

# Índice

<b>1. Preliminares y Comparación de Poblaciones</b>	<b>17</b>
1.1. Introducción . . . . .	17
1.2. Nombres nuevos para conocidos métodos clásicos . . . . .	17
1.3. Algunos elementos matemáticos básicos . . . . .	19
1.4. Algunos elementos básicos de los vectores aleatorios . . . . .	21
1.5. La distribución normal multivariante . . . . .	23
1.5.1. Análisis de la normalidad multivariante . . . . .	24
1.6. Comparación de dos poblaciones multivariantes . . . . .	26
1.6.1. Test $T^2$ de Hotelling . . . . .	26
1.6.2. Test de permutaciones . . . . .	28
1.7. Análisis de la Varianza Multivariante (MANOVA) . . . . .	29
1.8. Análisis de la Varianza Multivariante de Permutaciones (PER-MANOVA) . . . . .	32
<b>2. Análisis de Componentes Principales</b>	<b>35</b>
2.1. Introducción . . . . .	35
2.2. Determinación de las Componentes Principales . . . . .	37
2.3. Contribución de cada Componente Principal a la variabilidad total . . . . .	40
2.4. Componentes Principales Muestrales . . . . .	41
2.5. Estandarización . . . . .	42
2.6. Cálculo con R . . . . .	43
2.7. Elección del número de Componentes Principales . . . . .	47
2.8. Reducción en el número de variables . . . . .	50
2.9. Componentes Principales para datos bidimensionales . . . . .	51
2.9.1. Representaciones gráficas . . . . .	55
2.10. Scores . . . . .	59
2.11. Componentes Principales como transformaciones lineales ortogonales . . . . .	61
2.12. Detección de observaciones anómalas en datos multivariantes . . . . .	62
2.13. El biplot . . . . .	65

2.13.1. El triplot . . . . .	67
2.14. Determinación de clusters . . . . .	68
2.15. En búsqueda de la Proyección Óptima ( <i>Projection Pursuit</i> ) . . . . .	69
2.15.1. Clasificación con Proyección Óptima . . . . .	70
2.16. Referencias . . . . .	72
<b>3. Análisis de Correspondencias</b>	<b>75</b>
3.1. Introducción . . . . .	75
3.2. Análisis de Correspondencias bidimensional . . . . .	79
3.2.1. Cálculo con R . . . . .	86
3.2.2. Gráfico de correspondencias en tres dimensiones . . . . .	89
3.2.3. Dimensión de las coordenadas . . . . .	89
3.3. Análisis de Correspondencias múltiple . . . . .	93
3.3.1. Cálculo con R . . . . .	94
3.4. Referencias . . . . .	97
<b>4. Escalado Multidimensional</b>	<b>99</b>
4.1. Introducción . . . . .	99
4.2. Escalado Multidimensional Clásico: Métrico Euclídeo y no Euclídeo	101
4.2.1. Reconstrucción de la matriz de datos a partir de la matriz de distancias . . . . .	102
4.2.2. Matriz de proximidades Euclídea y no Euclídea . . . . .	105
4.2.3. Cálculo con R . . . . .	106
4.3. Escalado Multidimensional no Métrico . . . . .	109
4.4. Referencias . . . . .	111
<b>5. Análisis de Conglomerados</b>	<b>113</b>
5.1. Introducción . . . . .	113
5.2. Análisis cluster de casos . . . . .	115
5.2.1. Técnicas jerárquicas aglomerativas de formación de conglomerados . . . . .	116
5.2.2. Distancias y similitudes entre individuos . . . . .	119
5.2.3. Tipos de agrupamiento . . . . .	128
5.3. Análisis de Conglomerados con R . . . . .	136
5.3.1. Análisis cluster jerárquico aglomerativo . . . . .	136
5.4. Análisis cluster de variables . . . . .	148
5.5. Análisis cluster de bloques . . . . .	149
5.6. Métodos de optimización en el análisis cluster: Algoritmo $k$ -medias . . . . .	149
5.6.1. Minimización de la traza de $W$ . . . . .	151
5.6.2. Minimización del determinante de $W$ . . . . .	152
5.6.3. Maximización de la traza de $BW^{-1}$ . . . . .	152

---

5.6.4.	Algoritmo $k$ -medias-medias con R . . . . .	153
5.6.5.	Determinación del número de clusters . . . . .	157
5.7.	Técnicas inferenciales de formación de conglomerados . . . . .	160
5.7.1.	Elección del número de clusters . . . . .	161
<b>6.</b>	<b>Análisis Discriminante</b> . . . . .	<b>167</b>
6.1.	Introducción . . . . .	167
6.2.	Función discriminante lineal de Fisher . . . . .	169
6.2.1.	Utilización de probabilidades de priori . . . . .	172
6.2.2.	Cálculo con R . . . . .	173
6.3.	Valoración de la función discriminante . . . . .	176
6.4.	Función discriminante cuadrática . . . . .	179
6.4.1.	Cálculo con R . . . . .	180
6.5.	Referencias . . . . .	181
<b>7.</b>	<b>Análisis Factorial</b> . . . . .	<b>183</b>
7.1.	Introducción . . . . .	183
7.2.	Modelo del Análisis Factorial . . . . .	184
7.3.	Estimación de parámetros en el Modelo del Análisis Factorial . . . . .	186
7.3.1.	Análisis de Factores Principales . . . . .	188
7.3.2.	Análisis Factorial de Máxima Verosimilitud . . . . .	189
7.4.	Referencias . . . . .	189
<b>8.</b>	<b>Modelos Log-Lineales</b> . . . . .	<b>191</b>
8.1.	Introducción . . . . .	191
8.2.	Independencia condicionada . . . . .	195
8.3.	Tipos de Independencia . . . . .	202
8.4.	El modelo log-lineal como modelo lineal general . . . . .	210
8.4.1.	Comparación de modelos: Tests condicionales para modelos anidados . . . . .	213
8.5.	Modelos log-lineales con R . . . . .	215
<b>9.</b>	<b>Regresión Logística</b> . . . . .	<b>223</b>
9.1.	Introducción . . . . .	223
9.2.	Estimación y contraste . . . . .	227
9.3.	Regresión Logística con R . . . . .	227
9.4.	El modelo de regresión logística y el modelo log-lineal . . . . .	232
9.5.	Modelos de regresión Logit y Probit . . . . .	232
9.6.	Los modelos de regresión Logit y Probit como modelos lineales generalizados . . . . .	235
9.7.	Referencias . . . . .	236

<b>10.Regresión Poisson</b>	<b>237</b>
10.1. Introducción . . . . .	237
10.2. Estimación y contraste . . . . .	239
10.3. Regresión Poisson con R . . . . .	239
10.4. Bondad del ajuste . . . . .	242
10.5. Referencias . . . . .	245
<b>11.Regresión no Lineal y Regresión Suavizada</b>	<b>247</b>
11.1. Introducción . . . . .	247
11.2. Modelo de la Regresión no Lineal . . . . .	250
11.3. Regresión no Lineal con R . . . . .	251
11.3.1. Utilización de la función derivada . . . . .	254
11.3.2. Valores iniciales de los parámetros . . . . .	255
11.3.3. Análisis del modelo ajustado . . . . .	257
11.4. Regresión Suavizada . . . . .	259
11.4.1. Regresión Spline . . . . .	261
11.4.2. Regresión Spline con R . . . . .	262
<b>12.Análisis de la Varianza con Medidas Repetidas</b>	<b>265</b>
12.1. Introducción . . . . .	265
12.2. Análisis de la Varianza para un factor y Repetición de una variable	267
12.2.1. Fuentes de variación . . . . .	270
12.2.2. Tratamiento Informático con R . . . . .	277
12.3. Análisis de la Varianza para un factor y Repetición de dos va- riables . . . . .	280
12.4. Referencias . . . . .	281
<b>13.Análisis de Series Temporales</b>	<b>283</b>
13.1. Introducción . . . . .	283
13.1.1. Series temporales con R . . . . .	284
13.2. Elementos básicos en una Serie Temporal . . . . .	286
13.2.1. Tendencia . . . . .	287
13.2.2. Componente Cíclica . . . . .	288
13.2.3. Movimiento Estacional . . . . .	288
13.2.4. Clasificación de las Series temporales . . . . .	288
13.3. Filtrado lineal . . . . .	290
13.4. Series temporales estacionarias . . . . .	295
13.4.1. Procesos Autorregresivos de orden $p$ , $AR(p)$ . . . . .	296
13.4.2. Procesos de Medias Móviles de orden $q$ , $MA(q)$ . . . . .	296
13.4.3. Procesos Autorregresivos de Medias Móviles, $ARMA(p, q)$	297
13.5. Series temporales no estacionarias . . . . .	297

13.5.1. Procesos Autorregresivos Integrados de Medias Móviles, $ARIMA(p, d, q)$ . . . . .	297
13.6. Análisis de una serie temporal . . . . .	298
13.6.1. Identificación del modelo . . . . .	299
13.6.2. Estimación de parámetros . . . . .	302
13.6.3. Diagnósis . . . . .	305
13.6.4. Predicciones . . . . .	305
13.7. Modelos ARIMA . . . . .	307
13.7.1. Identificación del Modelo ARIMA . . . . .	307
13.7.2. Estimación de los parámetros . . . . .	313
13.7.3. Diagnósis . . . . .	315
13.7.4. Predicción . . . . .	317
13.7.5. Test de serie estacionaria . . . . .	319
13.7.6. Ejemplos . . . . .	320
13.8. Cointegración . . . . .	330
13.9. Modelos ARCH y GARCH . . . . .	333
13.10 Ejemplos de series climatológicas . . . . .	338
13.11 Serie Temporal Interrumpida . . . . .	348
13.12 Referencias . . . . .	352
<b>14. Control Estadístico de la Calidad</b>	<b>355</b>
14.1. Introducción . . . . .	355
14.2. Gráfico de control para la media . . . . .	356
<b>15. Data Mining</b>	<b>363</b>
15.1. Introducción y características del Data Mining . . . . .	363
15.1.1. Métodos de Aprendizaje Supervisado y de Aprendizaje no Supervisado . . . . .	364
15.2. El Data Mining y la Inferencia Estadística . . . . .	365
15.3. Tipos de Estructuras en la Base de Datos . . . . .	366
15.3.1. Data Snooping . . . . .	366
15.4. Tareas a realizar en Data Mining . . . . .	367
15.5. Componentes de un análisis Data Mining . . . . .	368
15.6. Estrategias de manejo de Bases de Datos de gran tamaño . . . . .	369
15.6.1. Procesamiento Analítico Automático ( <i>Online Analytical                 Processing OLAP</i> ) y Almacenamiento de Datos ( <i>Data                 Warehousing</i> ) . . . . .	370
15.7. Referencias . . . . .	371
<b>16. Técnicas Estadísticas para Datos Direccionales: Datos Cir- culares</b>	<b>373</b>
16.1. Introducción . . . . .	373

16.2. Análisis Descriptivo . . . . .	374
16.2.1. Coordenadas Polares . . . . .	375
16.2.2. Definición de dirección circular media . . . . .	376
16.2.3. Definiciones de dispersión en datos circulares . . . . .	376
16.3. Distribuciones circulares . . . . .	378
16.3.1. Distribución Uniforme Circular . . . . .	379
16.3.2. Distribución Cardioide . . . . .	379
16.3.3. Distribución de Carthwrite . . . . .	380
16.3.4. Distribución Normal Circular o von Mises . . . . .	381
16.3.5. Distribución Normal Envuelta ( <i>wrapped</i> ) . . . . .	381
16.3.6. Distribución de Cauchy Envuelta ( <i>wrapped</i> ) . . . . .	382
16.4. Inferencias sobre los parámetros de las distribuciones circulares	383
16.4.1. Estimación puntual . . . . .	383
16.4.2. Intervalos de confianza . . . . .	383
16.4.3. Tests para la dirección media y la concentración . . . . .	383
16.5. Tests de Uniformidad . . . . .	384
16.6. Referencias . . . . .	384
<b>17. Inferencias con Mixturas de Distribuciones</b>	<b>387</b>
17.1. Introducción . . . . .	387
17.2. Estimación de los parámetros . . . . .	388
17.2.1. Métodos Clásicos . . . . .	388
17.2.2. Intervalos bootstrap . . . . .	394
17.2.3. Métodos Robustos . . . . .	396
17.3. Revisión del Análisis Cluster . . . . .	397
17.4. Clasificación de individuos: Análisis Discriminante, Análisis de Mixturas, Análisis Cluster y Análisis con Componentes Princi- pales . . . . .	401
17.4.1. Clasificación con el Análisis de Componentes Principales	402
17.5. Referencias . . . . .	402
<b>18. Métodos Estadísticos para Datos de Alta Dimensión</b>	<b>405</b>
18.1. Introducción . . . . .	405
18.2. Regresión LASSO . . . . .	405
18.2.1. Método de ejecución . . . . .	409
18.2.2. Regresión LASSO con R . . . . .	409
18.3. Regresión Ridge . . . . .	419
18.3.1. Regresión Ridge con R . . . . .	420
18.4. Referencias . . . . .	423

---

<b>19. Modelos de Regresión Poisson cero-inflados</b>	<b>425</b>
19.1. Introducción . . . . .	425
19.2. Modelos de Regresión Poisson cero-inflados . . . . .	427
19.2.1. Formulación Matemática de los Modelos cero-inflados . . . . .	428
19.3. Modelos de Regresión Poisson cero-inflados con R . . . . .	429
19.3.1. Comparación de modelos . . . . .	433
19.4. Referencias . . . . .	433
<b>20. Modelos de Efectos Mixtos</b>	<b>435</b>
20.1. Introducción . . . . .	435
20.2. Formulación del Modelo de Efectos Mixtos . . . . .	436
20.2.1. Diversos Modelos de Efectos Mixtos Generalizados . . . . .	439
20.2.2. Modelos de Efectos Mixtos con R . . . . .	439
20.3. Modelos de Mínimos Cuadrados Generalizados . . . . .	439
20.3.1. Estructura de Varianza Fija . . . . .	441
20.3.2. Estructura de Varianza diferente por Grupos . . . . .	443
20.3.3. Estructura de Varianza Potencia . . . . .	445
20.3.4. Estructura de Varianza Exponencial . . . . .	447
20.3.5. Estructura de Varianza Constante más Potencia . . . . .	448
20.3.6. Estructura de Varianza Combinación . . . . .	448
20.4. Modelos para Datos Anidados . . . . .	450
20.5. Un Factor de Efectos Aleatorios . . . . .	452
20.6. Referencias . . . . .	454

# Capítulo 1

## Preliminares y Comparación de Poblaciones

### 1.1. Introducción

Hemos preferido comenzar el texto con un capítulo en el que se formalizarán algunos elementos matemáticos que se utilizarán en posteriores apartados, así como explicar cómo ejecutar la generalización multivariante de la Comparación de Poblaciones, es decir, el contraste en donde se establece la hipótesis nula de igualdad de la media multivariante de varias poblaciones en donde se ha observado esta variable multivariante.

### 1.2. Nombres nuevos para conocidos métodos clásicos

Los primeros capítulos del libro corresponden a lo que suele denominarse *Análisis Multivariante* porque nuestros datos serán observaciones de  $p$  variables aleatorias en los  $n$  individuos de la muestra, en lugar de observaciones de una sola variable aleatoria como ocurría en la mayoría de los métodos de *Análisis Univariante* estudiados en CB o EBR.

Por tanto, la *matriz de datos*, en donde aparecen recogidas las observaciones, es una matriz (es decir, una ordenación por filas y columnas) de la forma

$$\begin{array}{c} \text{Individuos} \end{array} \begin{array}{c} \text{Variables} \\ \left( \begin{array}{ccc} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{array} \right) \end{array}$$

En este tipo de análisis, al igual que ocurría en su homólogo *Análisis Univariante* estudiado en CB o EBR, caben dos formas posibles de estudio: el *Análisis Exploratorio de Datos*, en donde no se utilizan suposiciones ajenas a los datos, tales como modelos para las variables de donde se obtuvieron, y en donde se deja que éstos *hablen por sí mismos*. El propósito de este tipo de análisis es el de descubrir posibles patrones de comportamiento de los datos tales como simetrías, modelos probabilísticos, posibles grupos de datos homogéneos, etc. En él juega un papel especial el uso de gráficos. Los capítulos de Componentes Principales, Análisis de Correspondencias, Escalado Multidimensional y Análisis de Conglomerados serán básicamente de este tipo.

La otra posible vía de estudio de los datos, tanto en el caso univariante como en el multivariante, se denomina *Análisis Confirmatorio de Datos*, en el que se utiliza de forma destacada el contraste de hipótesis como herramienta estadística para la confirmación o rechazo de hipótesis sobre el modelo supuesto. En este caso, la suposición de una distribución normal multivariante para los datos es esencial. La utilización de Métodos Robustos en estas situaciones resulta muy interesante cuando esta suposición de normalidad no se pueda mantener o, al menos, resulte muy dudosa.

Pues bien, el *Análisis Exploratorio de Datos Multivariantes* recibe hoy en día el nombre de *Big Data Analysis* o también *Data Mining* (traducido en ocasiones por *Minería de Datos*), en donde el propósito será, como dijimos más arriba, explorar los datos sin suposiciones adicionales, buscando patrones de comportamiento, clasificaciones en grupos de datos, etc. Dado el gran volumen de datos con el que se suele trabajar en los tiempos actuales, otra característica de este tipo de análisis es el uso intensivo del ordenador, especialmente en la obtención de gráficos.

Una de las razones de realizar un Análisis Multivariante de datos (tanto exploratorio como confirmatorio) en lugar de  $p$  Análisis Univariantes, es el determinar relaciones entre las  $p$  variables de donde se obtuvieron los datos.

Si para descubrir estas estructuras o grupos, cuántos grupos hay, cuáles individuos pertenecen a cada grupo, etc., no utilizamos información previa referente a otros grupos similares de sujetos, se suele hablar de *Estadística no Supervisada*. Con objeto de buscar respuestas a esas preguntas pueden utilizarse ordenaciones, con un *Análisis de Componentes Principales*, o un *Multidimensional Scaling*, o clasificaciones con un *Análisis Cluster*.

Alternativamente, podemos conocer previamente los grupos en los que clasificar los datos, utilizando métodos de *Estadística Supervisada*, tales como el *Análisis Discriminante* o los *Modelos Lineales* o *Modelos Lineales Generalizados* (como los *Modelos log-lineales*, la *Regresión Logística* o la *Regresión Poisson*) o los *Modelos de Regresión no Lineal*.

Los temas sobre *Análisis de la Varianza con Medidas Repetidas*, el *Análisis de Series Temporales* y el *Control Estadístico de la Calidad* tienen un trata-

miento diferenciado del resto.

No obstante, los Métodos Estadísticos que estudiaremos en el libro lo serán de forma individual, ya que éstos no están diseñados habitualmente con un único propósito. Tan solo hemos pretendido enunciar aquí algunos de los nombres que suelen utilizarse hoy en día para asignar a grupos de Métodos Estadísticos y que pueden representar, en el mejor de los casos, el objetivo común para el que van a ser utilizados.

### 1.3. Algunos elementos matemáticos básicos

Esta sección puede resultar un tanto abstracta para lectores que quieren sólo utilizar las técnicas explicadas en el libro, pero nos ha parecido conveniente incluirla. Estos lectores pueden obviarla, al menos en una primera lectura.

Como dijimos más arriba, la matriz de datos está formada por las observaciones de las  $p$  variables en estudio en los  $n$  individuos de la muestra. Estas observaciones serán, por lo general, números reales, es decir, *escalares* aunque, como alguna variable puede ser del tipo cualitativo, como por ejemplo Color de los Ojos, en ocasiones los datos recogidos para esa variable y que forman la correspondiente columna de la matriz de datos, no serán escalares sino *valores* de la forma: Azul, Verde, Castaño, Azul, etc.

No obstante, si queremos utilizar potentes Métodos Estadísticos, las columnas de la matriz de datos deberán estar formadas por números reales, de manera que podamos utilizar técnicas matemáticas estándares. En ese caso, deberemos cuantificar las variables de tipo cualitativo con valores de tipo indicador: 0, 1, etc.

Los escalares los representaremos como hasta ahora, pero a las matrices (como la matriz de datos) las representaremos con letras negritas. Así, hablaremos de la matriz **A**, o de la matriz **B**, etc.

Si **A** es la matriz

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 4 \end{pmatrix}$$

en muchas ocasiones nos interesará trabajar con la matriz *traspuesta* de la anterior, que representaremos como  $\mathbf{A}^t$  y que se define como la matriz en la que sus filas están formadas por las columnas de la dada; es decir, en la que hemos traspuesto las filas y columnas. Así, la matriz traspuesta de la matriz **A** es

$$\mathbf{A}^t = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 3 & 4 \end{pmatrix}$$

ya que, por ejemplo, la que figuraba como primera fila, figura ahora como primera columna, la que figuraba como segunda columna es ahora la segunda fila, etc.

La *dimensión* de una matriz es el número de filas y de columnas por el que está formado (en ese orden). Así, la matriz  $\mathbf{A}$  tiene dimensión  $3 \times 2$  y la matriz  $\mathbf{A}^t$  dimensión  $2 \times 3$ . Una matriz se dice *cuadrada* si ambos valores de su dimensión son iguales; es decir, una matriz  $2 \times 2$  o una  $3 \times 3$  son matrices cuadradas y una  $2 \times 3$  no lo es. Si una matriz coincide con su traspuesta se dice que es *simétrica*.

Una matriz que aparece frecuentemente es la *matriz identidad*,  $\mathbf{I}$ , formada por unos en la diagonal principal y ceros en el resto,

$$\mathbf{I} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

La inversa de una matriz cuadrada  $\mathbf{A}$  se define como una matriz, a la que denominaremos  $\mathbf{A}^{-1}$ , tal que su producto por  $\mathbf{A}$  es la matriz identidad.

Además de los escalares y las matrices, trabajaremos en este texto con *vectores*, que van a ser ordenaciones de datos (habitualmente de tipo numérico), concebidos como columnas. Al igual que con las matrices, representaremos los vectores con letras negritas (de hecho se puede pensar en un vector formado por  $r$  escalares como en una matriz  $r \times 1$ ).

Si  $\mathbf{v}$  es el vector

$$\mathbf{v} = \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}$$

su traspuesto será el vector  $\mathbf{v}^t = (3, 1, 3)$ .

El producto de vectores y/o matrices tiene sentido sólo cuando el segundo valor de la dimensión del primer factor sea igual que el primer valor de la dimensión del segundo factor; el orden es relevante. Así, se puede (pre) multiplicar una matriz  $3 \times 2$  por una matriz  $2 \times 2$ , pero no al revés.

El producto del vector  $\mathbf{v}^t$  por el vector  $\mathbf{w}$ , ambos de longitud, digamos  $m$ , se define como

$$\mathbf{v}^t \mathbf{w} = (v_1, v_2, \dots, v_m) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = v_1 \cdot w_1 + v_2 \cdot w_2 + \cdots + v_m \cdot w_m = \sum_{i=1}^m v_i w_i.$$

La definición del producto de dos matrices y/o vectores  $\mathbf{A}$  y  $\mathbf{B}$  es (cuando se pueda definir el producto) una matriz (o un vector) tal que el elemento que ocupa el lugar  $(i, j)$  (es decir, el que ocupa la fila  $i$ -ésima y la columna  $j$ -ésima) es el resultado de multiplicar la fila  $i$ -ésima de la matriz  $\mathbf{A}$  por la columna  $j$ -ésima de la matriz  $\mathbf{B}$ , consideradas ambas como vectores, de la misma manera que en el párrafo anterior.

La dimensión de la matriz (vector) resultante es el primer valor de la dimensión del primer factor  $\times$  el segundo valor de la dimensión del segundo factor.

Así,  $\mathbf{AB}$  será igual a

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 \cdot 3 + 2 \cdot (-1) & 1 \cdot 1 + 2 \cdot 4 \\ 0 \cdot 3 + 3 \cdot (-1) & 0 \cdot 1 + 3 \cdot 4 \\ 2 \cdot 3 + 4 \cdot (-1) & 2 \cdot 1 + 4 \cdot 4 \end{pmatrix} = \begin{pmatrix} 1 & 9 \\ -3 & 12 \\ 2 & 18 \end{pmatrix}$$

y tendrá dimensión  $3 \times 2$ .

Es interesante que el lector repase cómo multiplicar o invertir matrices en el texto EBR-sección 1.3.3.

## 1.4. Algunos elementos básicos de los vectores aleatorios

Decir que observamos  $p$  variables aleatorias unidimensionales  $X_1, X_2, \dots, X_p$  es lo mismo que decir que observamos el *vector aleatorio*  $\mathbf{X}^t = (X_1, X_2, \dots, X_p)$ . Y al igual que las variables aleatorias unidimensionales tenían su *media* y su *varianza*, las variables aleatorias multidimensionales, o vectores aleatorios, tienen asociados el *vector de medias*, definido como el vector de las medias de las variables que forman el vector aleatorio,

$$\boldsymbol{\mu}^t = (E[X_1], \dots, E[X_p]) = (\mu_1, \dots, \mu_p)$$

y la *matriz de varianzas-covarianzas* (o simplemente matriz de covarianzas), que está formada por las covarianzas entre las variables del vector aleatorio, en donde la covarianza entre las variables  $X_i$  y  $X_j$  se define (CB-sección 4.3 o EBR-sección 4.3) como

$$\text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}$$

siendo la última igualdad, simplemente, una notación abreviada. Si  $i = j$  aparece la varianza de la variable

$$\sigma_{ii} = E[(X_i - \mu_i)^2] = \text{Var}(X_i) = \sigma_i^2$$

Por tanto, la matriz de covarianzas será

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

en donde suele ser  $n > p$ .

Una vez observadas la  $p$  variables en los  $n$  individuos de la muestra, y obtenida así la matriz de datos, el estimador natural del vector de medias poblacional  $\boldsymbol{\mu}$  es el *vector de medias muestrales*  $\bar{\mathbf{x}}$ , siendo

$$\bar{\mathbf{x}}^t = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

en donde  $\bar{x}_i$  es la media de los datos correspondientes a la variable  $i$ -ésima; es decir, la media aritmética de los datos de la columna  $i$ -ésima de la matriz de datos,

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$$

La matriz de varianzas-covarianzas poblacional  $\Sigma$  se estima mediante la *matriz de covarianzas muestral*

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^t$$

en donde  $\mathbf{x}_i$  es la  $i$ -ésima fila de la matriz de datos considerada como vector (es decir, como columna) aleatorio

$$\mathbf{x}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}$$

Por tanto, la matriz de varianzas-covarianzas muestral  $\mathbf{S}$  será la matriz de dimensión  $p \times p$

$$\begin{pmatrix} \sum_{i=1}^n \frac{(X_{i1} - \bar{x}_1)^2}{n-1} & \sum_{i=1}^n \frac{(X_{i1} - \bar{x}_1)(X_{i2} - \bar{x}_2)}{n-1} & \dots & \sum_{i=1}^n \frac{(X_{i1} - \bar{x}_1)(X_{ip} - \bar{x}_p)}{n-1} \\ \sum_{i=1}^n \frac{(X_{i2} - \bar{x}_2)(X_{i1} - \bar{x}_1)}{n-1} & \sum_{i=1}^n \frac{(X_{i2} - \bar{x}_2)^2}{n-1} & \dots & \sum_{i=1}^n \frac{(X_{i2} - \bar{x}_2)(X_{ip} - \bar{x}_p)}{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n \frac{(X_{ip} - \bar{x}_p)(X_{i1} - \bar{x}_1)}{n-1} & \sum_{i=1}^n \frac{(X_{ip} - \bar{x}_p)(X_{i2} - \bar{x}_2)}{n-1} & \dots & \sum_{i=1}^n \frac{(X_{ip} - \bar{x}_p)^2}{n-1} \end{pmatrix}$$

## 1.5. La distribución normal multivariante

Una suposición que habitualmente es necesario realizar, en los capítulos en los que efectuamos Análisis Confirmatorio, es que la variable aleatoria en observación  $p$ -dimensional,  $\mathbf{X} = (X_1, \dots, X_p)^t$  se distribuye según una *distribución normal multivariante*.

Diremos que  $\mathbf{X}$  sigue una *distribución normal multivariante* de dimensión  $p$  con vector de medias  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^t$  y matriz de covarianzas  $\boldsymbol{\Sigma}$ , de dimensión  $p \times p$ , si su función de densidad es

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\{(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}}.$$

Una cuestión central en el segundo volumen es la utilización de Métodos Robustos, para los cuales no es imprescindible tal suposición.

Si quiere visualizar una normal bivalente de vector de medias (0,0) y matriz de varianzas-covarianzas la matriz identidad, de dimensión 2, puede ejecutar en R la siguiente secuencia de instrucciones

```
> library(rgl)
> library(MASS)
> set.seed(31415)
> x<-rnorm(50)
> y<-rnorm(50)
> demo<-kde2d(x, y, n=40)
> xgrid <- demo$x
> ygrid <- demo$y
> z <- dnorm(xgrid)%*%t(dnorm(ygrid))
> spheres3d(x,y,rep(0,50),radius=0.1,color=4)
> surface3d(xgrid,ygrid,z*20,color=3,front="lines")
```

en donde suponemos se ha instalado la librería `rgl`, y en donde hemos fijado la *semilla* en el valor 31415, aunque este comando lo puede evitar obteniendo diferentes representaciones cada vez que ejecute las sentencias enteriores.

También puede modificar las medias y varianzas para obtener diferentes representaciones.

Una de las características de este dibujo es que, una vez obtenido, puede visualizarlo desde varias perspectivas moviendo el ratón del ordenador.

### 1.5.1. Análisis de la normalidad multivariante

Ya que el análisis de la normalidad multivariante de los datos es de gran importancia, parece razonable que estudiemos algún procedimiento para analizar tal suposición. En este sentido, es interesante saber que, si un vector aleatorio sigue una distribución normal multivariante, sus variables aleatorias marginales unidimensionales seguirán distribuciones normales univariantes, pero la afirmación contraria no es cierta necesariamente (sí es cierto, por ejemplo, si existe independencia entre las variables unidimensionales) por lo que, en general, no podremos comprobar la normalidad multivariante de un vector aleatorio analizando la normalidad de las variables unidimensionales que forman el vector; por esta razón, indicamos a continuación dos formas de analizar la normalidad multivariante de unos datos.

#### Test basado en la distancia de Mahalanobis

La distancia que existe desde los  $p$  valores observados en el individuo  $i$ -ésimo  $\mathbf{x}_i$  ( $i$ -ésima fila de la matriz de datos), al vector de valores promedio de la matriz de datos  $\bar{\mathbf{x}}$ , se suele formalizar mediante la denominada *distancia de Mahalanobis*  $d_i$  definida como

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

Pues bien, si  $\mathbf{x}_i$  procede de una normal multivariante de dimensión  $p$ , entonces se podría demostrar que  $d_i^2$  debe de tener, aproximadamente, una distribución  $\chi_p^2$ , por lo que, una forma de analizar si los datos observados  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , proceden de una normal multivariante de dimensión  $p$ , es analizar si puede admitirse que las distancias  $d_i^2$ ,  $i = 1, \dots, n$ , proceden de una  $\chi_p^2$ . Este análisis puede efectuarse mediante un test de bondad del ajuste de Kolmogorov-Smirnov (CB-sección 12.2 o EBR-sección 8.2.2).

#### Ejemplo 1.1

Con la función `mvrnorm`, ejecutando (2), podemos generar una muestra aleatoria de 300 datos procedentes una normal bivalente, con vector de medias (1,1), especificado en el argumento `mu`, y con matriz de varianzas-covarianzas el que aparece en `Sigma`,

$$\begin{pmatrix} 1'5 & 0'8 \\ 0'8 & 1'5 \end{pmatrix}$$

aunque antes deberemos abrir la librería `MASS` con (1) puesto que la función `mvrnorm` está en esta librería.

Una vez obtenidas las 300 distancias de Mahalanobis con (3), ejecutamos un test de Kolmogorov-Smirnov en (4).

```
> library(MASS) (1)
> X<-mvrnorm(300,mu=c(1,1),Sigma=matrix(c(1.5,0.8,0.8,1.5),ncol=2)) (2)
> di<-mahalanobis(X,colMeans(X),var(X)) (3)
> ks.test(di,"pchisq",2) (4)
```

One-sample Kolmogorov-Smirnov test

```
data: di
D = 0.03541, p-value = 0.8462 (5)
alternative hypothesis: two-sided
```

El p-valor dado en (5), asociado a dicho test, sugiere que aceptemos la hipótesis nula de que los datos se ajustan a la normal bivalente. Como se ha obtenido el vector  $\mathbf{X}$  mediante simulación, es posible que obtenga un valor algo distinto cada vez que ejecute ese comando.

### Test de Shapiro-Wilk multivariante

Este test es una generalización del test de normalidad de Shapiro-Wilk estudiado en EBR-suplemento. Se ejecuta con la función

`mshapiro.test(X)`

del paquete `mvnrmtest`. El único argumento de la función es una matriz numérica  $\mathbf{X}$  de, al menos, 3 columnas, es decir, que midamos al menos tres variables. Si  $\mathbf{X}$  no es un dato de la clase matriz, por ejemplo es un *data frame*, debemos transformarlo antes mediante la función `as.matrix`

#### Ejemplo 1.1 (continuación)

Si volvemos a generar una muestra de una normal multivariante con (1), y analizamos su normalidad multivariante con el test de Shapiro-Wilk multivariante mediante (2), el p-valor obtenido en (3) indica aceptar la normalidad de los datos así generados.

```
> library(MASS)
> Y<-mvrnorm(3,mu=c(1,1,1), (1)
+ Sigma=matrix(c(5,1.8,1.3,1.8,1.5,1.3,1.3,1.5),ncol=3)) (1)
> Y
      [,1]      [,2]      [,3]
[1,] 1.03871789 1.5784509 1.994743
[2,] 0.82411003 1.1840550 1.097008
[3,] 0.07795774 0.9076951 1.173389

> library(mvnrmtest)
> mshapiro.test(Y) (2)
```

Shapiro-Wilk normality test

```
data: Z
W = 0.98966, p-value = 0.8055 (3)
```

---

## 1.6. Comparación de dos poblaciones multivariantes

El parámetro de localización de una distribución es el valor que la resume, de manera que una forma habitual de comparar dos o más poblaciones es contrastar si los parámetros de localización de ellas pueden considerarse iguales.

En el caso de dos poblaciones unidimensionales hay una gran variedad de tests analizados en el texto EBR-capítulo 7 (o CB-capítulo 7), según si puede admitirse normalidad de ambas distribuciones, si las muestras son pequeñas o no, etc. Si las poblaciones a comparar son más de dos, la técnica del Análisis de la Varianza, estudiada en EBR-capítulo 9 (o CB-capítulo 8), es la adecuada, o la del Análisis de la Covarianza, EBR-suplemento (o CB-capítulo 11), si se considera alguna variable explicativa adicional.

Incluso en el caso de que no se pueda admitir un modelo para los datos a comparar, diversos tests no paramétricos fueron estudiados en EBR-capítulo 8 (o CB-capítulo 13).

Pero estas situaciones ya estudiadas corresponden a poblaciones univariantes. Si los datos son multivariantes deberemos utilizar técnicas nuevas. En este apartado veremos dos tests para comparar dos poblaciones multivariantes, uno paramétrico primero y, después, uno no paramétrico.

### 1.6.1. Test $T^2$ de Hotelling

El *test de la  $T^2$  de Hotelling* se utiliza para contrastar la hipótesis nula de que las medias (i.e., los vectores de medias) de dos poblaciones normales multivariantes pueden considerarse iguales, frente a la alternativa de no ser iguales. Obsérvese que en cuanto sea diferente la media de una (y sólo una) de las variables unidimensionales que componen el vector aleatorio, el test de Hotelling rechazará la hipótesis nula de poder admitirse que las medias multivariantes sean iguales. Por esta razón es muy difícil que se acepte una hipótesis nula con este test, además de que tener que verificarse la normalidad multivariante de las dos matrices de datos a comparar.

Dos requisitos adicionales para poder aplicar este test son que las matrices de varianzas-covarianzas de ambas poblaciones puedan considerarse iguales y, además, que el número de datos procedentes de ambas normales  $p$ -dimensionales, digamos  $n_1$  y  $n_2$  deben ser tales que  $n_1 + n_2 > p - 1$ .

Para ejecutar este test se utiliza la siguiente función de la librería `Hotelling`

de R,

```
hotelling.test(x,y,data)
```

en donde  $\mathbf{x}$  e  $\mathbf{y}$  son dos matrices de datos de dimensiones  $n_1 \times p$  y  $n_2 \times p$  procedentes de las poblaciones a comparar; y `data` son los datos en formato *data frame*, argumento que sólo es necesario si hay posible confusión en la procedencia de los dos grupos a comparar.

Esta función debemos ejecutarla pidiendo el resultado deseado con

```
> hotelling.test(x,y,data)$stats
> hotelling.test(x,y,data)$pval
```

si queremos obtener, respectivamente, el valor del estadístico de contraste y el p-valor del test.

### Ejemplo 1.2

Los datos `container.df` de la librería `Hotelling` corresponden a los niveles de concentración de nueve elementos (Titanio, Aluminio, Hierro, Manganeso, Magnesio, Calcio, Bario, Estroncio, y Circonio), recogidos de dos envases distintos identificados con la variable `gp`. Se desea averiguar si existen diferencias significativas entre ambos envases utilizando el test  $T^2$  de Hotelling.

Para ello, primero descomponemos los datos en los dos grupos a comparar ejecutando (1) y quitamos la primera columna en ambas matrices de datos, puesto que el grupo al que pertenecen ya lo sabemos, ejecutando (2).

Ejecutamos el test a continuación con (3), que nos proporciona el valor del estadístico de contraste y, con (4), que nos indica el p-valor, tan pequeño en este caso que claramente podemos rechazar la igualdad de todas las medias; es decir, la igualdad de los dos vectores de medias.

```
> library(Hotelling)
> data(container.df)
> Envase1<-container.df[container.df$gp == 1,] (1)
> Envase2<-container.df[container.df$gp == 2,] (1)
> Envase1<-Envase1[,-1] (2)
> Envase2<-Envase2[,-1] (2)
> hotelling.test(Envase1,Envase2)$stats (3)
$statistic
[1] 2043.033

$m
[1] 0.0617284

$df
[1] 9 10

$nx
[1] 10
```

```
$ny
[1] 10
```

```
$p
[1] 9
```

```
> hotelling.test(Envase1,Envase2)$pval (4)
[1] 4.232773e-09
```

---

### 1.6.2. Test de permutaciones

El *test de permutaciones* es el correspondiente test no paramétrico a aplicar cuando el test de Hotelling, acabado de estudiar, no se pueda utilizar, por ejemplo, porque no se verifica la normalidad multivariante de ambas matrices de datos. Para ejecutar este test se utiliza la misma función de R antes estudiada, introduciendo dos nuevos argumentos, `perm`, que debe establecerse en el valor TRUE para ejecutar este test, y `B`, número de permutaciones a utilizar, valor que debe ser alto,

```
hotelling.test(x, y, perm = TRUE, B = 10000, data)$pval
```

#### Ejemplo 1.2 (continuación)

---

Si analizamos la normalidad multivariante con los datos de la primera población a comparar, ejecutando (1), vemos que no se puede admitir su normalidad multivariante con un p-valor tan bajo y, por tanto, no se debe aplicar el test de Hotelling.

Si ejecutamos el test de permutaciones en (2) a los datos anteriores obtenemos que el p-valor es 0 concluyendo, de nuevo, con el rechazo de la igualdad de los vectores de medias de ambas poblaciones.

```
> di<-mahalanobis(Envase1,colMeans(Envase1),var(Envase1)) (1)
> ks.test(di,"pchisq",9) (1)
```

One-sample Kolmogorov-Smirnov test

```
data: di
D = 0.5241, p-value = 0.004387
alternative hypothesis: two-sided
```

```
> hotelling.test(Envase1,Envase2,perm=TRUE,B=10000)$pval (2)
[1] 0
```

---