	3.0	5.9	2.1

(a) Para cada una de las especies escribir la matriz de datos $\mathbf{X}$, calcular el vector de medias $\bar{\mathbf{x}}$ y calcular la matriz de covarianzas $\mathbf{S}$.

(b) Con el ordenador, repetir el apartado (a) usando los datos del fichero `iris.xls` (disponible en la web de la asignatura).

(c) Con el ordenador, obtener la matriz de diagramas de dispersión de las variables SL, SW, PL y PW, y describir los resultados obtenidos.

3. En las siguientes salidas de SPSS presentamos un análisis descriptivo numérico de los datos del fichero `limes.xls` que contiene las siguientes medidas de limas de Tahiti: día de recolección, diámetro del fruto, tamaño del fruto, peso del fruto, volumen del fruto, volumen del zumo, peso del zumo y peso de la cáscara.

	Media	Desviación típica	N
Diametro fruto	5,2868	,70505	99
Tamaño fruto	6,3887	,81695	97
Peso fruto	88,439	31,4848	96
Volumen fruto	103,57	34,200	97
Volumen zumo	31,05	14,254	97
Peso zumo	30,946	15,0697	99
Peso cascara	9,285	3,1917	98

		Diametro fruto	Tamaño fruto	Peso fruto	Volumen fruto	Volumen zumo	Peso zumo	Peso cascara
Diametro fruto	Correlación de Pearson	1	,890	,965	,955	,909	,915	,852
	Covarianza	,497	,504	21,210	23,241	8,810	9,695	1,916
	N	99	96	95	96	96	98	97
Tamaño fruto	Correlación de Pearson	,890	1	,931	,932	,841	,848	,838
	Covarianza	,504	,667	23,280	25,304	9,274	10,270	2,193
	N	96	97	94	94	94	96	96
Peso fruto	Correlación de Pearson	,965	,931	1	,990	,930	,936	,882
	Covarianza	21,210	23,280	991,295	1056,820	392,180	434,091	87,595
	N	95	94	96	93	94	95	94
Volumen fruto	Correlación de Pearson	,955	,932	,990	1	,924	,933	,905
	Covarianza	23,241	25,304	1056,820	1169,665	420,665	472,912	94,395
	N	96	94	93	97	94	96	95
Volumen zumo	Correlación de Pearson	,909	,841	,930	,924	1	1,000	,708
	Covarianza	8,810	9,274	392,180	420,665	203,177	203,259	30,733
	N	96	94	94	94	97	96	95
Peso zumo	Correlación de Pearson	,915	,848	,936	,933	1,000	1	,736
	Covarianza	9,695	10,270	434,091	472,912	203,259	227,097	35,241
	N	98	96	95	96	96	99	97
Peso cascara	Correlación de Pearson	,852	,838	,882	,905	,708	,736	1
	Covarianza	1,916	2,193	87,595	94,395	30,733	35,241	10,187
	N	97	96	94	95	95	97	98

(a) Calcular el vector de medias $\bar{\mathbf{x}}$, la matriz de covarianzas $\mathbf{S}$ y la de correlaciones $\mathbf{R}$.

(b) Con el ordenador, repetir el apartado (a) para cada mes de recolección usando los datos del fichero `limes.xls` (disponible en la web de la asignatura).

(c) En la Figura 1 se muestra la matriz de diagramas de dispersión para todos las limas recolectadas. ¿Existen relaciones lineales entre las variables?, ¿y no lineales?. ¿Crees que alguna variable se podría suprimir por ser redundante?. ¿Existen algunos frutos que tenga alguna dimensión atípica en relación al resto de los frutos?, ¿cómo son?

(d) En la Figura 2 se muestra un diagrama de caja multiple para los datos de las variables peso del fruto, volumen del fruto, volumen del zumo, y peso del zumo, por meses de recolección. ¿En qué mes crees que es mejor hacer la recolección si lo que se pretende es que las limas recolectadas tengan un aspecto similar?. ¿Se puede identificar en este gráfico alguno de los datos atípicos encontrados en el anterior?

Figura 1: Matriz de diagramas de dispersión.

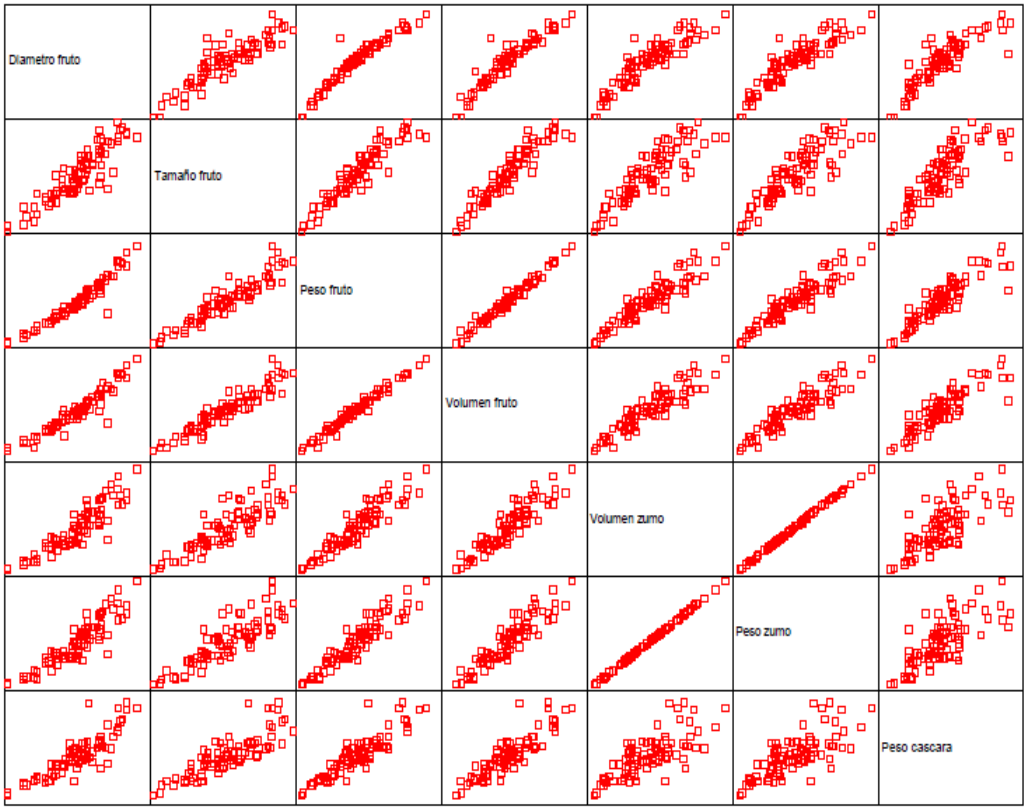
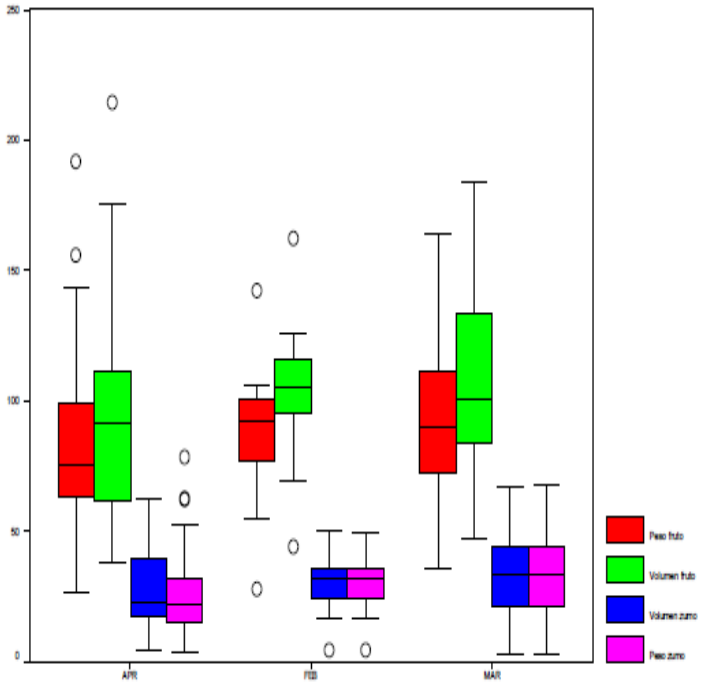


Figura 2: Diagramas de cajas por meses de recolección.



4. En las siguientes tablas presentamos las salidas de SPSS con los resultados de un estudio sobre una muestra de especímenes de anuros de la familia de los bufónidos del Parque Nacional Yasuní en Ecuador. Las variables que se consideran son:

- LRC: Longitud rostro-caudal.
- LM: Longitud de la mandíbula (de porción más anterior de la boca a articulación mandibular).
- AC: Ancho de la cabeza (a nivel de la articulación mandibular).
- ALC: Altura de la cabeza (desde el maxilar superior hasta el borde anterior del ojo).
- LF: Longitud del fémur.
- LTA: Longitud de la tibia.

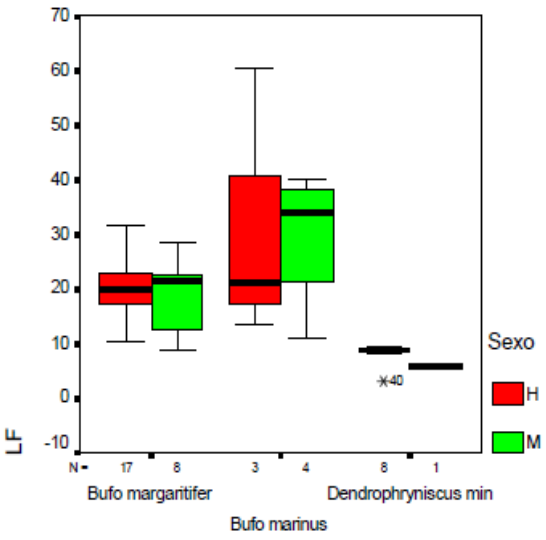
Estadísticos descriptivos

GP		N	Mínimo	Máximo	Media	Desv. típ.
Bufo margaritifer	LRC	25	23,50	78,00	48,5608	14,70828
	LM	25	6,59	23,40	14,7348	4,60823
	AC	25	8,82	29,33	17,9164	5,61413
	ALC	25	3,20	11,43	7,6568	1,97259
	LF	25	8,80	31,35	19,8424	5,98141
	LTA	25	8,77	31,29	20,0144	5,91083
	N válido (según lista)	25				
Bufo marinus	LRC	7	29,85	124,20	79,1386	36,49416
	LM	7	9,41	45,10	24,8729	12,44940
	AC	7	11,27	62,30	31,7329	18,01620
	ALC	7	5,14	22,64	12,1771	5,92320
	LF	7	11,04	60,32	30,5357	17,18499
	LTA	7	11,02	65,15	32,9257	18,61429
	N válido (según lista)	7				
Dendrophryniscus minutus	LRC	9	15,54	20,37	18,2156	1,38339
	LM	9	4,87	6,24	5,8344	,40952
	AC	9	5,08	6,95	5,7089	,51428
	ALC	9	1,91	6,95	2,9656	1,52106
	LF	9	3,14	9,33	7,8922	2,07508
	LTA	9	5,86	10,97	9,3144	1,48739
	N válido (según lista)	9				

Correlaciones

		LRC	LM	AC	ALC	LF	LTA
LRC	Correlación de Pearson	1	,984**	,972**	,941**	,968**	,971**
	N	41	41	41	41	41	41
LM	Correlación de Pearson	,984**	1	,984**	,948**	,975**	,983**
	N	41	41	41	41	41	41
AC	Correlación de Pearson	,972**	,984**	1	,955**	,973**	,980**
	N	41	41	41	41	41	41
ALC	Correlación de Pearson	,941**	,948**	,955**	1	,936**	,946**
	N	41	41	41	41	41	41
LF	Correlación de Pearson	,968**	,975**	,973**	,936**	1	,991**
	N	41	41	41	41	41	41
LTA	Correlación de Pearson	,971**	,983**	,980**	,946**	,991**	1
	N	41	41	41	41	41	41

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



(a) ¿De qué especie es el individuo con menor longitud del fémur?, ¿se puede saber su sexo?.

(b) ¿El vector de medias, en toda la muestra, de la variable vectorial definida por [ALC LF LTA]′ es aproximadamente igual a: [7,39 19,04 19,87]′?.

(c) ¿Qué dos variables están menos correlacionadas?, ¿podemos concluir que estas variables no están relacionadas linealmente?.

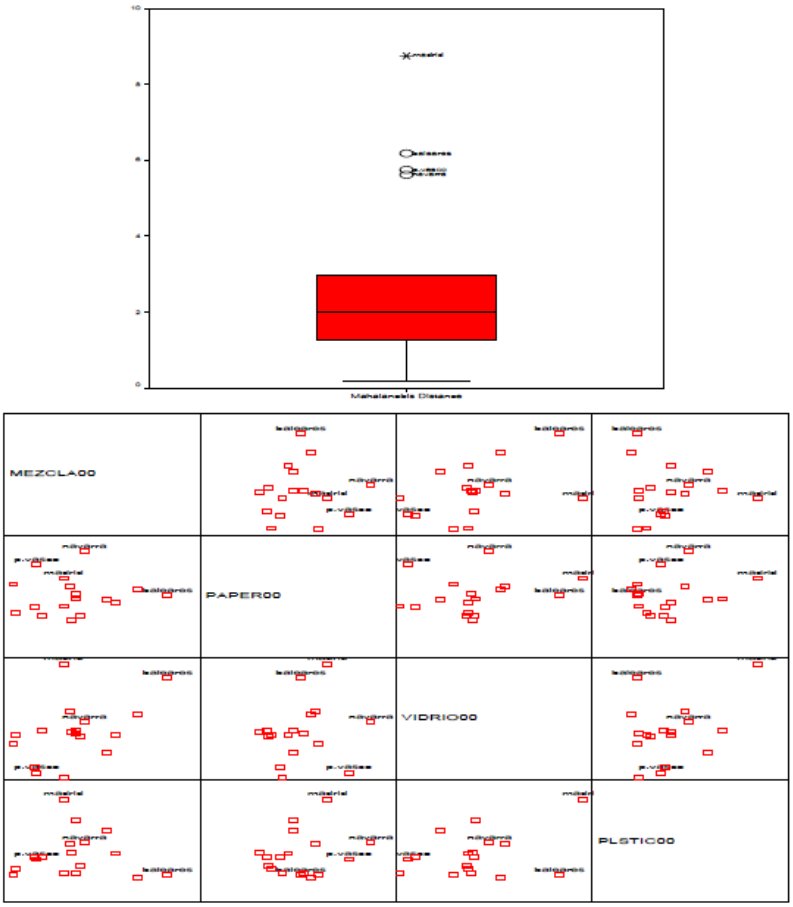
(d) Si nos piden que clasifiquemos a un nuevo individuo del que únicamente sabemos que es macho y que la longitud de su fémur es 10, ¿en qué especie lo clasificarías? ¿crees que es fácil que te equivoques?

(e) Y si del nuevo individuo lo único que sabemos es que la áltura de la cabeza es 9, ¿en qué especie lo clasificarías?

5. Se realiza un estudio sobre las tasas de reciclado de papel, vidrio y plástico y la tasa de residuos mezclados entre las comunidades autónomas españolas durante el año 2000. En cada Comunidad se ha elegido un grupo de 1000 familias que viven en zonas urbanas y en cada uno de ellos se ha medido la cantidad total (en Tm) de residuos (variable MEZCLA) que han producido y las cantidades totales de papel, vidrio y plástico que han reciclado. Las salidas de SPSS correspondientes a un análisis descriptivo de estos datos son las siguientes:

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
MEZCLA00	17	424	817	578,06	104,888
PAPER00	17	9,1	26,1	15,282	4,5709
VIDRIO00	17	10,2	21,4	14,741	2,9067
PLSTIC00	17	,0	32,2	9,612	8,8101
N válido (según lista)	17				



(a) ¿Existen cuatro comunidades con un comportamiento atípico respecto al tratamiento de los diferentes residuos?, ¿en todas ellas los niveles de reciclado (variables papel, vidrio y plástico) son superiores a los respectivos niveles promedio de las 17 comunidades?.

(b) ¿Las Comunidades Autónomas de Baleares, Navarra y País Vasco forman un grupo homogéneo o heterogéneo en términos de su respuesta a las variables analizadas en este estudio?.

(c) Calcular la media de material reciclado (papel+vidrio+plásticos) para las 17 comunidades autónomas.

(d) ¿Qué cantidad de plástico y vidrio se ha reciclado en la Comunidad de Madrid?.

(e) ¿Existe una correlación muy alta entre la cantidad de residuos producidos y la cantidad de papel, vidrio y plástico que se recicla?.

6. Para la base de datos del ejercicio 3, grafique e interprete un diagrama de estrellas y un diagrama de caras de Chernoff.