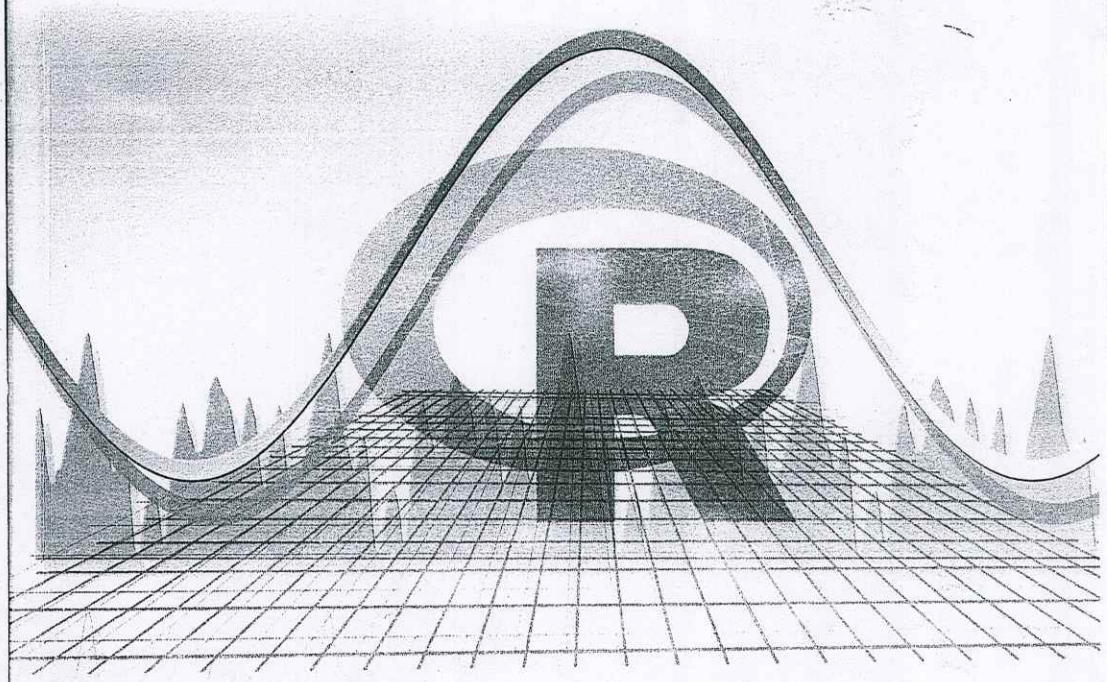
R EN PROFUNDIDAD

Programación, gráficos y estadística



Felicidad Marqués Asensio

Alfaomega

libros R

R en profundidad

Programación, gráficos y estadística

Felicidad Marqués Asensio

Alfaomega

libros R

Diseño de colección y pre-impresión: Grupo RC
Diseño de cubierta: Cuadratín

Datos catalográficos

Marqués, Felicidad
R en profundidad: programación, gráficos y estadística
Primera Edición
Alfaomega Grupo Editor, S.A. de C.V., México

ISBN: 978-607-622-973-6

Formato: 17 x 23 cm Páginas: 444

R en profundidad: programación, gráficos y estadística
Felicidad Marqués Asensio
ISBN: 978-84-944650-2-4 edición original publicada por RC Libros, Madrid, España.
Derechos reservados © 2017 RC Libros

Primera edición: Alfaomega Grupo Editor, México, julio 2017

© 2017 Alfaomega Grupo Editor, S.A. de C.V.
Dr. Isidoro Olvera (Eje 2 sur) No. 74, Col. Doctores, 06720, Ciudad de México.

Miembro de la Cámara Nacional de la Industria Editorial Mexicana
Registro No. 2317

Pág. Web: **http://www.alfaomega.com.mx**
E-mail: **atencionalcliente@alfaomega.com.mx**

ISBN: 978-607-622-973-6

Derechos reservados:
Esta obra es propiedad intelectual de su autor y los derechos de publicación en lengua española han sido legalmente transferidos al editor. Prohibida su reproducción parcial o total por cualquier medio sin permiso por escrito del propietario de los derechos del copyright.

Nota importante:
La información contenida en esta obra tiene un fin exclusivamente didáctico y, por lo tanto, no está previsto su aprovechamiento a nivel profesional o industrial. Las indicaciones técnicas y programas incluidos, han sido elaborados con gran cuidado por el autor y reproducidos bajo estrictas normas de control. ALFAOMEGA GRUPO EDITOR, S.A. de C.V. no será jurídicamente responsable por: errores u omisiones; daños y perjuicios que se pudieran atribuir al uso de la información comprendida en este libro, ni por la utilización indebida que pudiera dársele.

Edición autorizada para venta en México y todo el continente americano.

Impreso en México. Printed in Mexico.

Empresas del grupo:
México: Alfaomega Grupo Editor, S.A. de C.V. – Dr. Isidoro Olvera (Eje 2 sur) No. 74, Col. Doctores, C.P. 06720, Del. Cuauhtémoc, Ciudad de México. Tel.: (52-55) 5575-5022 – Fax: (52-55) 5575-2420 / 2490. Sin costo: 01-800-020-4396 E-mail: atencionalcliente@alfaomega.com.mx
Colombia: Alfaomega Colombiana S.A. – Calle 62 No. 20-46, Barrio San Luis, Bogotá, Colombia, Tels.: (57-1) 746 0102 / 210 0415 – E-mail: cliente@alfaomega.com.co
Chile: Alfaomega Grupo Editor, S.A. – Av. Providencia 1443. Oficina 24, Santiago, Chile
Tel.: (56-2) 2235-4248 – Fax: (56-2) 2235-5786 – E-mail: agechile@alfaomega.cl
Argentina: Alfaomega Grupo Editor Argentino, S.A. – Av. Córdoba 1215 piso 10, C.P. 1055, Buenos Aires, Argentina. – Tel./Fax: (54-11) 4811-0887 y 4811 7183 – E-mail: ventas@alfaomegaeditor.com.ar

ÍNDICE

CAPÍTULO 1. PRIMEROS PASOS EN R. ENTORNO DE TRABAJO 1
INTRODUCCIÓN A R 1
ENTORNO DE TRABAJO DE R 2
INSTALACIÓN DE PAQUETES ESTADÍSTICOS ADICIONALES 3
R STUDIO 5
IMPORTANDO Y EXPORTANDO DATOS 7
ARCHIVOS EN FORMATO R 7
ARCHIVOS EN FORMATO ASCII 9
ARCHIVOS EN FORMATO SPSS, SAS, STATA, S, MINITAB, SYSTAT, OCTAVE Y OTROS FORMATOS 11
CAPÍTULO 2. ELEMENTOS DEL LENGUAJE DE R. NÚMEROS, OPERADORES Y VARIABLES 15
NÚMEROS Y OPERACIONES ARITMÉTICAS EN R 15
NÚMEROS ENTEROS Y FUNCIONES DE VARIABLE ENTERA 16
NÚMEROS REALES Y FUNCIONES DE VARIABLE REAL 16
FUNCIONES TRIGONOMÉTRICAS 17
FUNCIONES HIPERBÓLICAS 18
FUNCIONES EXPONENCIALES Y LOGARÍTMICAS 18
FUNCIONES ESPECÍFICAS DE VARIABLE NUMÉRICA 19

VARIABLES VECTORIALES 20
ELEMENTOS DE VARIABLES VECTORIALES 22
FUNCIONES DE VARIABLES VECTORIALES 24
FACTORES 26
VARIABLES MATRICIALES 28
ELEMENTOS DE VARIABLES MATRICIALES 29
FUNCIONES MATRICIALES ESPECÍFICAS 32
FUNCIONES DE VARIABLES MATRICIALES 34
OPERADORES 38
OPERADORES ARITMÉTICOS 38
OPERADORES LÓGICOS 39
OPERADORES RELACIONALES 40
FUNCIONES DE INFORMACIÓN 42
FUNCIONES DE CONVERSIÓN 43
VARIABLES Y FUNCIONES DE CADENA 44
FUNCIONES DE SELECCIÓN Y MANIPULACIÓN 47
CAPÍTULO 3. LENGUAJE DE R. ESTRUCTURAS VECTORIALES Y MATRICIALES PARA ANÁLISIS DE DATOS 61
INTRODUCCIÓN 61
SECUENCIAS 61
ARRAYS 65
LISTAS 69
MARCOS DE DATOS 72
IMPORTACIÓN DE DATOS ESTADÍSTICOS COMO DATA FRAMES 78
IMPORTACIÓN DE ARCHIVOS DE TEXTO 78
IMPORTACIÓN DE ARCHIVOS EXCEL 80
IMPORTACIÓN DE ARCHIVOS SPSS 81
IMPORTACIÓN DE ARCHIVOS DE STATA 84
IMPOTACIÓN DE ARCHIVOS DE SAS 85
SERIES DE TIEMPO 85

© Alfaomega - RC Libros

CAPÍTULO 4. LENGUAJE DE R. FUNCIONES PREDEFINIDAS Y CREADAS A MEDIDA 93
FUNCIONES 93
DEFINICIÓN DE FUNCIONES 93
EL EDITOR DE R 94
FUNCIONES MATEMÁTICAS PREDEFINIDAS EN R 101
FUNCIONES MATEMÁTICAS ELEMENTALES DE VARIABLE REAL 101
FUNCIONES DE ANÁLISIS ESTADÍSTICO BÁSICO CON ARGUMENTO MATRICIAL 103
CAPÍTULO 5. LENGUAJE DE R. CONTROL DE FLUJO, CONDICIONALES Y OTRAS ESTRUCTURAS 117
CONTROL DE FLUJO 117
EL BUCLE FOR 117
EL BUCLE WHILE 119
LA ESTRUCTURA CONDICIONAL IF ELSE 121
LA ESTRUCTURA CONDICIONAL IFELSE 124
LA ESTRUCTURA SWITCH 125
NEXT 125
BREAK 126
LA ESTRUCTURA REPEAT 127
ALGORITMOS DE CÁLCULO NUMÉRICO 128
MÉTODO DE NEWTON PARA RESOLVER LA ECUACIÓN g(x)=0 128
MÉTODO DE SCHRODER PARA RESOLVER LA ECUACIÓN g(x)=0 130
MÉTODO DEL TRAPECIO PARA LA INTEGRACIÓN NUMÉRICA 130
MÉTODO DE SIMPSON PARA LA INTEGRACIÓN NUMÉRICA 133
CAPÍTULO 6. ANÁLISIS EXPLORATORIO DE DATOS A TRAVÉS DE R. EXPLORATORIO GRÁFICO 147
ANÁLISIS EXPLORATORIO DE DATOS 147
HISTOGRAMA DE FRECUENCIAS 148
DIAGRAMA DE TALLO Y HOJAS 150
GRÁFICO DE CAJA Y BIGOTES 151
GRÁFICO MÚLTIPLE DE CAJA Y BIGOTES 153

© Alfaomega - RC Libros

GRÁFICO DE SIMETRÍA Y NORMALIDAD 156
GRÁFICO DE DISPERSIÓN 156
GRÁFICOS PARA VARIABLES CUALITATIVAS 158
COMANDOS PARA GRÁFICOS EXPLORATORIOS EN R. SINTAXIS BÁSICA 161
COMANDOS PARA GRÁFICOS EXPLORATORIOS EN R. SINTAXIS AVANZADA 171
GRÁFICOS DE CAJA Y BIGOTES 171
GRÁFICOS DE BARRAS 175
GRÁFICOS DE SECTORES 179
HISTOGRAMAS 181
GRÁFICOS DE PUNTOS 184
CAPÍTULO 7. ANÁLISIS EXPLORATORIO FORMAL. ESTADÍSTICA DESCRIPTIVA CON R 195
ESTADÍSTICA DESCRIPTIVA 195
MEDIDAS DE CENTRALIZACIÓN: MEDIA, MEDIANA, MODA, CUANTILES, PERCENTILES Y MOMENTOS 195
MEDIDAS DE DISPERSIÓN 201
MEDIDAS DE FORMA 205
COEFICIENTE DE CORRELACIÓN LINEAL ENTRE DOS VARIABLES. REGRESIÓN SIMPLE MÍNIMO CUADRÁTICA 209
COEFICIENTE DE CORRELACIÓN POR RANGOS 212
COVARIANZA 212
DISTRIBUCIONES DE MÁS DE DOS DIMENSIONES: MATRIZ DE COVARIANZAS Y MATRIZ DE CORRELACIONES 213
FUNCIONES BÁSICAS DE R PARA ESTADÍSTICA DESCRIPTIVA 215
CAPÍTULO 8. PROBABILIDAD. DISTRIBUCIONES DISCRETAS Y CONTINUAS 225
CONCEPTO DE PROBABILIDAD: REGLA DE LAPLACE 225
PROBABILIDAD CONDICIONADA E INDEPENDENCIA: TEOREMA DE BAYES 226
VARIABLE ALEATORIA DISCRETA Y DISTRIBUCIÓN DE PROBABILIDAD DE TIPO DISCRETO 227
DISTRIBUCIONES DISCRETAS DE UNA Y DOS DIMENSIONES: CARACTERÍSTICAS 229
DISTRIBUCIONES BINOMIAL, DE POISSON, GEOMÉTRICA E HIPERGEOMÉTRICA 231
DISTRIBUCIÓN MULTINOMIAL 234

© Alfaomega - RC Libros

DISTRIBUCIÓN HIPERGEOMÉTRICA GENERALIZADA 235
DISTRIBUCIÓN UNIFORME DISCRETA 236
R Y LAS DISTRIBUCIONES DISCRETAS 237
VARIABLES ALEATORIAS CONTINUAS 240
CARACTERÍSTICAS DE LAS DISTRIBUCIONES CONTINUAS 242
VARIABLES ALEATORIAS CONTINUAS BIDIMENSIONALES. INDEPENDENCIA 243
DISTRIBUCIONES CONTINUAS RELEVANTES, NORMAL, EXPONENCIAL, $\chi 2$, T, F, ETC. 244
DESIGUALDAD DE TCHEBICHEV 248
TEOREMA CENTRAL DEL LÍMITE Y TEOREMA DE MOIVRE 249
R Y LAS DISTRIBUCIONES CONTINUAS 250
CAPÍTULO 9. INTERVALOS DE CONFIANZA Y TESTS DE HIPÓTESIS 275
ESTIMACIÓN POR INTERVALOS 275
INTERVALOS DE CONFIANZA PARA PARÁMETROS DE DISTRIBUCIONES NORMALES 276
INTERVALOS DE CONFIANZA PARA COMPARAR POBLACIONES NORMALES 277
RELACIÓN ENTRE INTERVALOS DE CONFIANZA Y CONTRASTES DE HIPÓTESIS 278
CONTRASTES DE HIPÓTESIS PARA POBLACIONES NORMALES 279
COMPARACIÓN DE DOS POBLACIONES NORMALES CON DATOS APAREADOS 281
EL LENGUAJE R Y LOS INTERVALOS DE CONFIANZA Y CONTRASTES DE HIPÓTESIS 284
CONTRASTES NO PARAMÉTRICOS 289
CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN: CONTRASTES DEL COEFICIENTE DE CORRELACIÓN Y MEDIDAS DE ASOCIACIÓN 289
CONTRASTE CHI-CUADRADO DE INDEPENDENCIA 291
CONTRASTES MÚLTIPLES DE HOMOGENEIDAD: TEST DE LA $\chi 2$ 293
CONTRASTES DE LA BONDAD DE AJUSTE: TEST DE LA CHI-CUADRADO 294
CONTRASTE DE KOLMOGOROV-SMIRNOV DE LA BONDAD DE AJUSTE 296
CONTRASTES DE ALEATORIEDAD: TEST DE RACHAS Y TEST DE DANIEL 298
EL LENGUAJE R Y LOS CONTRASTES NO PARAMÉTRICOS 300

© Alfaomega - RC Libros

CAPÍTULO 10. GRÁFICOS 2-D Y 3-D EN R **329**
GRAFICOS BÁSICOS EN R. EL COMANDO PLOT PARA GRÁFICOS DE SECUENCIA Y DISPERSIÓN 329
CURVAS PLANAS EN COORDENADAS PARAMÉTRICAS 333
CURVAS PLANAS EN COORDENADAS EXPLÍCITAS 333
CURVAS PLANAS EN COORDENADAS POLARES 339
GRÁFICOS DE SUPERFICIES 340
VENTANAS GRÁFICAS EN R 345
CAPÍTULO 11. TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN A TRAVÉS DE R **359**
TRANSFORMACIÓN DE DATOS MEDIANTE TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN 359
COMPONENTES PRINCIPALES 360
CÁLCULO DE LAS COMPONENTES PRINCIPALES 362
PUNTUACIONES O MEDICIÓN DE LAS COMPONENTES 364
NÚMERO DE COMPONENTES PRINCIPALES A RETENER 364
MATRIZ DE CARGAS FACTORIALES, COMUNALIDAD Y CÍRCULOS DE CORRELACIÓN 365
ROTACIÓN DE LAS COMPONENTES 367
ANÁLISIS FACTORIAL 367
CONTRASTES EN EL MODELO FACTORIAL 370
ROTACIÓN DE LOS FACTORES 372
INTERPRETACIÓN GRÁFICA DE LOS FACTORES 374
PUNTUACIONES O MEDICIÓN DE LOS FACTORES 374
ANÁLISIS DE COMPONENTES PRINCIPALES A TRAVÉS DE R 375
COMPONENTES PRINCIPALES A TRAVÉS DE COMANDOS 380
ANÁLISIS FACTORIAL A TRAVÉS DE R 383
ANÁLISIS FACTORIAL A TRAVÉS DE COMANDOS 385
CAPÍTULO 12. TÉCNICAS DE CLASIFICACIÓN Y SEGMENTACIÓN A TRAVÉS DE R **395**
EL ANÁLISIS CLÚSTER 395
CLÚSTERES JERÁRQUICOS, SECUENCIALES, AGLOMERATIVOS Y EXCLUSIVOS (S.A.H.N.) 397
EL DENDOGRAMA EN EL ANÁLISIS CLÚSTER JERÁRQUICO 398

© Alfaomega - RC Libros

ANÁLISIS CLÚSTER NO JERÁRQUICO 398
ANÁLISIS CLÚSTER JERÁRQUICO A TRAVÉS DE R 402
ANÁLISIS CLÚSTER A TRAVÉS DE COMANDOS 403
ANÁLISIS CLÚSTER NO JERÁRQUICO A TRAVÉS DE R 407
CAPÍTULO 13. MODELOS PREDICTIVOS A TRAVÉS DE R **411**
MODELO LINEAL DE REGRESIÓN MÚLTIPLE 411
HIPÓTESIS EN EL MODELO LINEAL 412
HIPÓTESIS RELATIVAS A LA PERTURBACIÓN ALEATORIA 412
HIPÓTESIS RELATIVAS A LOS REGRESORES 413
HIPÓTESIS RELATIVAS AL VECTOR DE PARÁMETROS 413
HIPÓTESIS RELATIVAS A LA FORMA FUNCIONAL 413
ESTIMACIÓN DEL MODELO LINEAL POR MÍNIMOS CUADRADOS ORDINARIOS MCO 414
ANÁLISIS DE LOS RESIDUOS 416
EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE EN R 418
REGRESIÓN POLINÓMICA 423
REGRESIÓN CON INTERACCIONES 424
REGRESIÓN SIN CONSTANTE 424
ÍNDICE ANALÍTICO **427**

© Alfaomega - RC Libros

CAPÍTULO 1

PRIMEROS PASOS EN R. ENTORNO DE TRABAJO

INTRODUCCIÓN A R

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación GNU del clásico lenguaje científico S-Plus. R y S-Plus son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico. R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas. Al igual que S-Plus, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C, C++ o Fortran que se cargan dinámicamente. Los

© Alfaomega - RC Libros

desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

R forma parte de un proyecto colaborativo y abierto. Sus usuarios pueden publicar paquetes que extienden su configuración básica. Dado el enorme número de nuevos paquetes, estos se han organizado en vistas (o temas), que permiten agruparlos según su naturaleza y función. Por ejemplo, hay grupos de paquetes relacionados con estadística bayesiana, econometría, series temporales, etc.

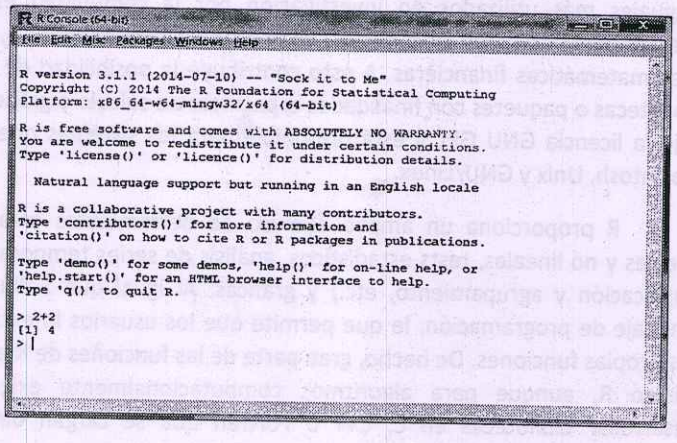
R, al igual que S-Plus, puede considerarse un lenguaje de programación orientado a objetos. Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. R también posee su propio formato para la documentación basado en LaTeX.

Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad.

R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como GNU Octave y su equivalente comercial, MATLAB.

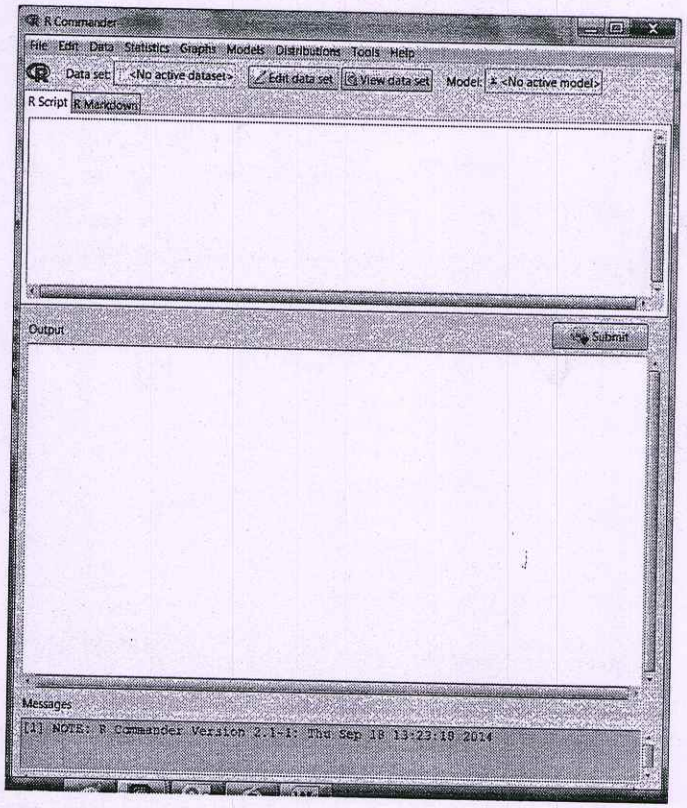
ENTORNO DE TRABAJO DE R

En R se puede trabajar en modo comando o en modo de ventanas interactivas. El modo comando es el modo por defecto (*R Console*).



© Alfaomega - RC Libros

El modo de ventanas interactivas *R Commander* es el que se obtiene después de instalar y leer el paquete *Rcmdir*. El R Commander aparece con 3 ventanas destinadas a instrucciones (*R Script*), resultados (*Output*) y mensajes de error o aclaratorios (*Messages*).

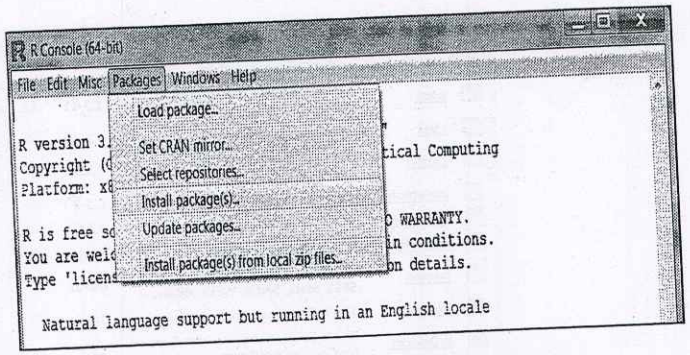


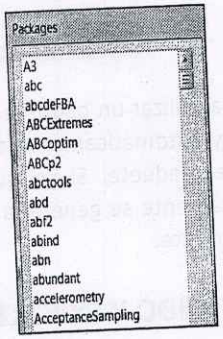
INSTALACIÓN DE PAQUETES ESTADÍSTICOS ADICIONALES

Hasta aquí hemos utilizado solamente el software básico de R. Pero existen multitud de paquetes adicionales que amplían el programa en muchas facetas de trabajo. Para instalar un paquete adicional, en el menú *Paquetes*, hacer clic en *Install package(s)*. A continuación se elige *Espejo* en la ventana *CRAN mirror* (a poder ser el más cercano), se pulsa *Ok* y en la pantalla *Packages* se eligen los paquetes a instalar.

© Alfaomega - RC Libros

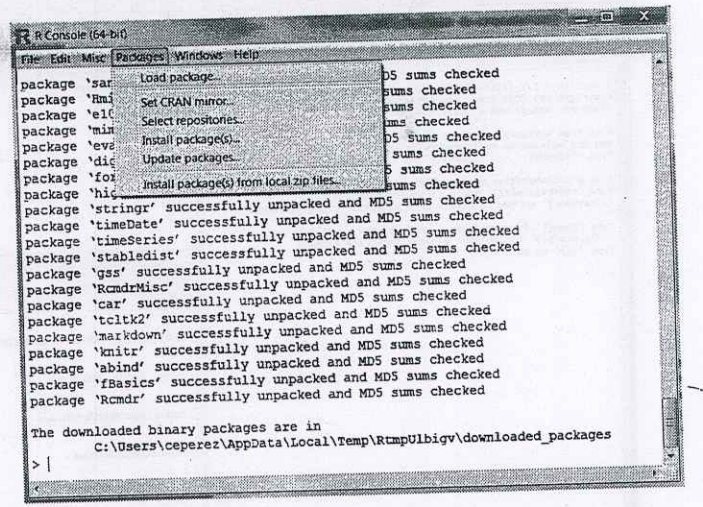
En nuestro caso, elegiremos inicialmente *fBasics* (estadística básica) y *Rcmdr* (ventana interactiva de comandos de R). Al hacer clic en Ok se instalan los paquetes.





De esta forma se van instalando paquetes de programas adicionales al paquete base. Este proceso de instalar paquetes se hace solo la primera vez que se instala un paquete. Una vez instalados, se elige en el menú *Packages* la opción *Load package* y en la lista de paquetes resultantes se eligen otra vez *fBasics* y *Rcmdr*, para ser utilizados. De esta forma se sitúan como activos los paquetes disponibles que tengamos instalados y queramos utilizar en una sesión de R. Los paquetes a usar deben ser cargados con *Load package* cuando se inicia la sesión de R.

© Alfaomega - RC Libros

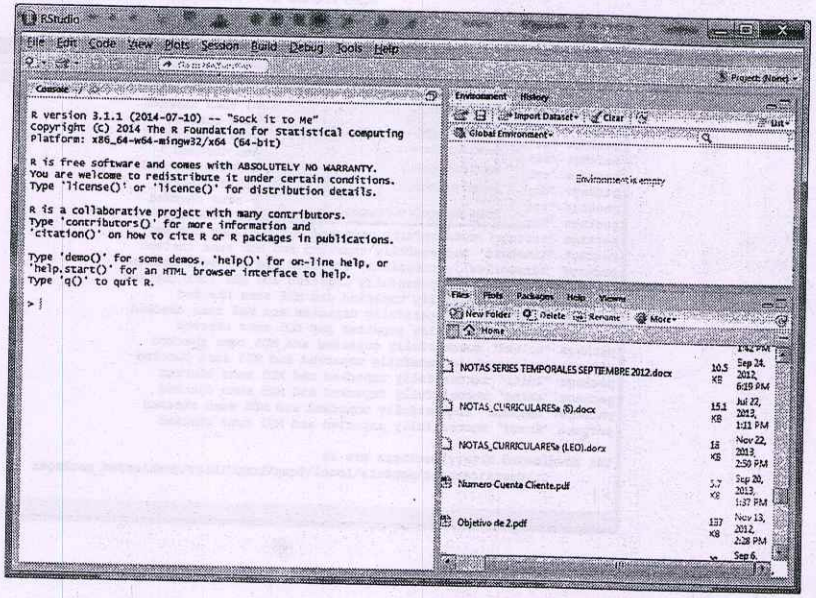


R STUDIO

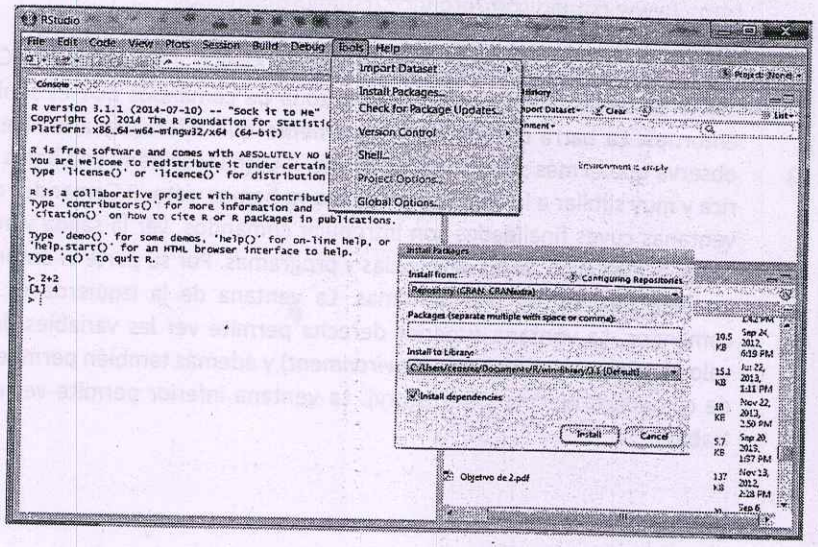
R Studio es un entorno de R que complementa *R Console* y *Rcomander*. Se trata de software libre que puede ser obtenido en la página oficial del programa http://www.rstudio.com/products/rstudio/download/

El entorno de trabajo de R Studio es similar a R Console y R Comander. A la izquierda de la pantalla se observa la zona de comandos que es similar en los tres entornos. La barra de menú es el elemento que diferencia a los tres entornos. Se observa que el más sencillo es R Console. R Comander tiene una barra de menús más rica y muy similar a la de R Studio. Como ya hemos visto, R Comander dispone de tres ventanas cuyas finalidades son introducir comandos, ver la salida y ver comentarios sobre la ejecución de las sentencias y programas. Por su parte el entorno de R Studio también dispone de tres ventanas. La ventana de la izquierda es la ventana de comandos. La ventana superior derecha permite ver las variables de entorno y su valor según se va trabajando (*Environment*) y además también permite ver el historial de comandos ejecutados (*History*). La ventana inferior permite ver el directorio de trabajo.

© Alfaomega - RC Libros

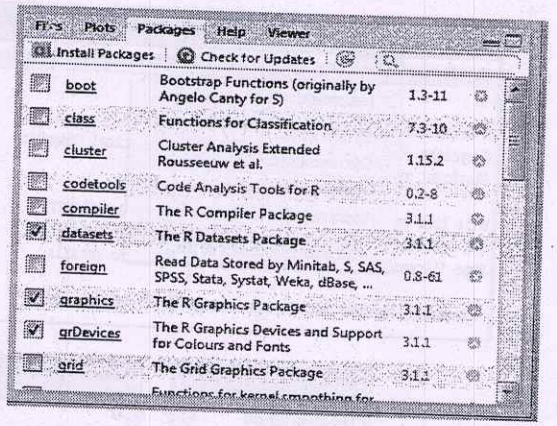


La opción *Install Packages* del menú *Tools* permite instalar paquetes adicionales en el entorno de R Studio. Basta poner su nombre en el campo *Packages*. También se puede elegir el origen de los paquetes y su destino en disco.



© Alfaomega - RC Libros

Una vez instalado un paquete, la solapa *Packages:* de la parte inferior derecha de la ventana de R Studio permite ver el listado de paquetes disponibles (están marcados en la lista) y los ya instalados. Para habilitar cualquiera de ellos basta marcarlo.



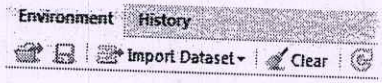
Para utilizar un paquete nuevo, por ejemplo *cluster*, basta marcarlo en la lista packages y automáticamente en la ventana de comandos se ejecuta el comando de lectura del paquete. Si se quita la marca de un paquete en la lista packages, automáticamente se genera la sintaxis en la ventana de comandos que inhabilita el uso del paquete.

IMPORTANDO Y EXPORTANDO DATOS

En R se pueden leer datos de ficheros externos con formatos adecuados para la mayoría de los paquetes estadísticos. Es posible leer ficheros ASCII, SPSS, STATA, etc. No obstante, R dispone de su propio formato de fichero con extensión RData. También es posible escribir ficheros en formato R.

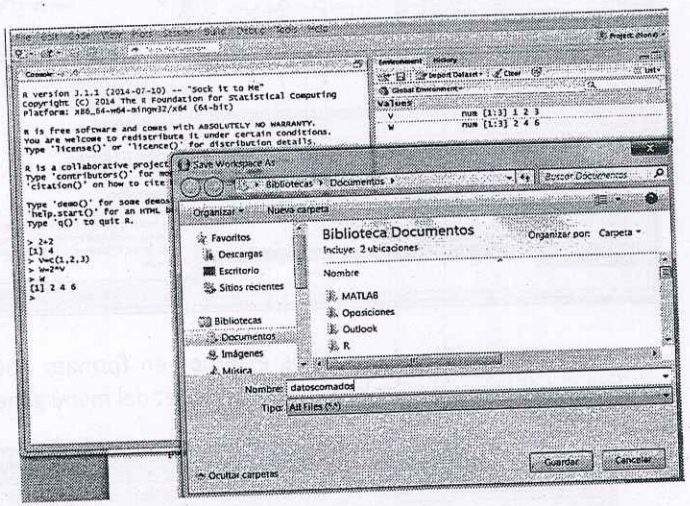
Archivos en formato R

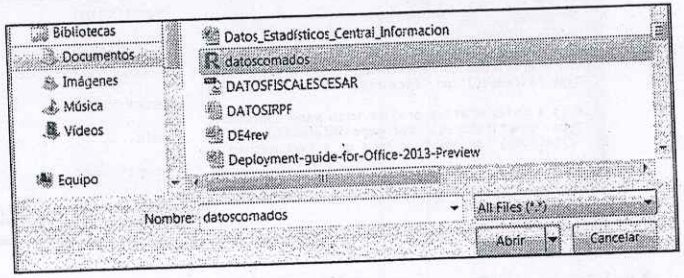
Las opciones de la barra de menú de la pantalla Environment del entorno de R Studio permiten realizar tareas de importación y exportación de datos.



© Alfaomega - RC Libros

El icono permite guardar la información actual de la ventana de comandos en un archivo en formato de R (extensión RData) que posteriormente puede ser cargado en memoria con el icono al iniciar otra sesión. Las figuras siguientes ilustran la tarea de guardar la información del entorno de trabajo en el fichero *datoscomandos.RData* y de su posterior lectura.

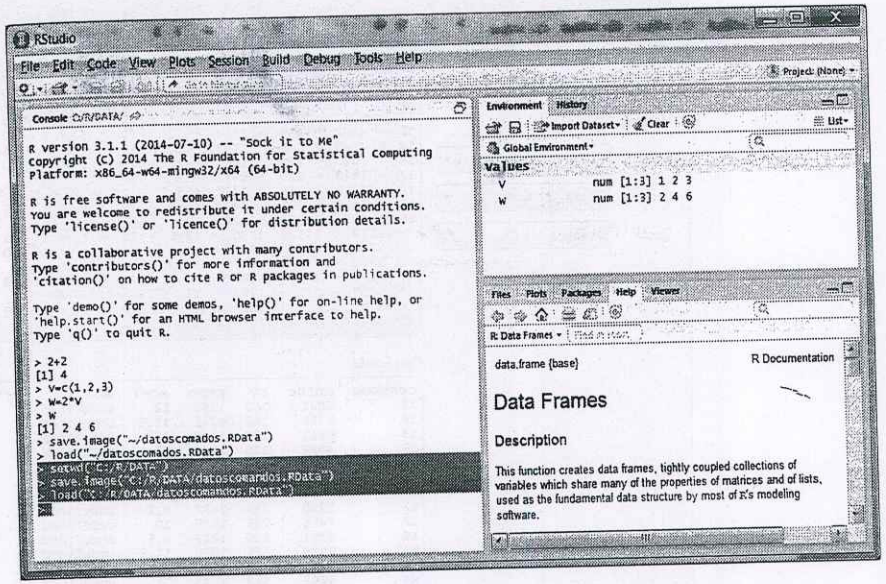




En la ventana de comandos se observa la sintaxis de R equivalente a las tareas de situar un directorio de trabajo, guardar el espacio de trabajo a un archivo de R y leer posteriormente esa información guardada del espacio de trabajo.

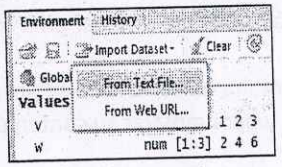
```
> setwd("C:/R/DATA")
> save.image("C:/R/DATA/datoscomandos.RData")
> load("C:/R/DATA/datoscomandos.RData")
```

© Alfaomega - RC Libros



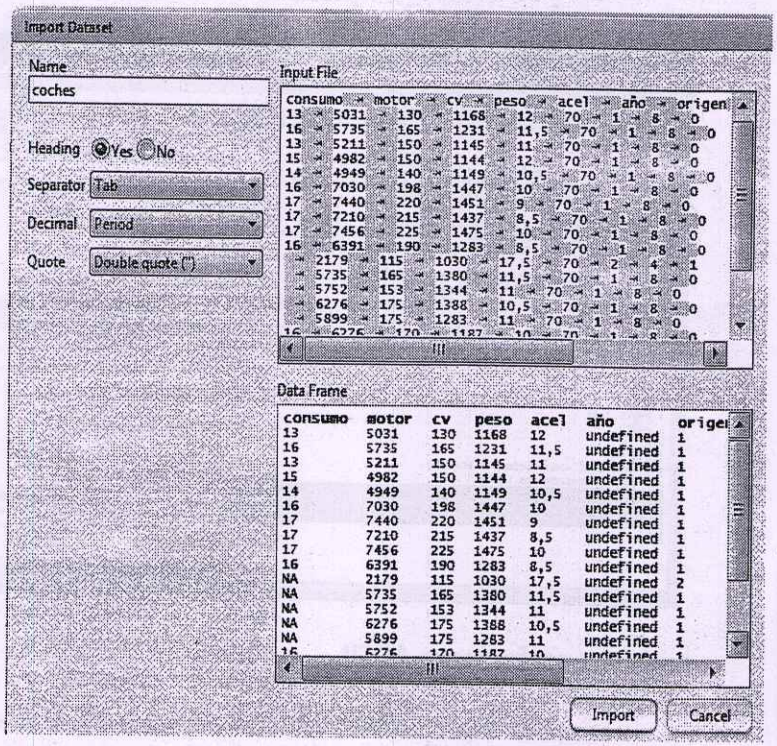
Archivos en formato ASCII

La opción *Import Dataset* del menú *Environment* permite importar ficheros externos en formato ASCII y ficheros web.



La opción *From Text File* de *Import Dataset* permite importar ficheros ASCII externos. Al realizar la importación, R presenta una imagen del contenido del fichero y las opciones de utilizar nombres de variables en la primera línea del fichero (*Heading*), elegir un separador de caracteres concreto (tabulador, espacio en blanco, coma y punto y coma), representar los decimales con puntos o con comas e introducir los caracteres entre comillas simples o dobles.

© Alfaomega - RC Libros

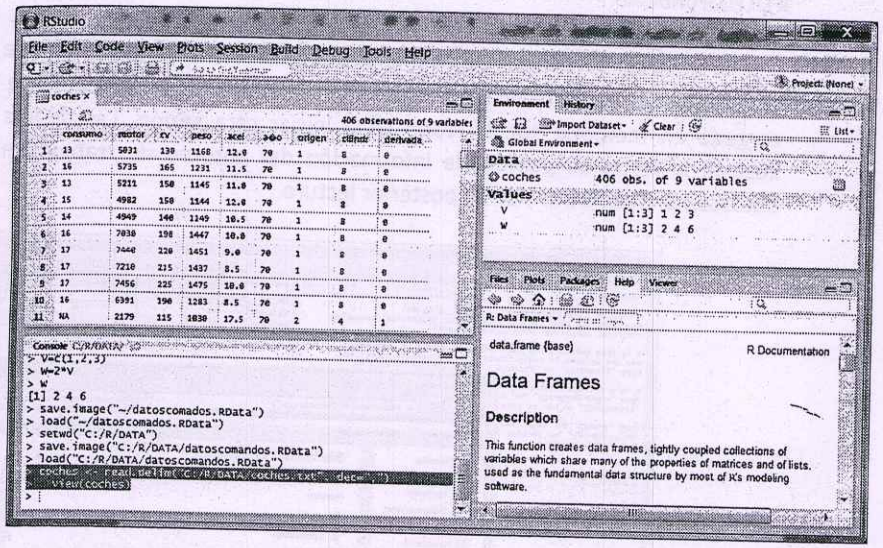


Al pulsar *Import* se produce la importación del fichero ASCII al espacio de trabajo de R y ya es posible trabajar en R con su información.

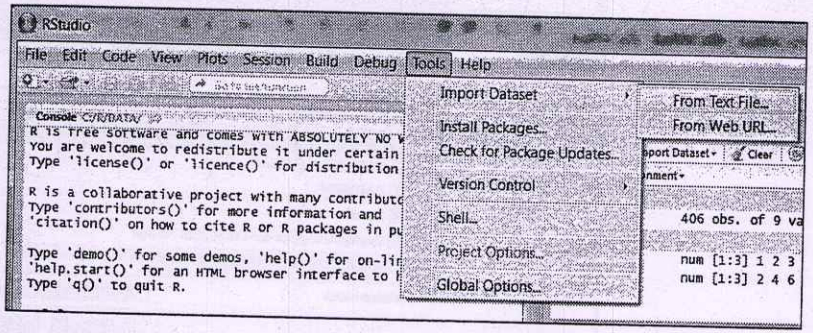
En la ventana de comandos se observa la sintaxis que se utiliza para importar un fichero ASCII y para mostrar su contenido en el espacio de trabajo.

```
> coches <- read.delim("C:/R/DATA/coches.txt", dec=",")
> View(coches)
```


© Alfaomega - RC Libros



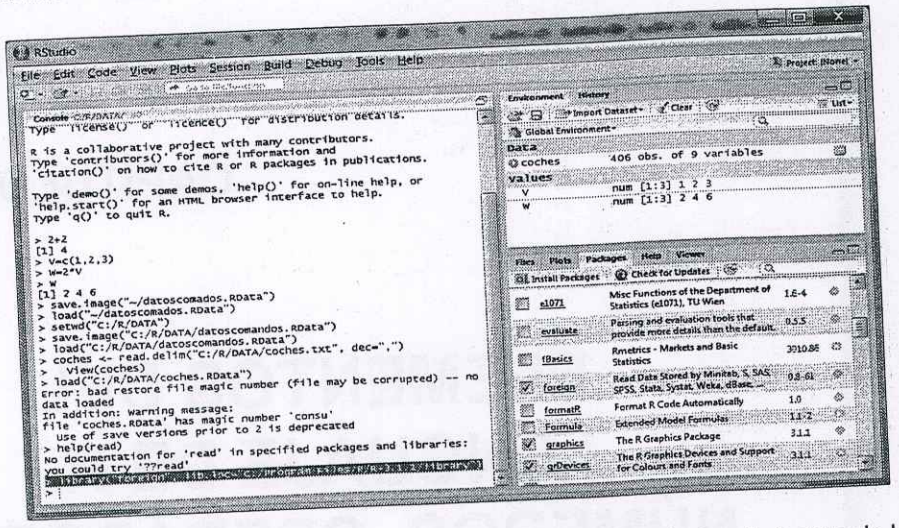
La importación de archivos externos en formato ASCII también se puede realizar desde la opción *Tools* → *Import Data Set* del menú general de R Studio.



Archivos en formato SPSS, SAS, STATA, S, MINITAB, SYSTAT, OCTAVE y otros formatos

R puede leer archivos en formatos SPSS, SAS, STATA, S, MINITAB, SYSTAT, OCTAVE y otros formatos a través de la librería *foreign*. Para ello se comienza habilitando la librería a través de la opción *Packages* del entorno de R Studio (o instalándola a través de la subopción *Install Packages* de la opción *Tools* del barra de menús de R Studio, en caso de que no esté instalada).

© Alfaomega - RC Libros

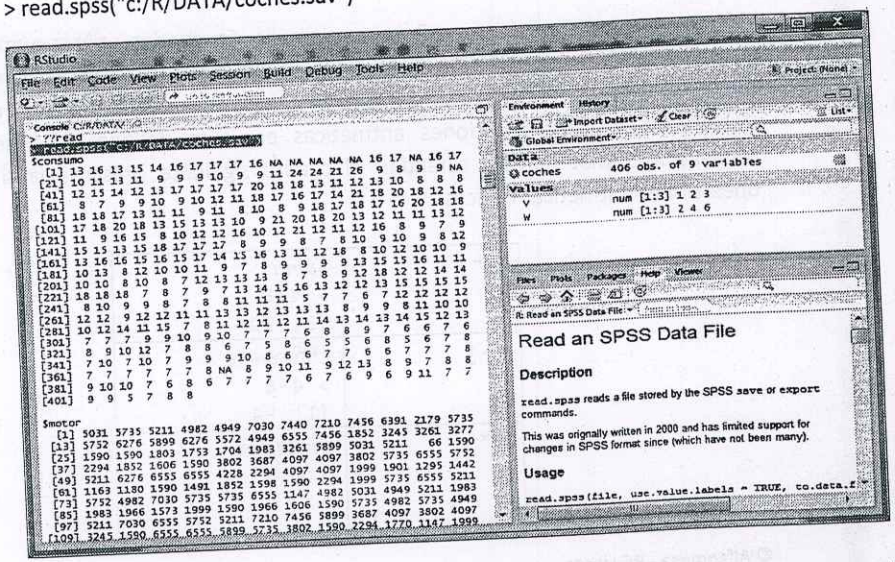


En la figura anterior se observa que la sintaxis equivalente a la lectura de la librería es la siguiente:

> library("foreign", lib.loc="C:/Program Files/R/R-3.1.1/library")

A continuación se puede utilizar la función *read.spss* para leer el fichero.

> read.spss("c:/R/DATA/coches.sav")



© Alfaomega - RC Libros

De la misma forma se pueden leer ficheros de los formatos SAS, STATA, S, SYSTAT, MINITAB, SAS Transport y OCTAVE utilizando respectivamente las funciones *read.sas7badt, read.dta, read.S, read.systat, read.mto, read.xport* y *read.octave* y read.dta de la librería *foreign*. La sintaxis más general para todas estas funciones es *read.formato("archivo")*. La sintaxis del comando read utilizada aquí es la más sencilla posible. Una sintaxis ampliada podría ser la siguiente:

read.spss(file, use.value.labels = TRUE, to.data.frame = FALSE, max.value.labels = Inf, trim.factor.names = FALSE, trim_values = TRUE, reencode = NA, use.missings = to.data.frame)

La opción *use.value.labels = TRUE* permite convertir los valores de las etiquetas en factores de R con sus correspondientes niveles.

La opción *to.data.frame = TRUE* permite convertir el fichero de datos en una estructura de marco de datos (*dataframe*), que será explicada posteriormente.

La opción *max.value.labels = inf* permite convertir a factores todas las variables independientemente de su etiqueta, ya que no se acotan sus valores.

La opción *trim.factor.names = TRUE* permite comprimir los espacios en los niveles de factor.

La opción *trim_values = TRUE* permite comprimir los espacios de los valores de las etiquetas.

La opción *reencode = NA* permite recodificar como NA la información que falta.

La opción *use.missings = to.data.frame* permite utilizar como desaparecidos los elementos que tienen el valor definido en *reencode* (habitualmente NA).

La sintaxis es similar para todos los comandos *read.formato(fichero)* indepedientemente de quién sea el formato.

© Alfaomega - RC Libros

CAPÍTULO 2

ELEMENTOS DEL LENGUAJE DE R. NÚMEROS, OPERADORES Y VARIABLES

NÚMEROS Y OPERACIONES ARITMÉTICAS EN R

El trabajo numérico en R es muy amplio. Se pueden utilizar números enteros, reales y complejos, que a su vez pueden ser argumentos de funciones dando lugar a las funciones enteras, reales y complejas. Por tanto, la variable compleja es un campo tratable en R. Las operaciones aritméticas en R se definen de acuerdo con las convenciones matemáticas estándar. El cuadro siguiente presenta la sintaxis de las operaciones aritméticas básicas:

x + y	*Suma*
	`> 4+8`
	`[1] 12`

x - y	*Diferencia*
	`> 4-8`
	`[1] -4`

x * y	*Producto* `> 4*8`

© Alfaomega - RC Libros

x/y	*División*
	`> 4/8`
	`[1] 0.5`

x ^y o xy**	*Potencia*
	`> 4**8`
	`[1] 65536`

NÚMEROS ENTEROS Y FUNCIONES DE VARIABLE ENTERA

R trabaja con números enteros y con funciones de variable entera de forma exacta. Independientemente del formato de presentación de los resultados, los cálculos son exactos. De todas formas, el comando *vpa* permite obtener salidas exactas con la precisión que se requiera.

En cuanto a las funciones con variable entera, las más importantes que contempla R son las siguientes (las expresiones entre comillas tienen formato cadena):

sign(n)	*Signo de n (1 si n>0 , -1 si n<0, n real)*
	`> sign(-8)`
	`[1] -1`

max(n1,n2)	*Máximo de los números n1 y n2*
	`> max(17,12)`
	`[1] 17`

min(n1,n2)	*Mínimo de los números n1 y n2*
	`> min(17,12)`
	`[1] 12`

factorial(n)	*Factorial de n (n(n-1)(n-2)...1)*
	`> factorial(4)`
	`[1] 24`

choose(n,k)	*Número combinatorio n sobre k*
	`> choose(5,2)`
	`[1] 10`

NÚMEROS REALES Y FUNCIONES DE VARIABLE REAL

Los números reales son la unión disjunta de los números racionales y los números irracionales. Un número racional es de la forma *p/q*, donde *p* es un entero y *q* otro entero. La forma en que R trata los racionales es distinta a la de la mayoría de calculadoras. Los números racionales son cocientes de enteros, y R también puede trabajar con ellos en modo exacto, de manera que el resultado de expresiones en las que intervienen números racionales es siempre otro número racional o entero.

© Alfaomega - RC Libros

R devuelve aproximaciones decimales de los resultados. Consideramos el siguiente ejemplo de operaciones entre números racionales:

```
> 1/6+1/5-2/10
[1] 0.1666667
```

R también trabaja con los números irracionales representando los resultados con la mayor precisión que puede. A continuación se presenta un ejemplo.

```
> sqrt(235)
[1] 15.32971
```

Existen constantes reales muy típicas que R representa como sigue:

pi	*Número π = 3,1415926*
	`> 2*pi`
	`[1] 6.283185`

exp(1)	*Número e = 2,7182818*
	`> exp(1)`
	`[1] 2.718282`

inf	*Infinito (por ejemplo 1/0)*
	`> 1/0`
	`[1] Inf`

NaN	*Indeterminación (por ejemplo 0/0)*
	`> 0/0`
	`[1] NaN`

NA	*Valor desaparecido*
NULL	*Objeto nulo*

R dispone de una gama de funciones de variable real predefinidas, que por supuesto serán válidas para variables racionales, irracionales y enteras. En los párrafos siguientes se presentan las más importantes.

Funciones trigonométricas

A continuación se presenta una tabla con las funciones trigonométricas y sus inversas que incorpora R ilustradas con ejemplos.

Función	*Inversa*

sin(x) *Seno*
`> sin(pi/2)`
`[1] 1`
	asin(x) *Arco seno*
	`> asin(1)`
	`[1] 1.570796`

© Alfaomega - RC Libros

cos(x) *Coseno*
`> cos(pi)`
`[1] -1`
	acos(x) *Arco coseno*
	`> acos(-1)`
	`[1] 3.141593`

tan(x) *Tangente*
`> tan(pi/4)`
`[1] 1`
	atan(x) y atan2(x)
	`> atan(1)`
	`[1] 0.7853982`

Funciones hiperbólicas

A continuación se presenta una tabla con las funciones hiperbólicas y sus inversas que incorpora R ilustradas con ejemplos.

Función	*Inversa*

sinh(x) *Seno hiperbólico*
`> sinh(2)`
`[1] 3.62686`
	asinh(x) *Arco seno hiperbólico*
	`> asinh(3.6269)`
	`[1] 2.000011`

cosh(x) *Coseno hiperbólico*
`> cosh(3)`
`[1] 10.06766`
	acosh(x) *Arco coseno hiperbólico*
	`> acosh(10.0677)`
	`[1] 3.000004`

tanh(x) *Tangente hiperbólica*
`> tanh(1)`
`[1] 0.7615942`
	atanh(x) *Arco tangente hiperbólica*
	`> atanh(0.7616)`
	`[1] 1.000014`

Funciones exponenciales y logarítmicas

A continuación se presenta una tabla con las funciones exponenciales y logarítmicas que incorpora R ilustradas con ejemplos.

Función	*Significado*

exp(x)	*Función exponencial en base e (e^x)*
	`> exp(log(7))`
	`[1] 7`

log(x)	*Función logaritmo en base e de x*
	`> log(exp(7))`
	`[1] 7`

log10(x)	*Función logaritmo en base 10 de x*
	`> log10(1000)`
	`[1] 3`

log2(x)	*Función logaritmo en base 2 de x*
	`> log2(2^8)`
	`[1] 8`

© Alfaomega - RC Libros

sqrt(x)	*Función raíz cuadrada de x*
	`> sqrt(2^8)`
	`[1] 16`

Funciones específicas de variable numérica

R incorpora un grupo de funciones de variable numérica cuya finalidad es obtener resultados aproximados de operaciones y redondeos, trabajar con signos, etc. Entre estas funciones tenemos las siguientes:

Función	*Significado*

abs(x)	*Valor absoluto del real x*
	`> abs(-8)`
	`[1] 8`

floor(x)	*El mayor entero menor o igual que el real x*
	`> floor(-23.557)`
	`[1] -24`

ceiling(x)	*El menor entero mayor o igual que el real x*
	`> ceiling(-23.557)`
	`[1] -23`

round(x)	*El entero más próximo al real x*
	`> round(-23.557)`
	`[1] -24`

round(x,n)	*Redondea los elementos de x a n decimales*

trunc(x)	*Elimina la parte decimal del real x*
	`> trunc(-23.557)`
	`[1] -23`

sign(x)	*Signo del real x (1 si x>0 , -1 si x<0)*
	`> sign(-23.557)`
	`[1] -1`

VARIABLES SIMPLES

MATLAB no requiere ningún tipo de comando para declarar variables. Sencillamente crea la variable mediante asignación directa de su valor. Por ejemplo:

```
> V=3
> V
[1] 3
```

© Alfaomega - RC Libros

La variable *V* valdrá 3 mientras no se cambie su valor mediante una nueva asignación. Una vez declarada la variable podemos utilizarla en cálculos.

```
> V^3
[1] 27

    27

> V+5
[1] 8
```

El valor asignado a una variable es permanente, hasta que no se cambie de forma expresa o hasta que no se salga de la presente sesión de R.

Si ahora escribimos:

```
> V=3+7
> V
[1] 10
```

la variable *V* pasa a valer 10 a partir de este momento, tal y como se observa en el cálculo siguiente:

```
> V^4
[1] 10000
```

Los nombres de las variables comienzan por una letra seguida de cualquier número de letras, dígitos o subrayados. También es muy importante señalar que R es sensible a mayúsculas y minúsculas; por lo tanto, una variable con mayúsculas es distinta a la misma variable con minúsculas.

VARIABLES VECTORIALES

Para representar un vector de *n* elementos se puede definir en R una variable de la siguiente forma:

V = c(v1, v2, v3, ...,vn)

Cuando se aplican la mayoría de los comandos y funciones de R sobre una variable vectorial el resultado que se obtiene es la aplicación del comando o función sobre cada elemento del vector:

© Alfaomega - RC Libros


```
> vector1=c(1,4,9,2.25,1/4)
> vector1
[1] 1.00 4.00 9.00 2.25 0.25

> sqrt(vector1)
[1] 1.0 2.0 3.0 1.5 0.5
```

También existen formas específicas de definición de variables vectoriales, entre las que se encuentran las siguientes:

variable=a:b	*Define el vector cuyos primero y último elementos son a y b, respectivamente, y los elementos intermedios se diferencian en una unidad*
	`> V1=1:6`
	`> V1`
	`[1] 1 2 3 4 5 6`

variable=seq(a:b)	*Define el vector cuyos primero y último elementos son a y b, respectivamente, y los elementos intermedios se diferencian en una unidad (secuencia)*
	`> V1=seq(1:6)`
	`> V1`
	`[1] 1 2 3 4 5 6`

variable=seq(a,b, by=s)	*Define el vector cuyos primero y último elementos son a y b, y los elementos intermedios se diferencian en la cantidad s especificada por el incremento*
	`> V2=seq(2,8,by=2)`
	`> V2`
	`[1] 2 4 6 8`

variable=seq(along=vector)	*Define el vector de números enteros consecutivos cuyos primero y último elementos son 1 y la longitud del vector*
	`> seq(along=V1)`
	`[1] 1 2 3 4 5 6`

variable=numeric(n)	*Crea el vector con n elementos nulos*
	`> V3=numeric(10)`
	`> V3`
	`[1] 0 0 0 0 0 0 0 0 0 0`

© Alfaomega - RC Libros

variable=rep(x,n)	*Repite n veces el valor x o el vector x*
	`> V4=rep(1,6)`
	`> V4`
	`[1] 1 1 1 1 1 1`
	`> v5=rep(c(1,2,3),3)`
	`> v5`
	`[1] 1 2 3 1 2 3 1 2 3`

variable= list(objeto1, …, objeton)	*Crea una lista con los objetos especificados*
	`> V6=list(a=c(1,2),b="hola",c=3)`
	`> V6`
	`$a`
	`[1] 1 2`
	`$b`
	`[1] "hola"`
	`$c`
	`[1] 3`

Elementos de variables vectoriales

MATLAB permite la selección de elementos de variables vectoriales mediante los siguientes comandos:

x[n]	*Devuelve el enésimo elemento del vector x*
	`> X=2:8`
	`> X`
	`[1] 2 3 4 5 6 7 8`
	`> X[3]`
	`[1] 4`

x[-n]	*Devuelve todos los valores del vector x menos el enésimo*
	`> X[-3]`
	`[1] 2 3 5 6 7 8`

© Alfaomega - RC Libros

x[1:n]	*Devuelve los n primeros elementos del vector x*
	`> X[1:3]`
	`[1] 2 3 4`

x[a:b]	*Devuelve los elementos del vector x entre el a-ésimo y el b-ésimo*
	`> X[2:3]`
	`[1] 3 4`

x[-(1:n)]	*Devuelve los elementos del vector x desde el n+1 hasta el último*
	`> X[-(1:3)]`
	`[1] 5 6 7 8`

x[-(a:b)]	*Devuelve los elementos del vector x que no están entre el a-ésimo y el b-ésimo*
	`> X[-(2:3)]`
	`[1] 2 5 6 7 8`

x[c(a,b,c)]	*Devuelve los elementos del vector x situados en el a-ésimo, b-ésimo y c-ésimo lugar*
	`> X[c(3,5,6)]`
	`[1] 4 6 7`

x[x>a]	*Devuelve los elementos del vector x que son mayores que el valor a*
	`> X[X>4]`
	`[1] 5 6 7 8`

x[x>a & x<b]	*Devuelve los elementos del vector x que son mayores que el valor a y menores que el valor b*
	`> X[X>4 & X<7]`
	`[1] 5 6`

x["nombre"]	*Devuelve elemento del vector o lista x denominado "nombre"*
	`> X=list(c(1,2,3), "Jhon",C=7)`
	`> X["C"]`
	`$C`
	`[1] 7`

© Alfaomega - RC Libros

x[a]	*Devuelve el elemento a en la lista x*
	`> X=list(c(1,2,3), "Jhon", 7)`
	`> X`
	`[[1]]`
	`[1] 1 2 3`
	`[[2]]`
	`[1] "Jhon"`
	`[[3]]`
	`[1] 7`
	`> X[2]`
	`[[1]]`
	`[1] "Jhon"`

Funciones de variables vectoriales

Las funciones reales de variable simple son ampliables a variable vectorial. Para aplicar una función de variable simple a una variable vectorial, basta aplicar la función a cada elemento del vector. A continuación se presentan ejemplos relativos a funciones trigonométricas y trigonométricas inversas.

Función	*Inversa*

sin(V) *Seno*
`> v=c(0, pi/2, pi,3*pi/2,2*pi)`
`> round(sin(v))`
`[1] 0 1 0 -1 0`
	asin(V) *Arco seno*
	`> w=c(0,1/2,1)`
	`> asin(w)`
	`[1] 0.0000000 0.5235988 1.5707963`

cos(V) *Coseno*
`> round(cos(v))`
`[1] 1 0 -1 0 1`
	acos(V) *Arco coseno*
	`> acos(w)`
	`[1] 1.570796 1.047198 0.000000`

tan(V) *Tangente*
`> round(tan(v))`
`[1] 0.000000e+00 1.633124e+16 0.000000e+00 5.443746e+15 0.000000e+00`
	atan(V) y atan2(V)
	`> atan(w)`
	`[1] 0.0000000 0.4636476 0.7853982`

© Alfaomega - RC Libros

sinh(V) *Seno hiperbólico*
`> sinh(w)`
`[1] 0.0000000 0.5210953 1.1752012`
	asinh(V) *Arco seno hiperbólico*
	`> asinh(w)`
	`[1] 0.0000000 0.4812118 0.8813736`

cosh(V) *Coseno hiperbólico*
`> cosh(w)`
`[1] 1.000000 1.127626 1.543081`
	acosh(V) *Arco coseno hiperbólico*
	`> acosh(w)`
	`[1] NaN NaN 0`

tanh(V) *Tangente hiperbólica*
`> tanh(w)`
`[1] 0.0000000 0.4621172 0.7615942`
	atanh(V) *Arco tangente hiperbólica*
	`> atanh(w)`
	`[1] 0.0000000 0.5493061 Inf`

A continuación se presentan ejemplos de funciones exponenciales, logarítmicas y otras funciones de variable vectorial:

Función	*Significado*

exp(V)	*Función exponencial en base e de los elementos del vector V*
	`> exp(log(c(1,2,3)))`
	`[1] 1 2 3`

log(V)	*Función logaritmo en base e de los elementos del vector V*
	`> log(exp(c(1,2,3)))`
	`[1] 1 2 3`

log10(V)	*Función logaritmo en base 10 de los elementos del vector V*
	`> log10((c(10,100,1000)))`
	`[1] 1 2 3`

log2(V)	*Función logaritmo en base 2 de los elementos del vector V*
	`> log2(c(2,4,8))`
	`[1] 1 2 3`

sqrt(V)	*Función raíz cuadrada de los elementos del vector V*
	`> sqrt(c(2,4,8))`
	`[1] 1.414214 2.000000 2.828427`

© Alfaomega - RC Libros

floor(V)	*El mayor entero menor o igual que cada elemento del vector V*
	`> floor(sqrt(c(2,4,8)))`
	`[1] 1 2 2`

ceiling(V)	*El menor entero mayor o igual que cada elemento del vector V*
	`> ceiling(sqrt(c(2,4,8)))`
	`[1] 2 2 3`

round(V)	*El entero más próximo a cada elemento de V*
	`> round(sqrt(c(2,4,8)))`
	`[1] 1 2 3`

trunc(V)	*Elimina la parte decimal de cada elemento de V*
	`> trunc(sqrt(c(2,4,8)))`
	`[1] 1 2 2`

abs(V)	*Valor absoluto de cada elemento de V*
	`> abs(c(-1,0,1,-6,2))`
	`[1] 1 0 1 6 2`

sign(V)	*Signo de cada elemento de V (1 si V_i>0 y -1 si V_i<0)*
	`> sign(c(-1,0,1,-6,2))`
	`[1] -1 0 1 -1 1`

factorial(V)	*Factorial de cada elemento de V*
	`> factorial(c(1,2,3,4,5,6))`
	`[1] 1 2 6 24 120 720`

length(V)	*Número de elementos del vector V*
	`> length(c(1,2,3,4,5,6))`
	`[1] 6`

sum(V)	*Suma de los elementos del vector V*
	`> sum(c(3,8,9,15))`
	`[1] 35`

prod(V)	*Producto de los elementos del vector V*
	`> prod(c(3,8,9,15))`
	`[1] 3240`

Factores

Un factor es un vector de cadenas de caracteres que se utiliza para representar datos categóricos, pudiendo incluir también los diferentes niveles de la

© Alfaomega - RC Libros

variable categórica. Los factores se crean con el comando *factor*. La sintaxis siguiente crea un factor a partir de los elementos de un vector:

```
> data = c(1,2,2,3,1,2,3,3,1,2,3,3,1)
> fdata = factor(data)
> fdata
 [1] 1 2 2 3 1 2 3 3 1 2 3 3 1
Levels: 1 2 3
```

Se observa que el factor anterior tiene tres niveles. Es posible cambiar los valores de los niveles de un factor con el comando *labels*. Por ejemplo, con la sintaxis siguiente podemos pasar los niveles del factor anterior a números romanos:

```
> rdata = factor(data,labels=c("I","II","III"))
> rdata
 [1] I   II  II  III I   II  III III I   II  III III I
Levels: I II III
```

También podríamos haber cambiado los niveles de un factor con el comando *levels*, a través de la siguiente sintaxis:

```
> levels(fdata) = c('I','II','III')
> fdata
 [1] I   II  II  III I   II  III III I   II  III III I
Levels: I II III
```

El comando *levels* se puede utilizar como argumento del comando *factor*.

```
> factor(1:3,labels=c("A","B","C"))
[1] A B C
Levels: A B C
```

El comando *exclude* permite excluir niveles de un factor a medida.

```
> factor(1:5,exclude=4)
[1] 1    2    3    <NA> 5
Levels: 1 2 3 5
```

El comando *ordered* permite especificar si los valores del factor están ordenados.

```
> fert = c(10,20,20,50,10,20,10,50,20)
> fert = factor(fert,levels=c(10,20,50),ordered=TRUE)
> fert
[1] 10 20 20 50 10 20 10 50 20
Levels: 10 < 20 < 50
```

© Alfaomega - RC Libros

VARIABLES MATRICIALES

Para representar una matriz de *n* elementos *v1, v2, ..., vn* con *r* filas y *c* columnas, se utiliza en R el comando *matrix* con la siguiente sintaxis:

M = matrix(c(v1, v2, v3, ...,vn), nrow=r, ncol=c)

Con esta definición los elementos de la matriz se sitúan consecutivamente por columnas hasta completar los *n = rxc* elementos que la constituyen.

```
> M <- matrix(c(1,2,3,4,5,6),nrow = 2, ncol = 3)
> M
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

Se puede prescindir de los literales *nrow* y *ncol*, quedando la sintaxis de definición de una matriz de dimensión (*r, c*) como sigue:

M = matrix(c(v1, v2, v3, ...,vn), r, c)

```
> M = matrix(c(1,2,3,4,5,6), 2, 3)
> M
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

Si queremos crear la matriz por filas se utiliza el comando *byrow* mediante la siguiente sintaxis:

M = matrix(c(v1, v2, v3, ...,vn), nrow=r, ncol=c, byrow=TRUE)

```
> M = matrix(c(1,2,3,4,5,6),nrow = 2, ncol = 3, byrow=TRUE)
> M
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

Si queremos añadirles nombres a los índices de columnas: *Naranjas, Plátanos* y *Melón*, y en las filas: *Supermercado* y *Tienda*, se utiliza el comando *dimnames* mediante la siguiente sintaxis:

© Alfaomega - RC Libros


```
> M = matrix(c(1,2,3,4,5,6),nrow = 2, ncol = 3, byrow=TRUE,
+ dimnames=list(c("Supermercado", "Tienda"), c("Naranjas",  +
"Plátanos","Melón")))
> M
            Naranjas Plátanos Melón
Supermercado        1        2     3
Tienda              4        5     6
```

También existen comandos específicos de definición de variables matriciales, entre los que se encuentran los siguientes:

rbind(x,y,z,...)	*Define la matriz cuyas filas son los vectores x, y, z, ...*
	`> rbind(1:3,11:13,21:23)`
	`[,1] [,2] [,3]`
	`[1,] 1 2 3`
	`[2,] 11 12 13`
	`[3,] 21 22 23`

cbind(x,y,z,...)	*Define la matriz cuyas columnas son los vectores x, y, z, ...*
	`> cbind(1:3,11:13,21:23)`
	`[,1] [,2] [,3]`
	`[1,] 1 11 21`
	`[2,] 2 12 22`
	`[3,] 3 13 23`

array(x, dim=c(m,n))	*Define la matriz de dimensión (m,n) con los elementos del vector x por columnas*
	`> M = array(c(1,2,3,4,5,6),dim=c(2,3))`
	`> M`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 4 6`

Elementos de variables matriciales

MATLAB permite la selección de elementos de variables matriciales mediante los siguientes comandos:

© Alfaomega - RC Libros

A[m,n]	*Define el elemento (m,n) de la matriz A (fila m y columna n)*
	`> A=rbind(c(3,5,7,4), c(1,6,8,9), c(2,6,8,1))`
	`> A`
	`[,1] [,2] [,3] [,4]`
	`[1,] 3 5 7 4`
	`[2,] 1 6 8 9`
	`[3,] 2 6 8 1`
	`> A[2,4]`
	`[1] 9`

A[a:b,c:d]	*Define la submatriz de A formada por las filas que hay entre la a-ésima y la b-ésima y por las columnas que hay entre la c-ésima y la d-ésima*
	`> A[1:2,2:3]`
	`[,1] [,2]`
	`[1,] 5 7`
	`[2,] 6 8`

A[,c:d]	*Define la submatriz de A formada por todas las filas de A y por las columnas que hay entre la c-ésima y la d-ésima*
	`> A[,2:4]`
	`[,1] [,2] [,3]`
	`[1,] 5 7 4`
	`[2,] 6 8 9`
	`[3,] 6 8 1`

A[a:b,]	*Define la submatriz de A formada por todas las columnas de A y por las filas que hay entre la a-ésima y la b-ésima*
	`> A[2:3,]`
	`[,1] [,2] [,3] [,4]`
	`[1,] 1 6 8 9`
	`[2,] 2 6 8 1`

A[, c(a,b,c,...)]	*Define la submatriz de A formada por las columnas a-ésima, b-ésima, c-ésima, etc., de la matriz A.*
	`> A[,c(1,3,4)]`
	`[,1] [,2] [,3]`
	`[1,] 3 7 4`
	`[2,] 1 8 9`
	`[3,] 2 8 1`

A[c(a,b,c,...),]	*Define la submatriz de A formada por las filas a-ésima, b-ésima, c-ésima, etc., de la matriz A.*

© Alfaomega - RC Libros

	`> A[c(1,3),]`
	`[,1] [,2] [,3] [,4]`
	`[1,] 3 5 7 4`
	`[2,] 2 6 8 1`

A[a,]	*Define la fila a-ésima de la matriz A*
	`> A[3,]`
	`[1] 2 6 8 1`

A[,b]	*Define la columna b-ésima de la matriz A*
	`> A[,3]`
	`[1] 7 8 8`

A[,]	*Equivale a la matriz A*
	`> A[,]`
	`[,1] [,2] [,3] [,4]`
	`[1,] 3 5 7 4`
	`[2,] 1 6 8 9`
	`[3,] 2 6 8 1`

rbind(A,B,C,...)	*Define la matriz formada por las submatrices A, B, C, ... situadas por orden una debajo de otra de forma consecutiva*
	`> M1=matrix(1:4,2,2)`
	`> M1`
	`[,1] [,2]`
	`[1,] 1 3`
	`[2,] 2 4`
	`> M2=matrix(5:8,2,2)`
	`> M2`
	`[,1] [,2]`
	`[1,] 5 7`
	`[2,] 6 8`
	`> M3=matrix(9:12,2,2)`
	`> M3`
	`[,1] [,2]`
	`[1,] 9 11`
	`[2,] 10 12`
	`> rbind(M1,M2,M3)`
	`[,1] [,2]`
	`[1,] 1 3`
	`[2,] 2 4`
	`[3,] 5 7`
	`[4,] 6 8`
	`[5,] 9 11`
	`[6,] 10 12`

© Alfaomega - RC Libros

cbind(A,B,C,...)	*Define la matriz formada por las submatrices A, B, C, ... situadas por orden una al lado de otra de forma consecutiva*
	`> cbind(M1,M2,M3)`
	`  [,1] [,2] [,3] [,4] [,5] [,6]`
	`[1,]  1  3  5  7  9  11`
	`[2,]  2  4  6  8  10  12`

Funciones matriciales específicas

En MATLAB se utilizan un grupo de funciones matriciales predefinidas que facilitan el trabajo en el campo matricial. Las más importantes son las siguientes:

diag(V)	*Crea una matriz diagonal con el vector V en la diagonal*
	`> diag(c(2,0,9,8,7))`
	`  [,1] [,2] [,3] [,4] [,5]`
	`[1,]  2  0  0  0  0`
	`[2,]  0  0  0  0  0`
	`[3,]  0  0  9  0  0`
	`[4,]  0  0  0  8  0`
	`[5,]  0  0  0  0  7`

diag(A)	*Extrae la diagonal de la matriz A como vector*
	`> A=rbind(c(1,3,5), c(2,0,8), c(-1,-3,2))`
	`> A`
	`  [,1] [,2] [,3]`
	`[1,]  1  3  5`
	`[2,]  2  0  8`
	`[3,] -1 -3  2`
	`> diag(A)`
	`[1] 1 0 2`

diag(n)	*Crea la matriz identidad de orden n*
	`> diag(4)`
	`  [,1] [,2] [,3] [,4]`
	`[1,]  1  0  0  0`
	`[2,]  0  1  0  0`
	`[3,]  0  0  1  0`
	`[4,]  0  0  0  1`

matrix(a,m,n)	*Crea la matriz de orden mxn con todos los elementos de la matriz iguales al valor a* `> matrix(1,4,5)`

© Alfaomega - RC Libros

	`  [,1] [,2] [,3] [,4] [,5]`
	`[1,]  1  1  1  1  1`
	`[2,]  1  1  1  1  1`
	`[3,]  1  1  1  1  1`
	`[4,]  1  1  1  1  1`
	`> matrix(8,4,5)`
	`  [,1] [,2] [,3] [,4] [,5]`
	`[1,]  8  8  8  8  8`
	`[2,]  8  8  8  8  8`
	`[3,]  8  8  8  8  8`
	`[4,]  8  8  8  8  8`

matrix(rnorm(a),m,n)	*Crea la matriz aleatoria de a elementos normales con dimensión (m,n) siendo a=mxn*
	`> matrix(rnorm(9),3,3)`
	`  [,1]  [,2]  [,3]`
	`[1,] 0.7753502 1.009282 -0.1153786`
	`[2,] -0.9162865 1.280172 1.7335183`
	`[3,] -0.2634779 -1.861913 0.1319272`

dim(A)	*Devuelve la dimensión de la matriz A*
	`> M=matrix(c(1,2,3,4,5,6),3,2)`
	`> M`
	`  [,1] [,2]`
	`[1,]  1  4`
	`[2,]  2  5`
	`[3,]  3  6`
	`> dim(M)`
	`[1] 3 2`

colSums(A)	*Devuelve la suma de las columnas de la matriz A*
	`> colSums(M)`
	`[1] 6 15`

rowSums(A)	*Devuelve la suma de las filas de la matriz A*
	`> rowSums(M)`
	`[1] 5 7 9`

t(A)	*Devuelve la matriz traspuesta de la matriz A*
	`> t(M)`
	`  [,1] [,2] [,3]`
	`[1,]  1  2  3`
	`[2,]  4  5  6`

det(A)	*Devuelve el determinante de la matriz cuadrada A*
	`> A=rbind(c(1,3,5), c(2,0,8), c(-1,-3,2))`
	`> A`

© Alfaomega - RC Libros

	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 0 8`
	`[3,] -1 -3 2`
	`> det(A)`
	`[1] -42`

solve(A)	*Devuelve la matriz inversa de la matriz cuadrada A*
	`> solve(A)`
	`[,1] [,2] [,3]`
	`[1,] -0.5714286 0.5000000 -0.57142857`
	`[2,] 0.2857143 -0.1666667 -0.04761905`
	`[3,] 0.1428571 0.0000000 0.14285714`

solve(A,b)	*Resuelve el sistema Ax=b*
	`> solve(A,c(1,2,3))`
	`[1] -1.2857143 -0.1904762 0.5714286`

eigen(A)	*Devuelve los autovalores y los autovectores de la matriz A*
	`> eigen(A)`
	`$values`
	`[1] 2.17917+5.115792i 2.17917-5.115792i -1.35834+0.000000i`
	`$vectors`
	`[,1] [,2] [,3]`
	`[1,] 0.5333555-0.2878809i 0.5333555+0.2878809i 0.8798028+0i`
	`[2,] 0.6346542+0.0000000i 0.6346542+0.0000000i -0.4533254+0i`
	`[3,] 0.0395385+0.4778151i 0.0395385-0.4778151i -0.1429795+0i`

kappa(A)	*Número de condición de la matriz de la matriz A*
	`> kappa(A)`
	`[1] 13.61015`

Funciones de variables matriciales

Las funciones reales de variable simple y vectorial son ampliables a variable matricial. Para aplicar una función de variable simple o vectorial a una variable matricial, basta aplicar la función a cada elemento de la matriz. A continuación se presentan ejemplos relativos a funciones trigonométricas y trigonométricas inversas.

Función	*Inversa*

sin(A) *Seno*
`> A=rbind(c(1,3,5), c(2,0,8), c(-1,-3,2))`
`> A`
	asin(A) *Arco seno*
	`> asin(A)`
	`[,1] [,2] [,3]`
	`[1,] 1.570796 NaN NaN`

© Alfaomega - RC Libros

`[,1] [,2] [,3]`
`[1,] 1 3 5`
`[2,] 2 0 8`
`[3,] -1 -3 2`
`> sin(A)`
`[,1] [,2] [,3]`
`[1,] 0.8414710 0.14112 -0.9589243`
`[2,] 0.9092974 0.00000 0.9893582`
`[3,] -0.8414710 -0.14112 0.9092974`
	`[2,] NaN 0 NaN`
	`[3,] -1.570796 NaN NaN`

cos(A) *Coseno*
`> cos(A)`
`[,1] [,2] [,3]`
`[1,] 0.5403023 -0.9899925 0.2836622`
`[2,] -0.4161468 1.0000000 -0.1455000`
`[3,] 0.5403023 -0.9899925 -0.4161468`
	acos(A) *Arco coseno*
	`> acos(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.000000 NaN NaN`
	`[2,] NaN 1.570796 NaN`
	`[3,] 3.141593 NaN NaN`

tan(A) *Tangente*
`> tan(A)`
`[,1] [,2] [,3]`
`[1,] 1.557408 -0.1425465 -3.380515`
`[2,] -2.185040 0.0000000 -6.799711`
`[3,] -1.557408 0.1425465 -2.185040`
	atan(A) y atan2(A)
	`> atan(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.7853982 1.249046 1.373401`
	`[2,] 1.1071487 0.000000 1.446441`
	`[3,] -0.7853982 -1.249046 1.107149`

sinh(A) *Seno hiperbólico*
`> sinh(A)`
`[,1] [,2] [,3]`
`[1,] 1.175201 10.01787 74.20321`
`[2,] 3.626860 0.00000 1490.47883`
`[3,] -1.175201 -10.01787 3.62686`
	asinh(A) *Arco seno hiperbólico*
	`> asinh(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.8813736 1.818446 2.312438`
	`[2,] 1.4436355 0.000000 2.776472`
	`[3,] -0.8813736 -1.818446 1.443635`

cosh(A) *Coseno hiperbólico*
`> cosh(A)`
`[,1] [,2] [,3]`
`[1,] 1.543081 10.06766 74.209949`
`[2,] 3.762196 1.00000 1490.479161`
`[3,] 1.543081 10.06766 3.762196`
	acosh(A) *Arco coseno hiperbólico*
	`> acosh(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.000000 1.762747 2.292432`
	`[2,] 1.316958 NaN 2.768659`
	`[3,] NaN NaN 1.316958`

tanh(A) *Tangente hiperbólica*
`> tanh(A)`
`[,1] [,2] [,3]`
`[1,] 0.7615942 0.9950548 0.9999092`
`[2,] 0.9640276 0.0000000 0.9999998`
`[3,] -0.7615942 -0.9950548 0.9640276`
	atanh(A) *Arco tangente hiperbólica*
	`> atanh(A)`
	`[,1] [,2] [,3]`
	`[1,] Inf NaN NaN`
	`[2,] NaN 0 NaN`
	`[3,] -Inf NaN NaN`

A continuación se presentan ejemplos de funciones exponenciales, logarítmicas y otras funciones de variable vectorial:

© Alfaomega - RC Libros

Función	*Significado*

exp(A)	*Función exponencial en base e de los elementos de la matriz A*
	`> A=rbind(c(1,3,5), c(2,0,8), c(-1,-3,2))`
	`> A`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 0 8`
	`[3,] -1 -3 2`
	`> exp(A)`
	`[,1] [,2] [,3]`
	`[1,] 2.7182818 20.08553692 148.413159`
	`[2,] 7.3890561 1.00000000 2980.957987`
	`[3,] 0.3678794 0.04978707 7.389056`

log(A)	*Función logaritmo en base e de los elementos de la matriz A*
	`> log(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.0000000 1.098612 1.6094379`
	`[2,] 0.6931472 -Inf 2.0794415`
	`[3,] NaN NaN 0.6931472`

log10(A)	*Función logaritmo en base 10 de los elementos de la matriz A*
	`> log10(A)`
	`[,1] [,2] [,3]`
	`[1,] 0.00000 0.4771213 0.69897`
	`[2,] 0.30103 -Inf 0.90309`
	`[3,] NaN NaN 0.30103`

log2(A)	*Función logaritmo en base 2 de los elementos de la matriz A*
	`> log2(A)`
	`[,1] [,2] [,3]`
	`[1,] 0 1.584963 2.321928`
	`[2,] 1 -Inf 3.000000`
	`[3,] NaN NaN 1.000000`

sqrt(A)	*Función raíz cuadrada de los elementos de la matriz A* `sqrt(A)`

© Alfaomega - RC Libros

	`[,1] [,2] [,3]`
	`[1,] 1.000000 1.732051 2.236068`
	`[2,] 1.414214 0.000000 2.828427`
	`[3,] NaN NaN 1.414214`

floor(A)	*El mayor entero menor o igual que cada elemento de la matriz A*
	`> floor(sqrt(A))`
	`[,1] [,2] [,3]`
	`[1,] 1 1 2`
	`[2,] 1 0 2`
	`[3,] NaN NaN 1`

ceiling(A)	*El menor entero mayor o igual que cada elemento de la matriz A*
	`> ceiling(sqrt(A))`
	`[,1] [,2] [,3]`
	`[1,] 1 2 3`
	`[2,] 2 0 3`
	`[3,] NaN NaN 2`

round(A)	*El entero más próximo a cada elemento de V*
	`> round(sqrt(A))`
	`[,1] [,2] [,3]`
	`[1,] 1 2 2`
	`[2,] 1 0 3`
	`[3,] NaN NaN 1`

trunc(A)	*Elimina la parte decimal de cada elemento de V*
	`> trunc(sqrt(A))`
	`[,1] [,2] [,3]`
	`[1,] 1 1 2`
	`[2,] 1 0 2`
	`[3,] NaN NaN 1`

abs(A)	*Valor absoluto de cada elemento de V*
	`> A=rbind(c(1,3,5), c(2,0,8), c(-1,-3,2))`
	`> A`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 0 8`
	`[3,] -1 -3 2`
	`> abs(A)`
	`[,1] [,2] [,3]`

© Alfaomega - RC Libros

	[1,] 1 3 5 [2,] 2 0 8 [3,] 1 3 2
sign(A)	*Signo de cada elemento de V (1 si V_i>0 y -1 si V_i<0)* > sign(A) [,1] [,2] [,3] [1,] 1 1 1 [2,] 1 0 1 [3,] -1 -1 1

OPERADORES

R es un lenguaje que incorpora operadores aritméticos, lógicos y relacionales de igual forma que cualquier otro lenguaje. A continuación se enumeran los tipos de operadores citados en el ámbito del lenguaje de R.

Operadores aritméticos

R incorpora los operadores aritméticos habituales (suma, resta, multiplicación y división) para trabajar con números. Pero a la vez amplía el sentido de estos operadores para el trabajo con escalares, vectores y matrices, tal y como se muestra en la tabla siguiente:

Operador	*Función que desempeña*
+	*Suma de escalares, vectores o matrices* > A=c(1,3,-2,6);B=c(4,-5,8,2); c=3; > V1=A+c > V1 [1] 4 6 1 9 > V2=A+B > V2 [1] 5 -2 6 8
-	*Resta de escalares, vectores o matrices* > V3=A-c > V3 [1] -2 0 -5 3 > V4=A-B > V4 [1] -3 8 -10 4
*	*Producto de escalares o de matrices (elemento a elemento) o de escalares por vectores o matrices*

© Alfaomega - RC Libros

	> V5=A*c > V5 [1] 3 9 -6 18
^ ó **	*Potencia escalar de un escalar, de un vector o de una matriz (en el caso del vector o matriz, se eleva cada elemento al escalar)* > V5^2 [1] 9 81 36 324 > V5**2 [1] 9 81 36 324 > c^3 [1] 27 > c**3 [1] 27
%*%	*Producto de matrices (producto matricial)* > A=matrix(rnorm(4),2,2) > A [,1] [,2] [1,] 1.082627 -1.4716030 [2,] -2.193202 0.7243937 > B=matrix(rnorm(4),2,2) > B [,1] [,2] [1,] 0.08075123 -1.689854 [2,] -0.06904861 1.087078 > A%*%B [,1] [,2] [1,] 0.1890356 -3.429229 [2,] -0.2271221 4.493664

Operadores lógicos

R también incorpora los operadores lógicos habituales utilizando la notación más habitual para los mismos. Habitualmente los resultados de los operadores lógicos suelen ser 1 si cierto y 0 si es falso. En la tabla siguiente se muestran estos operadores.

!=	*Negación lógica (NOT) o complementario de A* > 2!=1 [1] TRUE
A & B	*Conjunción lógica (AND) o intersección de A y B* > (2>3) & (5>1) [1] FALSE

© Alfaomega - RC Libros

A | B	*Disyunción lógica (OR) o unión de A y B*
	`> (2>3)||(5>1)`
	`[1] TRUE`

xor(A,B)	*OR exclusivo (XOR) o diferencia simétrica de A y B (vale TRUE si A o B, pero no ambos, son TRU)*
	`> xor((2>3),(5>1))`
	`[1] TRUE`

any(A,B,C,...)	*Vale TRUE si al menos una componente A, B, C,.. es TRUE*
	`> any((2>3),(5>1),(4>5))`
	`[1] TRUE`

all(A,B,C,...)	*Vale TRUE si todas las componentes A, B, C,.. son TRUE*
	`> all((2>3),(5>1),(4>5))`
	`[1] FALSE`

Operadores relacionales

Asimismo R se ocupa de las operaciones relacionales que ejecutan comparaciones elemento a elemento entre dos matrices y devuelven una matriz del mismo tamaño cuyos elementos son ceros si la correspondiente relación es cierta, o unos si la correspondiente relación es falsa. Los operadores relacionales también permiten comparar escalares con vectores o matrices, en cuyo caso se compara el escalar con todos los elementos de la matriz. La tabla siguiente muestra los operadores relacionales de R.

<	*Menor (para complejos solo afecta a partes reales)*
	`> 3<5`
	`[1] TRUE`

<=	*Menor o igual (para complejos solo afecta a partes reales)*
	`> 4>=6`
	`[1] FALSE`

>	*Mayor (para complejos solo afecta a partes reales)*
	`> X=matrix(3,4,2)`
	`> X`
	`[,1] [,2]`
	`[1,] 3 3`
	`[2,] 3 3`
	`[3,] 3 3`
	`[4,] 3 3`

© Alfaomega - RC Libros

	`> Y=matrix(c(1,2,3,4,5,6,7,8),4,2)`
	`> Y`
	`[,1] [,2]`
	`[1,] 1 5`
	`[2,] 2 6`
	`[3,] 3 7`
	`[4,] 4 8`
	`> X>Y`
	`[,1] [,2]`
	`[1,] TRUE FALSE`
	`[2,] TRUE FALSE`
	`[3,] FALSE FALSE`
	`[4,] FALSE FALSE`
	`> Y>3`
	`[,1] [,2]`
	`[1,] FALSE TRUE`
	`[2,] FALSE TRUE`
	`[3,] FALSE TRUE`
	`[4,] TRUE TRUE`

>=	*Mayor o igual (para complejos solo afecta a partes reales)*
	`> X>=Y`
	`[,1] [,2]`
	`[1,] TRUE FALSE`
	`[2,] TRUE FALSE`
	`[3,] TRUE FALSE`
	`[4,] FALSE FALSE`

x == y	*Igualdad (afecta a los números complejos)*
	`> X==Y`
	`[,1] [,2]`
	`[1,] FALSE FALSE`
	`[2,] FALSE FALSE`
	`[3,] TRUE FALSE`
	`[4,] FALSE FALSE`

x != y	*Desigualdad (afecta a los números complejos)*
	`> X!=Y`
	`[,1] [,2]`
	`[1,] TRUE TRUE`
	`[2,] TRUE TRUE`
	`[3,] FALSE TRUE`
	`[4,] TRUE TRUE`

© Alfaomega - RC Libros

FUNCIONES DE INFORMACIÓN

R dispone de varias funciones cuya finalidad es informar sobre la naturaleza de su argumento. Por ejemplo, la función *is.numeric(x)* devolverá TRUE si *x* es un valor numérico, y devolverá FALSE en caso contrario.

```
> v=4
> v
[1] 4
> is.numeric(v)
[1] TRUE
```

La función *is.matrix(x)* devolverá TRUE si *x* es una matriz y devolverá FALSE en caso contrario.

```
> M = matrix(c(1,2,3,4,5,6), 2, 3)
> M
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> is.matrix(M)
[1] TRUE
```

Para obtener el listado de funciones de información que contiene R basta teclear *methods(is)*.

```
> methods(is)
 [1] is.array                 is.atomic                is.call
 [4] is.character             is.complex               is.data.frame
 [7] is.double                is.element               is.empty.model
[10] is.environment           is.expression            is.factor
[13] is.finite                is.function              is.infinite
[16] is.integer               is.language              is.leaf
[19] is.list                  is.loaded                is.logical
[22] is.matrix                is.mts                   is.na
[25] is.na.data.frame         is.na.numeric_version    is.na.POSIXlt
[28] is.na<-                  is.na<-.default          is.na<-.factor
[31] is.name                  is.nan                   is.null
[34] is.numeric               is.numeric.Date          is.numeric.difftime
[37] is.numeric.POSIXt        is.numeric_version       is.object
[40] is.ordered               is.package_version       is.pairlist
[43] is.primitive             is.qr                    is.R
[46] is.raster                is.raw                   is.recursive
[49] is.relistable            is.single                is.stepfun
[52] is.symbol                is.table                 is.ts
[55] is.tskernel              is.unsorted              is.vector
```


© Alfaomega - RC Libros

FUNCIONES DE CONVERSIÓN

Las funciones de conversión transforman una variable de un determinado tipo a otro tipo especificado. Por ejemplo, si queremos transformar la variable numérica *v*=4 en variable carácter utilizaremos la función *as.character(v)*.

```
> v=4
> v
[1] 4
> as.character(v)
[1] "4"
```

Para obtener el listado de funciones de conversión que contiene R basta teclear *methods(as)*.

```
> methods(as)
 [1] as.array                      as.array.default
 [3] as.call                       as.character
 [5] as.character.condition        as.character.Date
 [7] as.character.default          as.character.error
 [9] as.character.factor           as.character.hexmode
[11] as.character.numeric_version  as.character.octmode
[13] as.character.POSIXt           as.character.srcref
[15] as.complex                    as.data.frame
[17] as.data.frame.array           as.data.frame.AsIs
[19] as.data.frame.character       as.data.frame.complex
[21] as.data.frame.data.frame      as.data.frame.Date
[23] as.data.frame.default         as.data.frame.difftime
[25] as.data.frame.factor          as.data.frame.integer
[27] as.data.frame.list            as.data.frame.logical
[29] as.data.frame.matrix          as.data.frame.model.matrix
[31] as.data.frame.numeric         as.data.frame.numeric_version
[33] as.data.frame.ordered         as.data.frame.POSIXct
[35] as.data.frame.POSIXlt         as.data.frame.raw
[37] as.data.frame.table           as.data.frame.ts
[39] as.data.frame.vector          as.Date
[41] as.Date.character             as.Date.date
[43] as.Date.dates                 as.Date.default
[45] as.Date.factor                as.Date.numeric
[47] as.Date.POSIXct               as.Date.POSIXlt
[49] as.dendrogram                 as.difftime
[51] as.dist                       as.double
[53] as.double.difftime            as.double.POSIXlt
[55] as.environment                as.expression
[57] as.expression.default         as.factor
[59] as.formula                    as.function
[61] as.function.default           as.graphicsAnnot
```

© Alfaomega - RC Libros

```
[63] as.hclust                    as.hexmode
[65] as.integer                   as.list
[67] as.list.data.frame           as.list.Date
[69] as.list.default              as.list.environment
[71] as.list.factor               as.list.function
[73] as.list.numeric_version      as.list.POSIXct
[75] as.logical                   as.logical.factor
[77] as.matrix                    as.matrix.data.frame
[79] as.matrix.default            as.matrix.noquote
[81] as.matrix.POSIXlt            as.name
[83] as.null                      as.null.default
[85] as.numeric                   as.numeric_version
[87] as.octmode                   as.ordered
[89] as.package_version           as.pairlist
[91] as.person                    as.personList
[93] as.POSIXct                   as.POSIXct.date
[95] as.POSIXct.Date              as.POSIXct.dates
[97] as.POSIXct.default           as.POSIXct.numeric
[99] as.POSIXct.POSIXlt           as.POSIXlt
[101] as.POSIXlt.character        as.POSIXlt.date
[103] as.POSIXlt.Date             as.POSIXlt.dates
[105] as.POSIXlt.default          as.POSIXlt.factor
[107] as.POSIXlt.numeric          as.POSIXlt.POSIXct
[109] as.qr                       as.raster
[111] as.raw                      as.relistable
[113] as.roman                    as.single
[115] as.single.default           as.stepfun
[117] as.symbol                   as.table
[119] as.table.default            as.ts
[121] as.vector                   as.vector.factor
```

VARIABLES Y FUNCIONES DE CADENA

Una variable carácter (cadena) es sencillamente una cadena de caracteres incluidos entre comillas simples que R trata de forma vectorial. La sintaxis general para la declaración de variables carácter en R es la siguiente:

c = "cadena de caracteres"

```
> c="cadena de caracteres"
> c
[1] "cadena de caracteres"
```

Entre las funciones de R que manejan variables carácter tenemos las siguientes:

© Alfaomega - RC Libros

nchar(cadena_caracteres)	*Devuelve el número de caracteres en la cadena*
	`> c`
	`[1] "cadena de carcateres"`
	`> c="cadena de caracteres"`
	`> nchar(c)`
	`[1] 20`

substring(x, n)	*Devuelve la subcadena de la cadena x que comienza en el carácter de lugar n y finaliza con la cadena*
	`> substring(c,10)`
	`[1] " caracteres"`

substring(x, n, m)	*Devuelve la subcadena de la cadena x que comienza en el carácter de lugar n y finaliza en el carácter de lugar m*
	`> substring(c,5,15)`
	`[1] "na de carac"`

strsplit(x, "cadena")	*Divide la cadena x en una lista de subcadenas separadas delimitadas por la cadena "cadena"*
	`> strsplit(c, "e")`
	`[[1]]`
	`[1] "cad" "na d" " caract" "r" "s"`

grep("patrón", "cadena")	*Comprueba si el patrón especificado se encuentra contenido en la cadena, en cuyo caso devuelve 1*
	`> grep("c",c)`
	`[1] 1`
	`> grep("b",c)`
	`integer(0)`

regexpr("patrón", "cadena")
o
regexec("patrón", "cadena")
	Devuelve la posición en la cadena de la primera ocurrencia del patrón dado dentro de la cadena
	`> regexpr("a", c)`
	`[1] 2`

© Alfaomega - RC Libros

	`attr(,"match.length")`
	`[1] 1`
	`attr(,"useBytes")`
	`[1] TRUE`
	
	`> regexpr("d",c)`
	`[1] 3`
	`attr(,"match.length")`
	`[1] 1`
	`attr(,"useBytes")`
	`[1] TRUE`
	
	`> regexec("de",c)`
	`[[1]]`
	`[1] 3`
	`attr(,"match.length")`
	`[1] 2`
	
	`> regexec("ca",c)`
	`[[1]]`
	`[1] 1`
	`attr(,"match.length")`
	`[1] 2`

gregexpr("patrón", "cadena")	*Devuelve todas las posiciones en la cadena de las ocurrencias del patrón dado dentro de la cadena*
	
	`> gregexpr("a", c)`
	`[[1]]`
	`[1] 2 6 12 14`
	`attr(,"match.length")`
	`[1] 1 1 1 1`
	`attr(,"useBytes")`
	`[1] TRUE`

sub("patrón", "valorreemplazo", "cadena")	*Reemplaza en la cadena el patrón especificado por el valor de reemplazo para la primera ocurrencia*
	
	`> sub("a","x",c)`
	`[1] "cxdena de caracteres"`

gsub("patrón", "valorreemplazo", "cadena")	*Reemplaza en la cadena el patrón especificado por el valor de reemplazo para todas las ocurrencias*
	
	`> gsub("a","x",c)`
	`[1] "cxdenx de cxrxcteres"`

© Alfaomega - RC Libros

tolower("cadena")	*Convierte a minúsculas todos los caracteres de la cadena*
	
	`> tolower("CADENA")`
	`[1] "cadena"`

toupper("cadena")	*Convierte a mayúsculas todos los caracteres de la cadena*
	
	`> toupper("cadena")`
	`[1] "CADENA"`

match("cadena",table)	*Devuelve un vector con la posición de la primera coincidencia exacta de la cadena completa en la tabla*
	
	`> match("mean", c("median", "mean", "mode", "mean"))`
	`[1] 2`

pmatch("cadena",table)	*Devuelve un vector con la posición de la primera coincidencia de parte de la cadena en la tabla*
	
	`> pmatch("med", c("mean", "median", "mode"))`
	`[1] 2`

paste("cadena", array, sep= "caracteres de separación")	*Sitúa la cadena como elementos del array dado separados por el carácter de separación especificado*
	
	`> paste("A", 1:6, sep = "")`
	`[1] "A1" "A2" "A3" "A4" "A5" "A6"`

FUNCIONES DE SELECCIÓN Y MANIPULACIÓN

A continuación se presentan un grupo de funciones que sirven para seleccionar y manipular los elementos de un vector o array.

© Alfaomega - RC Libros

which.max(v)	*Devuelve el índice del mayor elemento de V*
	`> V=c(10,25,3,8,49,12,3,NA)`
	`> V`
	`[1] 10 25  3  8 49 12  3 NA`
	`> which.max(V)`
	`[1] 5`

which.min(v)	*Devuelve el índice del menor elemento de V*
	`> which.min(V)`
	`[1] 3`

which(v==a)	*Devuelve los índices de los elementos de V que coinciden con a*
	`> which(V==3)`
	`[1] 3 7`

rev(v)	*Invierte el orden de los elementos de V*
	`> rev(V)`
	`[1] NA  3 12 49  8  3 25 10`

sort(v)	*Ordena los elementos de V en orden creciente*
	`> sort(V)`
	`[1]  3  3  8 10 12 25 49`

rev(sort(v))	*Ordena los elementos de V en orden decreciente*
	`> rev(sort(V))`
	`[1] 49 25 12 10  8  3  3`

unique(v)	*Devuelve el vector sin elementos repetidos*
	`> unique(V)`
	`[1] 10 25  3  8 49 12 NA`

subset(v, condición)	*Devuelve los elementos de V que cumplen la condición*
	`> subset(V, V<10)`
	`[1] 3 8 3`

© Alfaomega - RC Libros

sample(v,n)	*Devuelve una muestra aleatoria simple sin reemplazamiento de n elementos del vector V*
	`> sample(V,4)`
	`[1] 49  3  8 10`

sample(v,n, replace=TRUE)	*Devuelve una muestra aleatoria simple con reemplazamineto de n elementos del vector V*
	`> sample(V,4,replace=TRUE)`
	`[1] 25 49  8  3`

table(v)	*Devuelve una tabla con las frecuencias de los diferentes valores de V*
	`> table(V)`
	`V`
	`  3  8 10 12 25 49`
	`  2  1  1  1  1  1`

na.omit(v)	*Suprime las observaciones con valores desaparecidos de V*
	`> na.omit(V)`
	`[1] 10 25  3  8 49 12  3`
	`attr(,"na.action")`
	`[1] 8`
	`attr(,"class")`
	`[1] "omit"`

na.fail(v)	*Devuelve un mensaje de error si V contiene algún elemento desaparecido*
	`> na.fail(V)`
	`Error in na.fail.default(V) :`
	`missing values in object`

Las funciones de selección y manipulación anteriores también pueden aplicarse a matrices y arrays. A continuación se presentan ejemplos:

© Alfaomega - RC Libros

```
> A=matrix(c(10,25,3,40,5,6,7,8,9),3,3)
> A
     [,1] [,2] [,3]
[1,]   10   40    7
[2,]   25    5    8
[3,]    3    6    9

> which.max(A)
[1] 4

> which.min(A)
[1] 3

> rev(A)
[1]  9  8  7  6  5 40  3 25 10

> sort(A)
[1]  3  5  6  7  8  9 10 25 40

> rev(sort(A))
[1] 40 25 10  9  8  7  6  5  3

> which(A==5)
[1] 5

> which(A==25)
[1] 2

> na.omit(A)
     [,1] [,2] [,3]
[1,]   10   40    7
[2,]   25    5    8
[3,]    3    6    9

> na.fail(A)
     [,1] [,2] [,3]
[1,]   10   40    7
[2,]   25    5    8
[3,]    3    6    9

> table(A)
A
 3  5  6  7  8  9 10 25 40
 1  1  1  1  1  1  1  1  1
```


© Alfaomega - RC Libros


```
> subset(A<10)
      [,1]  [,2] [,3]
[1,] FALSE FALSE TRUE
[2,] FALSE  TRUE TRUE
[3,]  TRUE  TRUE TRUE

> sample(A,4)
[1] 5 8 9 7
```

Ejercicio 2-1. Calcular el valor que ofrece R por defecto para el factorial de 23. Calcularlo también con 15, con 20 y con 22 cifras significativas.

```
> factorial(23)
[1] 2.585202e+22

> options(digits=15)
> factorial(23)
[1] 2.5852016738885e+22

> options(digits=20)
> factorial(23)
[1] 25852016738884978212844

> options(digits=22)
> factorial(23)
[1] 25852016738884978212844
```

*Ejercicio 2-2. Dadas las variables vectoriales a=[π,2π,3π,4π,5π] y b=[e,2e,3e,4e,5e] calcular c= Sen(a)+b, d=Cosh(a), e=Ln(b), f=c*d, g=c/d, h=d^2. Los productos, cocientes y potencias se consideran elemento a elemento.*

```
> a=c(pi,2*pi,3*pi,4*pi,5*pi)
> b=c(exp(1),2*exp(1),3*exp(1),4*exp(1),5*exp(1))
> a
[1]  3.141593  6.283185  9.424778 12.566371 15.707963
> b
[1]  2.718282  5.436564  8.154845 10.873127 13.591409

> c=sin(a)+b
> c
[1]  2.718282  5.436564  8.154845 10.873127 13.591409

> d=cosh(a)
> d
```

© Alfaomega - RC Libros

```
[1] 1.159195e+01 2.677468e+02 6.195824e+03 1.433757e+05
3.317812e+06
> e=log(b)
> e
[1] 1.000000 1.693147 2.098612 2.386294 2.609438

> f=c*d
> f
[1] 3.151020e+01 1.455622e+03 5.052599e+04 1.558942e+06
4.509374e+07

> g=c/d
> g
[1] 2.344973e-01 2.030487e-02 1.316184e-03 7.583663e-05
4.096498e-06

> h=d^2
> h
[1] 1.343734e+02 7.168833e+04 3.838823e+07 2.055658e+10
1.100788e+13
```

Ejercicio 2-3. Dado el vector de los 10 primeros números naturales, hallar:

Su sexto elemento.
Sus elementos situados entre el cuarto y el séptimo ambos inclusive.
Sus elementos situados entre el segundo y el noveno ambos inclusive de tres en tres.
Los elementos del apartado anterior, pero de mayor a menor.

```
> x=(1:10)
> x
 [1]  1  2  3  4  5  6  7  8  9 10

> x[6]
[1] 6
```

Hemos obtenido el sexto elemento del vector x.

```
> x[4:7]
[1] 4 5 6 7
```

Hemos obtenido los elementos del vector x situados entre el cuarto y el séptimo, ambos inclusive.

```
> seq(2,9, by=3)
[1] 2 5 8
```


© Alfaomega - RC Libros

Hemos obtenido los elementos del vector x situados entre el segundo y el noveno, ambos inclusive, pero separados de tres en tres unidades.

```
> seq(9,2, by=-3)
[1] 9 6 3
```

Hemos obtenido los elementos del vector *x* situados entre el noveno y el segundo, ambos inclusive, pero separados de tres en tres unidades y empezando por el noveno.

Ejercicio 2-4. Dada la matriz 2x3 cuyas filas son los seis primeros números impares consecutivos, se pide:

Anular su elemento (2,3), transponerla y adosarle la matriz identidad de orden 3 a su derecha.

Si llamamos C a la matriz obtenida anteriormente, construir una matriz D extrayendo las columnas impares de la matriz C, una matriz E formada por la intersección de las dos primeras filas de C y sus columnas tercera y quinta, y una matriz F formada por la intersección de las dos primeras filas y las tres últimas columnas de la matriz C.

Construir la matriz diagonal G tal que los elementos de su diagonal principal son los mismos que los de la diagonal principal de D.

Construimos la matriz H, formada por la intersección de la primera y tercera filas de C y su segunda, tercera y quinta columnas.

Consideramos en primer lugar la matriz *2x3* cuyas filas son los 6 primeros impares consecutivos:

```
> A=matrix(c(1,3,5,7,9,11), 2,3,byrow=TRUE)
> A
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    7    9   11
```

Ahora vamos a anular el elemento *(2,3)*, o sea, su último elemento:

```
> A[2,3]=0
> A
     [,1] [,2] [,3]
```

© Alfaomega - RC Libros

```
[1,]    1    3    5
[2,]    7    9    0
```

A continuación, consideramos la matriz *B* transpuesta de *A*:

```
> B=t(A)
> B
     [,1] [,2]
[1,]    1    7
[2,]    3    9
[3,]    5    0
```

Ahora construimos una matriz *C*, formada por la matriz *B* y la matriz identidad de orden 3 adosada a su derecha:

```
> C=cbind(B,diag(3))
> C
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    7    1    0    0
[2,]    3    9    0    1    0
[3,]    5    0    0    0    1
```

Vamos a construir una matriz *D* extrayendo las columnas impares de la matriz *C*, una matriz *E* formada por la intersección de las dos primeras filas de *C* y sus columnas tercera y quinta, y una matriz *F* formada por la intersección de las dos primeras filas y las tres últimas columnas de la matriz *C*:

```
> D=C[,c(1,3,5)]
> D
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    3    0    0
[3,]    5    0    1

> E=C[c(1,2),c(3,5)]
> E
     [,1] [,2]
[1,]    1    0
[2,]    0    0

> F=C[c(1,2),c(3,4,5)]
> F
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
```

Ahora construimos la matriz diagonal *G* tal que los elementos de su diagonal principal son los mismos que los de la diagonal principal de *D*:

© Alfaomega - RC Libros


```
> G=diag(diag(D))
> G
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    0    0
[3,]    0    0    1
```

A continuación, construimos la matriz *H*, formada por la intersección de la primera y tercera filas de *C* y su segunda, tercera y quinta columnas:

```
> H=C[c(1,3),c(2,3,5)]
> H
     [,1] [,2] [,3]
[1,]    7    1    0
[2,]    0    0    1
```

Ejercicio 2-5. Construir una matriz I formada por la matriz identidad de orden 5x4 y las matrices nula y unidad del mismo orden adosadas a su derecha. A continuación, extraer la primera fila de I y, por último, formar la matriz J con las filas impares y las columnas pares de Y y calculamos su orden (tamaño).

Adicionalmente, construir una matriz aleatoria K de orden 3x4 e invertir primero el orden de sus filas, después el orden de sus columnas y después el orden de sus filas y columnas a la vez.

Ahora construimos una matriz *I* formada por la matriz identidad de orden *5x4* y las matrices nula y unidad del mismo orden adosadas a su derecha. A continuación, extraemos la primera fila de *I* y, por último, formamos la matriz *J* con las filas impares y las columnas pares de *Y* y calculamos su orden (tamaño).

```
> I=cbind(diag(5)[,c(1,2,3,4)],matrix(0,5,4),matrix(1,5,4))
> I
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,]    1    0    0    0    0    0    0    0    1    1    1    1
[2,]    0    1    0    0    0    0    0    0    1    1    1    1
[3,]    0    0    1    0    0    0    0    0    1    1    1    1
[4,]    0    0    0    1    0    0    0    0    1    1    1    1
[5,]    0    0    0    0    0    0    0    0    1    1    1    1

> I[1,]
 [1] 1 0 0 0 0 0 0 0 1 1 1 1

> J=I[c(1,3,5),c(2,4,6,8,10,12)]
> J
     [,1] [,2] [,3] [,4] [,5] [,6]
```

© Alfaomega - RC Libros

```
[1,]    0    0    0    0    1    1
[2,]    0    0    0    0    1    1
[3,]    0    0    0    0    1    1

> dim(J)
[1] 3 6
```

A continuación, construimos una matriz aleatoria *K* de orden *3x4* e invertimos primero el orden de sus filas, después el orden de sus columnas y después el orden de sus filas y columnas a la vez. Por último, hallamos la matriz *L* de orden *4x3* cuyas columnas resultan de tomar los elementos de las columnas de *K* consecutivamente.

```
> K=matrix(rnorm(12),3,4)
> K
           [,1]        [,2]       [,3]       [,4]
[1,] 0.0367322 -0.07320238 -0.5295449 -0.9802087
[2,] 1.0106856 -1.67161887  1.0675619 -2.0160255
[3,] 0.1834108 -0.15990453  0.6908918 -1.4977978

> K[c(3,2,1),]
           [,1]        [,2]       [,3]       [,4]
[1,] 0.1834108 -0.15990453  0.6908918 -1.4977978
[2,] 1.0106856 -1.67161887  1.0675619 -2.0160255
[3,] 0.0367322 -0.07320238 -0.5295449 -0.9802087

> K[,c(4,3,2,1)]
           [,1]       [,2]        [,3]       [,4]
[1,] -0.9802087 -0.5295449 -0.07320238 0.0367322
[2,] -2.0160255  1.0675619 -1.67161887 1.0106856
[3,] -1.4977978  0.6908918 -0.15990453 0.1834108

> K[c(3,2,1),c(4,3,2,1)]
           [,1]       [,2]        [,3]       [,4]
[1,] -1.4977978  0.6908918 -0.15990453 0.1834108
[2,] -2.0160255  1.0675619 -1.67161887 1.0106856
[3,] -0.9802087 -0.5295449 -0.07320238 0.0367322
```

Ejercicio 2.6. Dada la matriz cuadrada de orden 3, cuyas filas son los 9 primeros números naturales, obtener su inversa, su transpuesta y su diagonal. Obtener la suma de los elementos de la primera fila y la suma de los elementos de la diagonal. Extraer la submatriz cuya diagonal son los elementos a_{11} y a_{22} y extraer también la submatriz cuyos elementos de la diagonal son a_{11} y a_{33}. Finalmente, hallar los autovalores y autovectores de la matriz inicial.

© Alfaomega - RC Libros


```
> M=matrix(c(1,2,3,4,5,6,7,8,9), 3,3,byrow=TRUE)
> M
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
> A=solve(M)
Error in solve.default(M) :
  system is computationally singular: reciprocal condition
number = 1.54198e-18
```

La matriz no tiene inversa porque su determinante es cero.

```
> round(det(M))
[1] 0
```

Los siguientes apartados del ejercicio se resuelven a continuación.

```
> B=t(M)
> B
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> V=diag(M)
> V
[1] 1 5 9

> s=M[1,1]+M[1,2]+M[1,3]
> s
[1] 6
> sd=M[1,1]+M[2,2]+M[3,3]

> sd
[1] 15
> SM=M[c(1,2),c(1,2)]
> SM
     [,1] [,2]
[1,]    1    2
[2,]    4    5

> SM1=M[c(1,3),c(1,3)]
> SM1
     [,1] [,2]
```

© Alfaomega - RC Libros

```
[1,]    1    3
[2,]    7    9

> eigen(M)
$values
[1]  1.611684e+01 -1.116844e+00 -1.303678e-15

$vectors
           [,1]        [,2]       [,3]
[1,] -0.2319707 -0.78583024  0.4082483
[2,] -0.5253221 -0.08675134 -0.8164966
[3,] -0.8186735  0.61232756  0.4082483
```

Ejercicio 2-7. Dada la matriz cuadrada de orden 3, cuyas filas son los 9 primeros números naturales (sin el cero), identificar sus valores menores que 5.

Si consideramos ahora el vector cuyos elementos son los 9 primeros números naturales (sin el cero) y el vector con los mismos elementos de mayor a menor, identificar los valores resultantes de restar a los elementos del segundo vector la cantidad 1 si el correspondiente elemento del primer vector es mayor que 2, o la cantidad 0 si es menor o igual que 2.

```
> X=5*diag(3)
> X
     [,1] [,2] [,3]
[1,]    5    0    0
[2,]    0    5    0
[3,]    0    0    5

> X>=matrix(c(1,2,3,4,5,6,7,8,9),3,3,byrow=TRUE)
      [,1]  [,2]  [,3]
[1,]  TRUE FALSE FALSE
[2,] FALSE  TRUE FALSE
[3,] FALSE FALSE FALSE
```

Los elementos de la matriz *X* que son mayores o iguales que los de la matriz [1 2 3; 4 5 6; 7 8 9] se corresponden con TRUE en la matriz respuesta. El resto de los elementos se corresponden con FALSE (el resultado de la operación sería el mismo si comparamos el escalar 5 con la matriz [1 2 3; 4 5 6; 7 8 9]).

© Alfaomega - RC Libros


```
> X=5
> X>=matrix(c(1,2,3,4,5,6,7,8,9),3,3,byrow=TRUE)
      [,1]  [,2]  [,3]
[1,]  TRUE  TRUE  TRUE
[2,]  TRUE  TRUE FALSE
[3,] FALSE FALSE FALSE
```

A continuación, resolvemos la segunda parte del ejercicio:

```
> A=1:9
> A
[1] 1 2 3 4 5 6 7 8 9

> B=10-A
> B
[1] 9 8 7 6 5 4 3 2 1

> Y=A>4
> Y
[1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
> Z=B-(A>2)
> Z
[1] 9 8 6 5 4 3 2 1 0
```

Los valores de *Y* iguales a TRUE se corresponden con los elementos de A mayores que 4. Los valores de *Z* resultan de restar a los correspondientes elementos de *B* la cantidad 1 si el correspondiente elemento de *A* es mayor que 2, o la cantidad 0 si el correspondiente elemento de *A* es menor o igual que 2.

Ejercicio 2-8. Si consideramos ahora el vector A cuyos elementos son los 9 primeros números naturales (sin el cero), identificar sus valores mayores que 2 y menores que 6.

Si llamamos P al vector de la misma dimensión que A que asocia un 1 a los resultados identificados en el apartado anterior y un cero al resto, identificar cuándo A o P, pero no ambos, valen 1.

```
> A=1:9;P=(A>2)&(A<6)
> A
[1] 1 2 3 4 5 6 7 8 9
> P
[1] FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
```

© Alfaomega - RC Libros

Devuelve TRUE cuando *A* es mayor que 2 y menor que 6, y devuelve FALSE en caso contrario.

Ahora resolvemos el segundo apartado del problema.

```
> Q=(A>=1)&(A<6)
> Q
[1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE

> xor(A,Q)
[1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

Los valores TRUE del último resultado son la solución pedida.

© Alfaomega - RC Libros

CAPÍTULO 3

LENGUAJE DE R. ESTRUCTURAS VECTORIALES Y MATRICIALES PARA ANÁLISIS DE DATOS

INTRODUCCIÓN

En este capítulo se abordan conceptos tan esenciales dentro de la programación en R como son las secuencias, los arrays, las listas y los marcos de datos, que permiten un trabajo fluido en análisis vectorial y matricial y en todo tipo de trabajo con variables. Los elementos de los vectores, las matrices y otras estructuras superiores tienen asignados índices, a partir de los cuales se manejan convenientemente y de forma ordenada sus elementos.

SECUENCIAS

Una secuencia es sencillamente una relación ordenada de objetos. Los

© Alfaomega - RC Libros

a:b	*Devuelve una secuencia con los enteros entre a y b separados de unidad en unidad.*
	`> S1=3:7`
	`> S1`
	`[1] 3 4 5 6 7`

seq(a,b,by=s)	*Devuelve la secuencia de números entre a y b separados de s en s unidades*
	`> S2=seq(1,9,by=2)`
	`> S2`
	`[1] 1 3 5 7 9`

seq(a,b,s)	*Devuelve la secuencia de números entre a y b separados de s en s unidades*
	`> S3=seq(1,9,2)`
	`> S3`
	`[1] 1 3 5 7 9`

seq(along=vector)	*Devuelve la secuencia de números enteros consecutivos que comienza en 1 y finaliza en la longitud del vector*
	`> S4=seq(along=c(25,30,35,40,45))`
	`> S4`
	`[1] 1 2 3 4 5`

rep(a,n)	*Crea la secuencia que repite el elemento a n veces*
	`> S5=rep(2,5)`
	`> S5`
	`[1] 2 2 2 2 2`

rep(vector,n))	*Crea la secuencia que repite el vector n veces*
	`> S6=rep(c(1,2,3),3)`
	`> S6`
	`[1] 1 2 3 1 2 3 1 2 3`

© Alfaomega - RC Libros

seq(along=matriz)	*Devuelve la secuencia de números enteros consecutivos que comienza en 1 y finaliza en la mxn, siendo m y n las dimensiones de la matriz*
	`>A=rbind(c(4,3,5),c(3,2,1),c(8,6,7))`
	`> A`
	`[,1] [,2] [,3]`
	`[1,] 4 3 5`
	`[2,] 3 2 1`
	`[3,] 8 6 7`
	`> S7=seq(along=A)`
	`> S7`
	`[1] 1 2 3 4 5 6 7 8 9`

rep(matriz,n))	*Crea la secuencia que repite los elementos de la matriz por columnas n veces*
	`> S8=rep(A,2)`
	`> S8`
	`[1] 4 3 8 3 2 6 5 1 7 4 3 8 3 2 6 5 1 7`

seq(length=n, from=a, to=b)	*Devuelve la secuencia de n números reales igualmente espaciados que comienza en a y finaliza en b*
	`> seq(length=9, from=1, to=5)`
	`[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0`

sequence(a)	*Crea la secuencia de enteros consecutivos que comienza en 1 y termina en a.*
	`> sequence(4)`
	`[1] 1 2 3 4`

sequence(a:b)	*Crea la secuencia de a+b enteros consecutivos que comienza en 1 y termina en a y vuelve a comenzar en 1 y termina en b*
	`> sequence(4:5)`
	`[1] 1 2 3 4 1 2 3 4 5`

© Alfaomega - RC Libros

sequence(c(a,b))	*Crea la secuencia de a+b enteros consecutivos que comienza en 1 y termina en a y vuelve a comenzar en 1 y termina en b*
	
	`> sequence(c(10,5))`
	`[1] 1 2 3 4 5 6 7 8 9 10 1 2 3 4 5`

gl(a,b)	*Crea la secuencia de a*b enteros consecutivos que comienza en 1 y termina en a y que se repite cada uno b veces. Suele usarse para definir los niveles de un factor*
	
	`> gl(3, 5)`
	`[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3`
	`Levels: 1 2 3`

gl(a,b, length=n)	*Crea la secuencia de n enteros consecutivos que comienza en 1 y termina en a y que se repite cada uno b veces. Suele usarse para definir niveles de factores*
	
	`> gl(3, 5, length=30)`
	`[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 1`
	`1 1 1 1 2 2 2 2 2 3 3 3 3 3`
	`Levels: 1 2 3`

gl(a,b, label=vectordeetiquetas)	*Crea la secuencia de las °a etiquetas definidas en el vector de etiquetas y repite cada una de ellas b veces*
	
	`> gl(2, 6, label=c("Macho", "Hembra"))`
	`[1] Macho Macho Macho Macho Macho Macho`
	`[7] Hembra Hembra Hembra Hembra Hembra Hembra`
	`Levels: Macho Hembra`

Como las secuencias se comportan como vectores, todas los comandos y funciones de R aplicables a vectores también son aplicables a secuencias, tal y como se observa en los ejemplos siguientes:

```
> S1+S2
[1]  4  7 10 13 16
```


© Alfaomega - RC Libros


```
> S3^2
[1]  1  9 25 49 81

> log(S4)
[1] 0.0000000 0.6931472 1.0986123 1.3862944 1.6094379
> exp(S5)

[1] 7.389056 7.389056 7.389056 7.389056 7.389056
```

El operador ':' tiene prioridad sobre otros operadores aritméticos en una expresión:

```
> 1:10-1
[1] 0 1 2 3 4 5 6 7 8 9

> 1:(10-1)
[1] 1 2 3 4 5 6 7 8 9
```

La función *seq* también puede generar secuencias de números reales:

```
> seq(1, 5, 0.5)
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

donde el primer número indica el principio de la secuencia; el segundo, el final, y el tercero, el incremento que se debe usar para generar la secuencia. También se puede usar:

ARRAYS

Un array es una colección de elementos individuales bien ordenada. Se trata sencillamente de una lista de elementos y cada uno de los cuales está asociado a un entero positivo llamado índice del elemento, que representa el lugar (posición) que ocupa dentro de la lista. Lo esencial es que cada elemento está asociado a un índice, que también puede ser cero o negativo, que le identifica plenamente, de tal forma que, para realizar cambios en algún elemento del array, basta con referirse a sus índices.

Los arrays pueden ser de una o varias dimensiones, según tengan uno o varios conjuntos de índices que identifiquen a sus elementos. Los comandos más importantes que habilita R para el trabajo con arrays son los siguientes:

© Alfaomega - RC Libros

array(vector de datos, vector de dimensiones))	*Crea el array cuyo vector de datos y vector de dimensiones son los especificados.*
	Los ejemplos siguientes crean arrays de dos dimensiones
	`> array(1:5,c(2,3))`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 4 1`
	`> array(1:5,c(2,2))`
	`[,1] [,2]`
	`[1,] 1 3`
	`[2,] 2 4`
	El ejemplo siguiente crea un array tridimensional
	`> array(1:5,c(2,3,3))`
	`, , 1`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 4 1`
	`, , 2`
	`[,1] [,2] [,3]`
	`[1,] 2 4 1`
	`[2,] 3 5 2`
	`, , 3`
	`[,1] [,2] [,3]`
	`[1,] 3 5 2`
	`[2,] 4 1 3`

array(vector de datos, dim= vector de dimensiones))	*Sintaxis equivalente a la anterior*
	`> array(1:5,dim=c(2,3))`
	`[,1] [,2] [,3]`

© Alfaomega - RC Libros

	`[1,] 1 3 5`
	`[2,] 2 4 1`
	`> array(1:5,dim=c(2,3,3))`
	`, , 1`
	`[,1] [,2] [,3]`
	`[1,] 1 3 5`
	`[2,] 2 4 1`
	`, , 2`
	`[,1] [,2] [,3]`
	`[1,] 2 4 1`
	`[2,] 3 5 2`
	`, , 3`
	`[,1] [,2] [,3]`
	`[1,] 3 5 2`
	`[2,] 4 1 3`

array(vector de datos, vector de dos dimensiones))	*Crea la matriz cuyos elementos son los especificados en el vector de datos por columnas y con las dimensiones que se indican en el vector de dimensiones.* ***La matriz es un caso particular de array***
	`> A=array(c(9,8,9,6,1,3), c(2,3))`
	`> A`
	`[,1] [,2] [,3]`
	`[1,] 9 9 1`
	`[2,] 8 6 3`

array(vector de datos, dim= vector de dos dimensiones))	*Crea la matriz cuyos elementos son los especificados en el vector de datos por columnas y con las dimensiones que se indican en el vector de dimensiones*
	`> A=array(c(9,8,9,6,1,3), dim=c(2,3))`
	`> A`
	`[,1] [,2] [,3]`
	`[1,] 9 9 1`
	`[2,] 8 6 3`

© Alfaomega - RC Libros

array(vector de datos, dim=(1,length(vector de dats))	*Crea el vector cuyos elementos son los especificados en el vector de datos.* ***El vector es un caso particular de array***
	 `> A=array(c(9,8,9,6,1,3), dim=c(1,6))`
	 `> A`
	 `[,1] [,2] [,3] [,4] [,5] [,6]`
	 `[1,] 9 8 9 6 1 3`

array(vector de datos, (1,length(vector de dats))	*Crea el vector cuyos elementos son los especificados en el vector de datos*
	 `> A=array(c(9,8,9,6,1,3), c(1,6))`
	 `> A`
	 `[,1] [,2] [,3] [,4] [,5] [,6]`
	 `[1,] 9 8 9 6 1 3`

dim(X)	*Devuelve las dimensiones del array X*
	 `> A=array(1:5,dim=c(2,3,3));`
	 `> dim(A)`
	 `[1] 2 3 3`

aperm(array)	*Transpone un array*
	 `> A=array(1:5,dim=c(2,3,3))`
	 `> A`
	 `, , 1`
	 `[,1] [,2] [,3]`
	 `[1,] 1 3 5`
	 `[2,] 2 4 1`
	 `, , 2`
	 `[,1] [,2] [,3]`
	 `[1,] 2 4 1`
	 `[2,] 3 5 2`
	 `, , 3`
	 `[,1] [,2] [,3]`
	 `[1,] 3 5 2`
	 `[2,] 4 1 3`
	 `> AT=aperm(A)`

© Alfaomega - RC Libros


```
> AT
, , 1

     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    1
[3,]    3    5    2

, , 2

     [,1] [,2] [,3]
[1,]    2    4    1
[2,]    3    5    2
[3,]    4    1    3
```

Las variables indexadas pueden utilizarse en expresiones aritméticas y el resultado es una variable indexada formada a partir de las operaciones elemento a elemento de los vectores subyacentes. Los atributos dim de los operandos deben ser iguales en general y coincidirán con el vector de dimensiones del resultado. Así pues, si A, B y C son variables indexadas similares, entonces

> D = 2*A*B + C + 1

almacena en D una variable indexada similar, cuyo vector de datos es el resultado de las operaciones indicadas sobre los vectores de datos subyacentes a A, B y C.

Las reglas exactas correspondientes a los cálculos en que se mezclan variables indexadas y vectores deben ser estudiadas con detenimiento.

Tal y como ya se ha puesto de manifiesto, es importante resaltar que los vectores y las matrices son casos particulares de arrays.

LISTAS

En R, una lista es un objeto que consiste en una colección ordenada de objetos, conocidos como componentes. No es necesario que los componentes sean de la misma naturaleza, por lo que una lista puede estar compuesta de, por ejemplo, un vector numérico, un valor lógico, una matriz y una función. Este hecho diferencia a las listas de los vectores. Los comandos más importantes que habilita R para el trabajo con arrays son los siguientes:

© Alfaomega - RC Libros

list(elementos)	*Crea la lista con los elementos especificados*
	`> Lista=list(nombre="Pedro", esposa="María", no.hijos=3, edad.hijos=c(4,7,9))`
	`> Lista`
	`$nombre`
	`[1] "Pedro"`
	`$esposa`
	`[1] "María"`
	`$no.hijos`
	`[1] 3`
	`$edad.hijos`
	`[1] 4 7 9`

lista[[i]]	*Devuelve el elmento de i-ésimo de primer nivel de la lista*
	`> Lista[[2]]`
	`[1] "María"`
	`> Lista[[4]]`
	`[1] 4 7 9`

lista[[i]][j]	*Devuelve el elemento j-ésimo de segundo nivel del elemento iésimo de primer nivel de la lista*
	`> Lista[[4]][2]`
	`[1] 7`
	El 7 es el segundo elemento del vector (segundo nivel) que constituye el cuarto elemento de primer nivel de la lista

lista$elemento	*Devuelve el elemento especificado de primer nivel de la lista sin tener que utilizar su número.*
	`> Lista$esposa`
	`[1] "María"`
	`> Lista$nombre`
	`[1] "Pedro"`
	`> Lista$edad.hijos`
	`[1] 4 7 9`

© Alfaomega - RC Libros

lista[["elemento"]]	*Devuelve el elemento especificado de primer nivel de la lista sin tener que utilizar su número.*
	`> Lista[["esposa"]]`
	`[1] "María"`
	`> Lista[["nombre"]]`
	`[1] "Pedro"`
	`> Lista[["edad.hijos"]]`
	`[1] 4 7 9`

lista[i]	*Devuelve la sublista iésima de la lista*
	`> Lista[4]`
	`$edad.hijos`
	`[1] 4 7 9`

lenght(lista)	*Devuelve el número de elementos de primer nivel de la lista*
	`> length(Lista)`
	`[1] 4`

lista.ABC = c(lista.A, lista.B, lista.C)	*Devuelve la lista resultado de concatenar las listas especificadas*

Es muy importante distinguir claramente entre Lst[[1]] y Lst[1]. `[[. . .]]' es el operador utilizado para seleccionar un solo elemento, mientras que `[. . .]' es un operador general de indexado que devuelve una sublista de la lista. Esto es, Lst[[1]] es el primer objeto de la lista Lst, y si es una lista con nombres, el nombre no está incluido. Por su parte, Lst[1], es una sublista de la lista Lst que consiste en la primera componente. Si la lista tiene nombre, este se transfiere a la sublista.

Los nombres de los componentes pueden abreviarse hasta el mínimo de letras necesarios para identificarlos de modo exacto. Así, en

```
> Lista = list(coeficientes=c(1.3,4), covarianza=.87)
```

`Lst$coeficientes` puede especificarse mediante `Lista$coe`, y `Lista$covarianza` como `Lista$cov`.

© Alfaomega - RC Libros

MARCOS DE DATOS

La estructura marco de datos o "data frame" es una lista cuyos componentes deben ser vectores (numéricos, carácter, lógicos), factores, matrices numéricas, listas u otros marcos de datos. Si son vectores, deben ser todos de igual longitud, y si son matrices, deben tener todas el mismo número de filas.

Es frecuente estructurar un data frame como un conjunto de vectores columna de igual longitud. Podemos imaginar el marco de datos como una matriz rectangular, cuyas columnas son sus variables y que pueden ser de diferente tipo: numéricas, factores, lógicas, etc.

Para construir una variable de tipo marco de datos (data frame) se puede utilizar la función *data.frame()*. Veamos un ejemplo de data frame con 2 columnas, construido a partir de un vector numérico *v*, que será una columna de nombre valor y otro vector que será otra columna de nombre *caso* formada con datos cualitativos.

```
> v=c(2.1,1.4,6,3.5,8);v
[1] 2.1 1.4 6.0 3.5 8.0

>
DatosEstruc=data.frame(valor=v,caso=c("alto","bajo","medio","al
to","medio"))

> DatosEstruc

  valor  caso
1   2.1  alto
2   1.4  bajo
3   6.0  medio
4   3.5  alto
5   8.0  medio
```

Nos referimos separadamente a cada columna separando con el signo $ el nombre del data frame y el de la columna (que es como la denominación de la variable asociada).

```
> DatosEstruc$valor;DatosEstruc$caso

[1] 2.1 1.4 6.0 3.5 8.0
[1] alto  bajo  medio alto  medio
Levels: alto bajo medio
```


© Alfaomega - RC Libros

Para ver la dimensión de un objeto, en este caso el data frame DatosEstruc, se puede emplear la función *dim*:

```
> dim(DatosEstruc)

[1] 5 2
Es decir, tiene 2 columnas, y cada una 5 filas.
```

La función de R summary() aplicada al data frame nos da el resumen estadístico de sus variables (columnas).

```
> summary(DatosEstruc)
     valor          caso
 Min.   :1.4   alto :2
 1st Qu.:2.1   bajo :1
 Median :3.5   medio:2
 Mean   :4.2
 3rd Qu.:6.0
 Max.   :8.0
```

Veamos algunas instrucciones para acceder a la información del data frame. Se puede entender que se trabaja con el data frame como si fuese una matriz con filas y columnas:

```
> DatosEstruc[1,2]
[1] alto
Levels: alto bajo medio

> DatosEstruc[0,]
[1] valor caso
<0 rows> (or 0-length row.names)

> DatosEstruc[1,]
  valor caso
   1   2.1 alto

> DatosEstruc[1:3,]
  valor  caso
1   2.1  alto
2   1.4  bajo
3   6.0 medio
> DatosEstruc[1:3,2]
[1] alto  bajo  medio
Levels: alto bajo medio
```

© Alfaomega - RC Libros

```
> nuevo_orden=c(2,1,4,3,5);nuevo_orden
[1] 2 1 4 3 5

> nuevo_orden=c(2,1,4,3,5);nuevo_orden
[1] 2 1 4 3 5

> DatosEstruc_ordenado=DatosEstruc[nuevo_orden,];
> DatosEstruc_ordenado
  valor  caso
2   1.4  bajo
1   2.1  alto
4   3.5  alto
3   6.0 medio
5   8.0 medio
```

Veamos cómo generar desde el entorno R Console un vector de nombre aleatNormal con números aleatorios que proceden de una distribución normal de media 2 y desviación típica 4. Calculamos la media y la desviación tipica muestrales empleando diferentes funciones de R:

```
> aleatNormal=rnorm(15,2,4);aleatNormal
 [1] -1.919925930  0.880112011  2.915664570 -1.827781189  0.371865595
 [6]  1.737658371  1.299319151 -0.747209025  5.202930682 -2.966065683
[11]  1.353941243  0.002939512  0.946853283 -5.835235018  3.582367120

> mean(aleatNormal);sd(aleatNormal)
[1] 0.3331623
[1] 2.745425
> summary(aleatNormal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.8350 -1.2870  0.8801  0.3332  1.5460  5.2030
> numSummary(aleatNormal)

   mean       sd       0%      25%      50%      75%     100%    n
0.3331623 2.745425 -5.835235 -1.287495 0.880112 1.545800 5.202931 15
```

Los comandos más importantes que habilita R para el trabajo con marcos de datos son los siguientes:

A=data.frame(x,y,z,..)	*Crea una data frame cuyos elementos son las columnas x, y, z,..* `> datos1=data.frame(Peso=c(90,120,66),`

© Alfaomega - RC Libros

	`Altura=c(1.90,1.87,1.70))`
	`> datos1`
	`  Peso Altura`
	`1  90  1.90`
	`2  120  1.87`
	`3  66  1.70`

B=data.frame(A,x,y,..)	*Crea un data frame que contiene otros data frames como elementos (A es un data frame) y otros componentes adicionales (x,y,..). Se pueden anidar dataframes dentro de otros dataframes*
	`> datos2=data.frame(datos1,Sexo=c("Hombre", "Hombre", "Mujer"))`
	`> datos2`
	`  Peso Altura  Sexo`
	`1  90  1.90 Hombre`
	`2  120  1.87 Hombre`
	`3  66  1.70  Mujer`

A[n]	*Devuelve el elemento n-ésimo del data frame A*
	`> datos2[3]`
	`  Sexo`
	`1 Hombre`
	`2 Hombre`
	`3  Mujer`

A$elemento	*Extrae el elemento especificado del dataframe A y lo transforma en un vector.*
	`> Peso=datos2$Peso`
	`> Peso`
	`[1]  90 120  66`
	Esto es importante porque a partir de este momento al elemento extraído del data frame se le pueden aplicar todos los comandos y funciones relativos a vectores

I()	*Permite introducir un vector en un data frame.*
	`> nombres=c("Pepe","Paco","Pepita")`
	`> nombres`

© Alfaomega - RC Libros

	`[1] "Pepe" "Paco" "Pepita"`
	`> datos3=data.frame(datos2, Nombres=I(nombres))`
	`> datos3`
	`  Peso Altura  Sexo Nombres`
	`1  90  1.90 Hombre  Pepe`
	`2  120  1.87 Hombre  Paco`
	`3  66  1.70  Mujer  Pepita`

subset(A,select=elemento)	*Extrae el elemento especificado del data frame.*
	`> datos3`
	`  Peso Altura  Sexo Nombres`
	`1  90  1.90 Hombre  Pepe`
	`2  120  1.87 Hombre  Paco`
	`3  66  1.70  Mujer  Pepita`
	`> subset(datos3,select=c(Sexo,Nombres))`
	`  Sexo Nombres`
	`1 Hombre  Pepe`
	`2 Hombre  Paco`
	`3  Mujer  Pepita`

transform(A, transformación)	*Realiza la transformación especificada en un elemento del data frame A*
	`> transform(datos3,logPeso=log(Peso))`
	`  Peso Altura  Sexo Nombres  logPeso`
	`1  90  1.90 Hombre  Pepe 4.499810`
	`2  120  1.87 Hombre  Paco 4.787492`
	`3  66  1.70  Mujer  Pepita 4.189655`
	`> transform(datos3,IMC=Peso/(Altura)^2)`
	`  Peso Altura  Sexo Nombres  IMC`
	`1  90  1.90 Hombre  Pepe 24.93075`
	`2  120  1.87 Hombre  Paco 34.31611`
	`3  66  1.70  Mujer  Pepita 22.83737`

dim(A)	*Devuelve la dimensión del data frame A*
	`> dim(datos3)`
	`[1] 3 4`

© Alfaomega - RC Libros

attach(A)	*Habilita los elementos del data frame A para que sean accesibles como vectores sin necesidad de utilizar el comando A$elemento*
	`> attach(datos3)`
	`The following objects are masked _by_ .GlobalEnv:`
	`  Nombres, Peso, Sexo`
	`> Peso`
	`[1]  90 120  66`
	`> sum(Peso)`
	`[1] 276`

dettach(A)	*Deshabilita los elementos del data frame A para que sean accesibles como vectores sin necesidad de utilizar el comando A$elemento*

expand.grid(x,y,..)	*Crea un dataframe con todas las combinaciones de los valores de x,y,..*
	`> datos4=expand.grid(c(1,2,3),c(10,20))`
	`> datos4`
	`  Var1 Var2`
	`1  1  10`
	`2  2  10`
	`3  3  10`
	`4  1  20`
	`5  2  20`
	`6  3  20`
	`> datos5=expand.grid(a=c(60,80), p=c(100, 300), sexo=c("Macho", "Hembra"))`
	`> datos5`
	`  a  p  sexo`
	`1 60 100  Macho`
	`2 80 100  Macho`
	`3 60 300  Macho`
	`4 80 300  Macho`
	`5 60 100 Hembra`
	`6 80 100 Hembra`
	`7 60 300 Hembra`
	`8 80 300 Hembra`

© Alfaomega - RC Libros

summarize(A)	*Halla estadísticos descriptivos para todas las variables columna del data frame A*

```
> summary(datos4)
      Var1           Var2
 Min.   :1.00   Min.   :10
 1st Qu.:1.25   1st Qu.:10
 Median :2.00   Median :15
 Mean   :2.00   Mean   :15
 3rd Qu.:2.75   3rd Qu.:20
 Max.   :3.00   Max.   :20
```

IMPORTACIÓN DE DATOS ESTADÍSTICOS COMO DATA FRAMES

Importación de archivos de texto

Ya sabemos que R admite la importación directa de archivos de texto ASCII. La sintaxis para la importación es la siguiente:

mydata=read.table("c:/mydata.txt", header=TRUE, sep=",")

La sintaxis anterior importa el archivo de nombre mydata.txt situado en C:/ manteniendo las cabeceras del fichero como nombres de sus variables y considerando que el separador entre elementos es la coma. No obstante, si el fichero está encolumnado (variables como columnas), no es necesario especificar el separador.

Cuando se importa un fichero ASCII, el lenguaje R lo considera con una estructura de data frame. Eso significa que solo podrán aplicarse comandos y funciones relativos a data frames. Veamos un ejemplo:

```
> datos = read.table("H:/R/data/zonasmadR.txt", header=TRUE)
> datos
           b   pt  p14  p65   p10 anal  nes  ocu ocuin ocuser  tec  pd   tm
1     Centro 166.5 23.3 38.1 152.8  4.2 21.4 54.1   7.6   41.7  8.8 0.8 10.3
2 Arganzuela 121.1 23.5 18.4 106.1  2.0 16.5 69.4   7.6   28.6  7.2 0.6  8.4
3     Retiro 126.0 27.2 16.8 109.2  1.2 28.1 39.9   6.3   30.1 10.4 1.9  4.7
4  Salamanca 180.0 30.5 33.4 162.1  0.5 45.3 57.5   7.6   45.1 16.1 2.6  5.4
5  Chamartín 180.0 30.5 16.1 130.3  1.3 39.3 48.1   7.2   35.8 14.5 2.8  4.8
6     Tetuán 164.2 31.3 23.5 145.1  4.2 24.2 52.3   9.6   37.9  9.6 1.1 12.2
7   Chamberi 182.7 29.4 35.0 165.4  1.8 47.2 59.4   7.5   46.4 17.1 2.4  6.1
8 Fuencarral 176.2 51.3 15.6 142.2  3.6 21.6 95.6  10.3   38.7 11.1 1.6 14.3
9    Moncloa 108.4 23.4 13.4  94.2  1.5 32.5 34.6   5.3   26.0  8.7 1.4  5.5
10    Latina 289.5 79.5 23.1 239.7  6.0 22.7 86.6  17.7   59.8 10.4 1.3 26.4
11 Carabanchel 255.9 60.5 24.1 218.3  7.3 16.6 77.4  19.4   50.1  7.5 1.0 28.2
12 Villaverde 195.0 48.5 16.1 166.1  8.3  9.0 56.6  19.3   30.7  3.7 0.5 26.3
13   Mediodía 171.7 49.3 11.1 139.9  9.8  5.5 48.5  13.3   28.1  2.9 0.3 22.9
14   Vallecas 186.2 42.2 20.3 159.8 10.3  7.2 53.7  13.6   32.5  3.1 0.3 23.8
15 Moratalaz 145.9 40.8 10.9 121.4  3.9 10.1 73.7  16.5   49.0 11.6 2.1 19.7
16 CiudadLineal 135.1 55.3 21.9 201.5  4.3 28.2 73.7  16.5   49.0 11.6 2.1 19.7
17    SanBlas 137.7 32.1 10.3 118.5  6.0  6.3 41.4  12.2   24.1  2.9 0.2 18.2
18 Hortaleza 167.9 51.4 10.1 132.6  4.0 15.5 51.6  12.3   33.7  7.9 1.4 15.8
```


© Alfaomega - RC Libros


```
> class(datos)
[1] "data.frame"
```

Podemos aplicar al data frame la función summary.

```
> summary(datos)

          b            pt              p14              p65              p10
 Arganzuela  : 1   Min.   :108.4   Min.   :23.30   Min.   :10.10   Min.   : 94.2
 Carabanchel : 1   1st Qu.:139.8   1st Qu.:29.68   1st Qu.:13.95   1st Qu.:123.6
 Centro      : 1   Median :169.8   Median :36.45   Median :17.60   Median :143.7
 CiudadLineal: 1   Mean   :171.7   Mean   :40.56   Mean   :19.90   Mean   :150.3
 Chamartín   : 1   3rd Qu.:182.0   3rd Qu.:50.80   3rd Qu.:23.40   3rd Qu.:164.6
 Chamberi    : 1   Max.   :289.5   Max.   :79.50   Max.   :38.10   Max.   :239.7
 (Other)     :12
      anal             nes              ocu              ocuin            ocuser
 Min.   : 0.500   Min.   : 5.50   Min.   :34.60   Min.   : 5.30   Min.   :24.10
 1st Qu.: 1.850   1st Qu.:11.45   1st Qu.:49.27   1st Qu.: 7.60   1st Qu.:30.25
 Median : 4.100   Median :21.50   Median :55.35   Median :11.25   Median :36.85
 Mean   : 4.456   Mean   :22.07   Mean   :59.67   Mean   :11.66   Mean   :38.18
 3rd Qu.: 6.000   3rd Qu.:28.18   3rd Qu.:72.62   3rd Qu.:15.78   3rd Qu.:46.08
 Max.   :10.300   Max.   :47.20   Max.   :95.60   Max.   :19.40   Max.   :59.80

      tec              pd              tm
 Min.   : 2.900   Min.   :0.200   Min.   : 4.700
 1st Qu.: 7.275   1st Qu.:0.650   1st Qu.: 6.675
 Median : 9.200   Median :1.350   Median :15.050
 Mean   : 9.172   Mean   :1.356   Mean   :15.150
 3rd Qu.:11.475   3rd Qu.:2.050   3rd Qu.:22.100
 Max.   :17.100   Max.   :2.800   Max.   :28.200
```

Sin embargo, no podemos utilizar sus variables individualmente como vectores para trabajar con ellas. Si intentamos sumar los elementos de pt obtenemos error.

```
> sum(pt)
Error in sum(pt) : invalid 'type' (closure) of argument
```

Para resolver este problema, basta con definir las variables del data frame como vectores.

```
> pt=datos$pt
> sum(pt)
[1] 3090
```

© Alfaomega - RC Libros

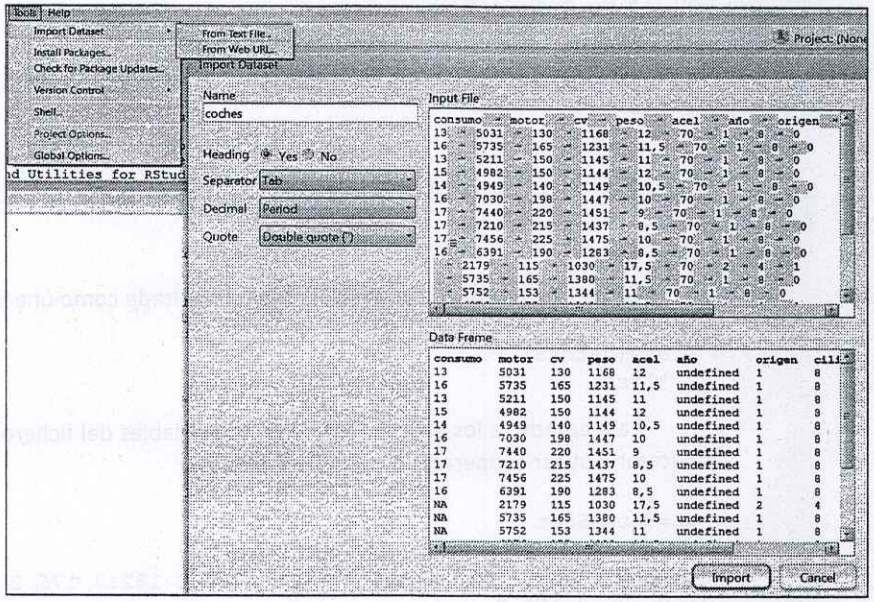
Pero la tarea anterior habría que realizarla variable a variable. Sin embargo, como ya hemos visto, pueden transformarse todas las columnas del data frame a vectores con la función *attach*.

```
> attach(datos);
```

Ahora ya podemos utilizar cualquier columna del data frame como un vector.

```
> sum(ocuin)
[1] 209.8
```

La importación del fichero ASCII también se puede hacer desde el entorno de RStudio mediante la opción *Import Dataset → From Text File*. En la zona inferior de la figura siguiente se observa que la importación se realiza con estructura de Data Frame. Al hacer clic en *Import*, se realiza la importación.



Importación de archivos Excel

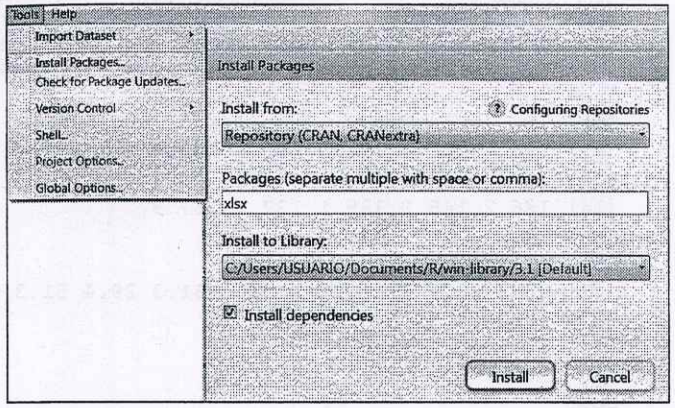
Para importar ficheros de Excel, es necesario instalar y cargar previamente la librería *xlsx*.

© Alfaomega - RC Libros

Para instalar la librería *xls* , se utiliza la siguiente sintaxis:

```
> install.packages("xlsx")
```

También puede utilizarse la opción *Install packages* del menú *Tools* de RStudio, según se indica en la figura siguiente:



También se puede utilizar la opción *Install Packages* del menú *Packages* de *RConsole*.

Una vez instalada la librería, se habilita mediante la sintaxis:

```
> library(xlsx)
```

Para realizar la importación del archivo *zonasmad.xls* situado en la ruta *H:\R/datos*, se utiliza la siguiente sintaxis:

```
> datosE = read.xlsx("H:/R/data/zonasmad.xls",1)
```

Con la sintaxis anterior se lee la primera hoja del archivo Excel. Para leer otra hoja cualquiera se utiliza la siguiente sintaxis:

```
> datosE = read.xlsx("H:/R/data/zonasmad.xls",sheetName =
"nombrehoja"))
```

Importación de archivos SPSS

Para importar ficheros de IBM SPSS conservando etiquetas y sus valores es necesario instalar y cargar previamente la librería *foreing*. La sintaxis es:

© Alfaomega - RC Libros

```
> install(foreign)
> library(foreign)
> datosS=read.spss("H:/R/data/zonasmad.sav", use.value.labels =
TRUE)
> datosS
$b
 [1] "Centro          " "Arganzuela       " "Retiro           "
 [4] "Salamanca       " "Chamartín        " "Tetuán          "
 [7] "Chamberi        " "Fuencarral      " "Moncloa         "
[10] "Latina          " "Carabanchel     " "Villaverde      "
[13] "Mediodía        " "Vallecas        " "Moratalaz       "
[16] "Ciudad Lineal   " "San Blas        " "Hortaleza       "

$pt
 [1] 166.5 121.1 126.0 180.0 180.0 164.2 182.7 176.2 108.4 289.5
255.9 195.0 171.7
[14] 186.2 145.9 135.1 137.7 167.9

$p14
 [1] 23.3 23.5 27.2 30.5 30.5 31.3 29.4 51.3 23.4 79.5 60.5 48.5
49.3 42.2 40.8
[16] 55.3 32.1 51.4

$p65
 [1] 38.1 18.4 16.8 33.4 16.1 23.5 35.0 15.6 13.4 23.1 24.1 16.1
11.1 20.3 10.9
[16] 21.9 10.3 10.1

$p10
 [1] 152.8 106.1 109.2 162.1 130.3 145.1 165.4 142.2  94.2 239.7
218.3 166.1 139.9
[14] 159.8 121.4 201.5 118.5 132.6

$anal
 [1]  4.2  2.0  1.2   NA  1.3  4.2  1.8  3.6  1.5  6.0  7.3  8.3
9.8 10.3  3.9
[16]  4.3  6.0  4.0


$nes
 [1] 21.4 16.5 28.1 45.3 39.3 24.2 47.2 21.6 32.5 22.7 16.6  9.0
5.5  7.2 10.1
[16] 28.2  6.3 15.5

$ocu
```


© Alfaomega - RC Libros


```
 [1] 54.1 69.4 39.9 57.5 48.1 52.3 59.4 95.6 34.6 86.6 77.4 56.6
48.5 53.7 73.7
[16] 73.7 41.4 51.6

$ocuin
 [1]  7.6  7.6  6.3  7.6  7.2  9.6  7.5 10.3  5.3 17.7 19.4 19.3
13.3 13.6 16.5
[16] 16.5 12.2 12.3

$ocuser
 [1] 41.7 28.6 30.1 45.1 35.8 37.9 46.4 38.7 26.0 59.8 50.1 30.7
28.1 32.5 49.0
[16] 49.0 24.1 33.7

$tec
 [1]  8.8  7.2 10.4 16.1 14.5  9.6 17.1 11.1  8.7 10.4  7.5  3.7
2.9  3.1 11.6
[16] 11.6  2.9  7.9

$pd
 [1] 0.8 0.6 1.9 2.6 2.8 1.1 2.4 1.6 1.4 1.3 1.0 0.5 0.3 0.3 2.1
2.1 0.2 1.4

$tm
 [1] 10.3  8.4  4.7  5.4  4.8 12.2  6.1 14.3  5.5 26.4 28.2 26.3
22.9 23.8 19.7
[16] 19.7 18.2 15.8
```

Comprobamos que la importación se ha realizado como una lista.

```
> class(datosS)
[1] "list"
```

Para acceder a los elementos de la lista (variables del fichero) como vectores es necesario utizar el operador $ variable a variable.

```
> pt=datosS$pt
> pt
 [1] 166.5 121.1 126.0 180.0 180.0 164.2 182.7 176.2 108.4 289.5
255.9 195.0 171.7
[14] 186.2 145.9 135.1 137.7 167.9
> sum(pt)
[1] 3090
```

© Alfaomega - RC Libros

También podemos utilizar la función attach para habilitar simultáneamente todas las variables del fichero como vectores.

```
> attach(datosS)

> sum(ocu)
[1] 1074.1
```

También podíamos haber importado el fichero de SPSS como un data frame mediante la siguiente sintaxis:

```
> datosS=read.spss("H:/R/data/zonasmad.sav", use.value.labels =
TRUE, to.data.frame = TRUE)
```

Importación de archivos de STATA

Para importar ficheros de STATA conservando etiquetas y sus valores es necesario instalar y cargar previamente la librería *foreign*. La sintaxis es:

```
> install(foreign)
> library(foreign)
> datosSTATA=read.dta("H:/R/data/zonasmad.dta")
> datosSTATA
```

	b	pt	p14	p65	p10	anal	nes	ocu	ocuin	ocuser	tec	pd	tm
1	Centro	166.5	23.3	38.1	152.8	4.2	21.4	54.1	7.6	41.7	8.8	0.8	10.3
2	Arganzuela	121.1	23.5	18.4	106.1	2.0	16.5	69.4	7.6	28.6	7.2	0.6	8.4
3	Retiro	126.0	27.2	16.8	109.2	1.2	28.1	39.9	6.3	30.1	10.4	1.9	4.7
4	Salamanca	180.0	30.5	33.4	162.1	NA	45.3	57.5	7.6	45.1	16.1	2.6	5.4
5	Chamartín	180.0	30.5	16.1	130.3	1.3	39.3	48.1	7.2	35.8	14.5	2.8	4.8
6	Tetuán	164.2	31.3	23.5	145.1	4.2	24.2	52.3	9.6	37.9	9.6	1.1	12.2
7	Chamberi	182.7	29.4	35.0	165.4	1.8	47.2	59.4	7.5	46.4	17.1	2.4	6.1
8	Fuencarral	176.2	51.3	15.6	142.2	3.6	21.6	95.6	10.3	38.7	11.1	1.6	14.3
9	Moncloa	108.4	23.4	13.4	94.2	1.5	32.5	34.6	5.3	26.0	8.7	1.4	5.5
10	Latina	289.5	79.5	23.1	239.7	6.0	22.7	86.6	17.7	59.8	10.4	1.3	26.4
11	Carabanchel	255.9	60.5	24.1	218.3	7.3	16.6	77.4	19.4	50.1	7.5	1.0	28.2
12	Villaverde	195.0	48.5	16.1	166.1	8.3	9.0	56.6	19.3	30.7	3.7	0.5	26.3
13	Mediodía	171.7	49.3	11.1	139.9	9.8	5.5	48.5	13.3	28.1	2.9	0.3	22.9
14	Vallecas	186.2	42.2	20.3	159.8	10.3	7.2	53.7	13.6	32.5	3.1	0.3	23.8
15	Moratalaz	145.9	40.8	10.9	121.4	3.9	10.1	73.7	16.5	49.0	11.6	2.1	19.7
16	Ciudad Lineal	135.1	55.3	21.9	201.5	4.3	28.2	73.7	16.5	49.0	11.6	2.1	19.7
17	San Blas	137.7	32.1	10.3	118.5	6.0	6.3	41.4	12.2	24.1	2.9	0.2	18.2
18	Hortaleza	167.9	51.4	10.1	132.6	4.0	15.5	51.6	12.3	33.7	7.9	1.4	15.8

En este caso la importación se realiza en modo vectorial, de modo que todas las variables son ya accesibles como vectores.

© Alfaomega - RC Libros


```
> pt
 [1] 166.5 121.1 126.0 180.0 180.0 164.2 182.7 176.2 108.4 289.5
255.9 195.0 171.7
[14] 186.2 145.9 135.1 137.7 167.9
> sum(pt)
[1] 3090
```

Importación de archivos de SAS

Para importar ficheros de SAS es necesario instalar y cargar previamente la librería *foreign*. La sintaxis es:

```
> install(foreign)
> library(foreign)
> datosSAS=read.ssd("H:/R/data/", "zonasmad")
```

SERIES DE TIEMPO

La función *ts* crea un objeto de clase "ts" (serie de tiempo) a partir de un vector (serie de tiempo inivariante) o una matriz (serie multivariada). La sintaxis y las opciones de la función *ts* se presentan a continuación:

```
ts(data = NA, start = 1, end = numeric(0), frequency = 1,
deltat = 1, class, names)
```

Las opciones tienen la siguiente interpretación:

`data` un vector o una matriz

`start` el tiempo de la primera observación ya sea un número o un vector con dos enteros (ver ejemplo más abajo)

`end` el tiempo de la última observación especificado de la misma manera que `start`

`frequency` el número de observaciones por unidad de tiempo

`deltat` la fracción del periodo de muestreo entre observaciones sucesivas (ej. `1/12` para datos mensuales); únicamente se debe especificar o `frequency` o `deltat`

`class` clase que se debe asignar al objeto; por defecto es "ts" para una serie univariada, y `c("mts", "ts")` para una serie multivariada

© Alfaomega - RC Libros

`names` para una serie multivariada, un vector de tipo carácter con los nombres de las series individuales; por defecto los nombres de las columnas de `data`, o `Serie 1, Serie 2, . . .`

A continuación se presentan algunos ejemplos con `ts()`:

```
> ts(1:10, start = 1959)
Time Series:
Start = 1959
End = 1968
Frequency = 1
[1] 1 2 3 4 5 6 7 8 9 10

> ts(1:47, frequency = 12, start = c(1959, 2))
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1959 1 2 3 4 5 6 7 8 9 10 11
1960 12 13 14 15 16 17 18 19 20 21 22 23
1961 24 25 26 27 28 29 30 31 32 33 34 35
1962 36 37 38 39 40 41 42 43 44 45 46 47

> ts(1:10, frequency = 4, start = c(1959, 2))
Qtr1 Qtr2 Qtr3 Qtr4
1959 1 2 3
1960 4 5 6 7
1961 8 9 10

> ts(matrix(rpois(36, 5), 12, 3), start=c(1961, 1),
frequency=12)
Series 1 Series 2 Series 3
Jan 1961 8 5 4
20
Feb 1961 6 6 9
Mar 1961 2 3 3
Apr 1961 8 5 4
May 1961 4 9 3
Jun 1961 4 6 13
Jul 1961 4 2 6
Aug 1961 11 6 4
Sep 1961 6 5 7
Oct 1961 6 5 7
Nov 1961 5 5 7
Dec 1961 8 5 2
```


© Alfaomega - RC Libros

Ejercicio 3-1. Consideramos la secuencia de enteros del cero al seis. Hallar la secuencia que contenga sus cuadrados.

```
> S1=seq(1:6)
> S1
[1] 1 2 3 4 5 6
> S2=seq(1:6)^2
> S2
[1]  1  4  9 16 25 36
```

Ejercicio 3-2. Calcular: $\sum ne^{-n}$ *y* $\prod n^2 2^n/e^n$ *(n=1..10)*

```
> sum((1:10)*exp(1:10)^(-(1:10)))
[1] 0.4048814

> prod((1:10)^2*2^(1:10)/exp(1)^(1:10))
[1] 616565.6
```

Ejercicio 3-3. Calcular: $\sum_{k=1}^{100} k^3$, $\sum_{k=1}^{100} k^4$ *y* $\sum_{k=1}^{100} k^5$

```
> sum(1:100)^3
[1] 128787625000

> options(digits=20)
> sum(1:100)^4
[1] 650377506250000.00000

> sum(1:100)^5
[1] 3284406406562500096
```

Ejercicio 3-4. Calcular la media aritmética y la media geométrica de los 1000 primeros números enteros positivos.

```
> M=sum(1:100)/100
> M
[1] 50.5

> G=prod(1:100)^(1/100)
> G
[1] 37.99269
```

© Alfaomega - RC Libros

Ejercicio 3-5. Consideremos un array tridimensional cuyos elementos sean tres matrices cuadradas de orden 3 formadas por los 27 primeros números enteros positivos. Definir el array y hallar la matriz inversa del producto de sus elementos. Hallar también los autovalores y autovectores de la suma de sus elementos.

Comenzamos definiendo adecuadamente el array tridimensional.

```
> A=array(1:27, dim=c(3,3,3))

> A
, , 1

     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

, , 2

     [,1] [,2] [,3]
[1,]   10   13   16
[2,]   11   14   17
[3,]   12   15   18

, , 3

     [,1] [,2] [,3]
[1,]   19   22   25
[2,]   20   23   26
[3,]   21   24   27
```

A continuación extraemos sus elementos (tres matrices cuadradas de orden 3).

```
> M1=A[,,1]
> M1
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> M2=A[,,2]
> M2
```


© Alfaomega - RC Libros


```
     [,1] [,2] [,3]
[1,]   10   13   16
[2,]   11   14   17
[3,]   12   15   18

> M3=A[,,3]
> M3
     [,1] [,2] [,3]
[1,]   19   22   25
[2,]   20   23   26
[3,]   21   24   27
```

Ahora calculamos la matriz inversa del producto de los tres elementos.

```
> solve(M1*M2*M3)
            [,1]       [,2]        [,3]
[1,]  0.16393199 -0.2873789  0.12738046
[2,] -0.20605196  0.3538731 -0.15417236
[3,]  0.07342013 -0.1250817  0.05434675
```

Por último, calculamos los valores y vectores propios de la suma de los tres elementos.

```
> eigen(M1+M2+M3)
$values
[1]  1.272729e+02 -1.272856e+00 -6.997266e-16

$vectors
           [,1]        [,2]       [,3]
[1,] -0.5356313 -0.79526971  0.4082483
[2,] -0.5763903 -0.09853771 -0.8164966
[3,] -0.6171493  0.59819429  0.4082483
```

Ejercicio 3-6. Generar un array tridimensional que contenga los ocho primeros números naturales no negativos y cuyos tres índices varían de 1 a 2. Acceder a sus elementos y sumarlos. Generar el array cuyos elementos sean los cuadrados del array anterior y sumarlos.

```
> A=array(1:8, dim=c(2,2,2))
> A
, , 1

     [,1] [,2]
[1,]    1    3
```

© Alfaomega - RC Libros

```
[2,]    2    4

, , 2

     [,1] [,2]
[1,]    5    7
[2,]    6    8

> A1=A[,,1]
> A1
     [,1] [,2]
[1,]    1    3
[2,]    2    4

> A2=A[,,2]

> A2
     [,1] [,2]
[1,]    5    7
[2,]    6    8

> S=A1+A2
> S
     [,1] [,2]
[1,]    6   10
[2,]    8   12

> B=A^2
> B
, , 1

     [,1] [,2]
[1,]    1    9
[2,]    4   16

, , 2

     [,1] [,2]
[1,]   25   49
[2,]   36   64

> B1=B[,,1]
> B1
     [,1] [,2]
[1,]    1    9
```


© Alfaomega - RC Libros


```
[2,]    4   16

> B2=B[,,2]
> B2
     [,1] [,2]
[1,]   25   49
[2,]   36   64

> S2=B1+B2
> S2
     [,1] [,2]
[1,]   26   58
[2,]   40   80
```

Ejercicio 3-7. A continuación se presenta una tabla con los alumnos de una clase, sus tallas y sus pesos. Se trata de calcular el peso medio y la altura media de los alumnos de la clase.

Alumno	Altura	Peso
Pérez	1.81	85
González	1.70	72
Rodríguez	1.75	80
Alonso	1.90	95
López	1.83	85
Tello	1.75	72
Domínguez	1.81	84

Comenzaremos asignando la información a un data frame de R.

```
> Altura=c(1.81,1.70,1.75,1.90,1.83,1.75,1.81)
> Peso=c(85,72,80,95,85,72,84)
> Datos=data.frame(Alumno,Altura,Peso)
> Datos
    Alumno Altura Peso
1    Pérez   1.81   85
2 González   1.70   72
3 Rodríguez  1.75   80
4   Alonso   1.90   95
5    López   1.83   85
6    Tello   1.75   72
7 Domínguez  1.81   84
```

© Alfaomega - RC Libros

A continuación habilitamos las variables del data frame para que puedan ser tratadas como vectores.

```
> attach(Datos)
The following objects are masked _by_ .GlobalEnv:

    Altura, Alumno, Peso
```

Finalmente calculamos la altura media y el peso medio.

```
> ALTURAMEDIA=sum(Altura)/length(Altura)
> ALTURAMEDIA
[1] 1.792857

> PESOMEDIO=sum(Peso)/length(Peso)
> PESOMEDIO
[1] 81.85714
```


© Alfaomega - RC Libros

CAPÍTULO 4

LENGUAJE DE R. FUNCIONES PREDEFINIDAS Y CREADAS A MEDIDA

FUNCIONES

El concepto de función, lo mismo que el concepto de variable, es fundamental a la hora de trabajar con un software matemático. Evidentemente, el concepto teórico de función matemática es fijo e independiente del paquete de software. Pero la forma de implementar y manejar las funciones es muy característica de cada programa en particular. R permite definir y manejar las funciones de modo muy simple. A su vez, R incorpora un conjunto extenso de funciones, que abordaremos a lo largo de los capítulos.

DEFINICIÓN DE FUNCIONES

Ya hemos visto que R incorpora múltiples funciones para usar en el trabajo cotidiano con el programa. Pero, además, el programa ofrece la posibilidad de definir funciones a medida. La sintaxis para definir una función es la siguiente:

© Alfaomega - RC Libros

Nombrefuncion = function(argumento1, argumento2,...,argumenton)
{
sentencias
return(object)
}

El objeto devuelto puede ser cualquier tipo de dato (vector, matriz, array, lista, factor, data frame, etc.).

EL EDITOR DE R

Las funciones se pueden escribir sobre la ventana de comandos de R, pero eso es incómodo, sobre todo para posibles correcciones de errores. Por esta razón, se utiliza el editor de R, que permite escribir el programa en modo texto, con facilidad para realizar correcciones en caso de errores.

El editor de R se invoca desde RStudio mediante *File → New File → R Script*, tal y como se indica en la Figura 4-1.

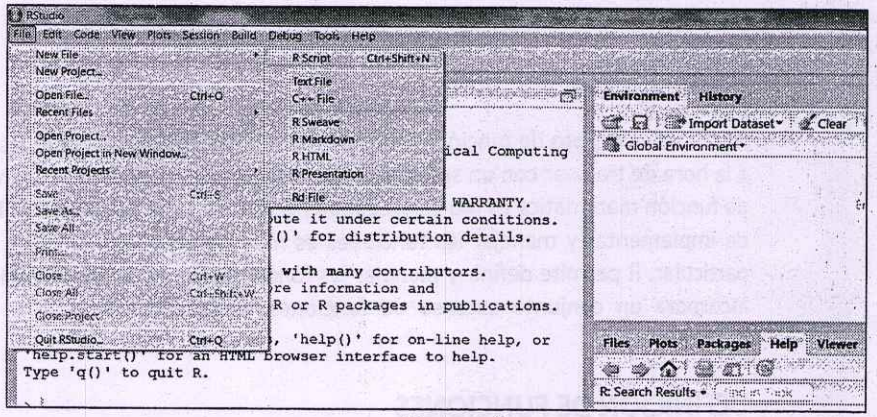


Figura 4-1

Se obtiene el editor (Figura 4-2), para poder escribir la sintaxis de funciones y programas. El botón *Run* permite la ejecución de la sintaxis y el botón permite guardar la sintaxis.

© Alfaomega - RC Libros

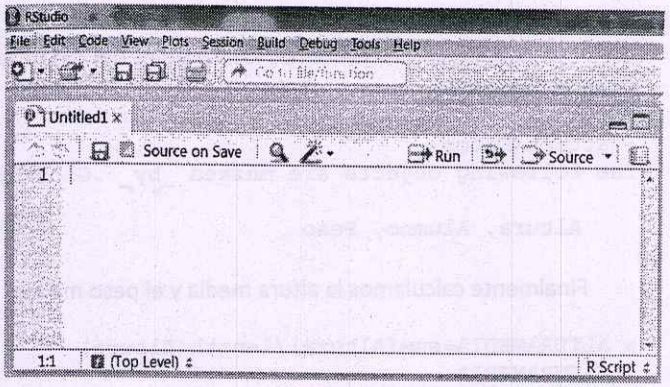


Figura 4-2

Como ejemplo vamos a construir una función que realice la suma de dos números cualesquiera. Escribimos la sintaxis en el editor, la seleccionamos y hacemos clic en el botón *Run* (Figura 4-3).

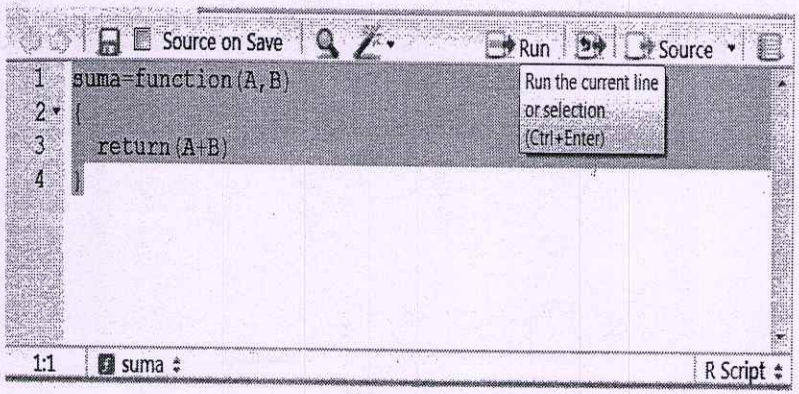


Figura 4-3

Sobre RConsole (ventana de comandos de R) vemos reflejada la sintaxis que hemos ejecutado desde el editor (Figura 4-4). Si no aparecen errores en la consola, la sintaxis del programa es correcta.

© Alfaomega - RC Libros

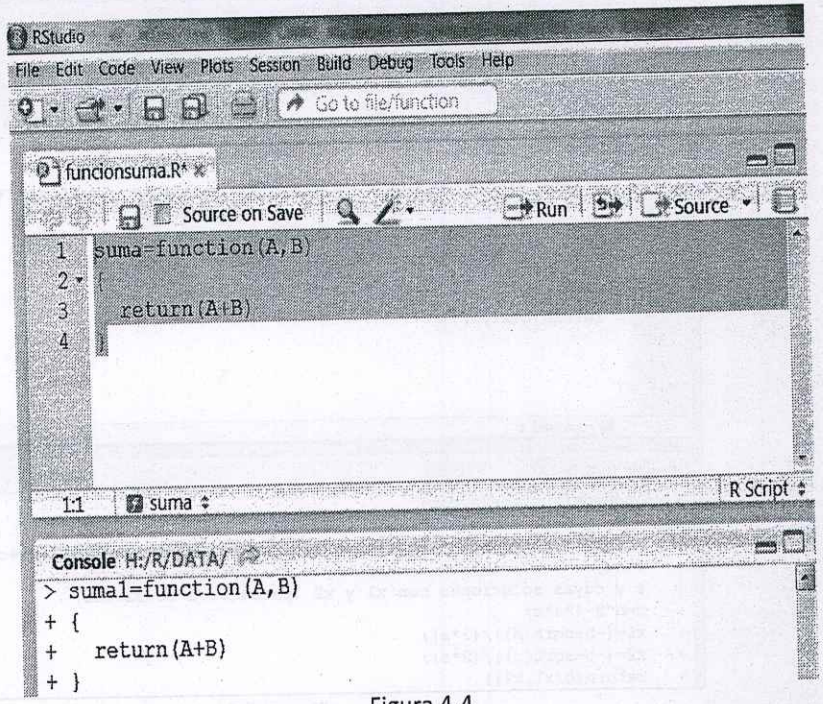


Figura 4-4

Ahora ya podemos ejecutar la función desde la ventana de comandos de R.

```
> suma(3,8)
[1] 11
> suma(-1,75)
[1] 74
```

Como no se han puesto restricciones sobre los argumentos de la función, sería posible también sumar objetos de cualquier tipo, por ejemplo, vectores o matrices. Veamos algunos ejemplos:

```
> suma(c(1,2,3),c(4,5,6))
[1] 5 7 9

> A=matrix(c(1,2,3,4,5,6),3,2)
> A
     [,1] [,2]
```


© Alfaomega - RC Libros


```
[1,]    1    4
[2,]    2    5
[3,]    3    6

> B=matrix(c(11,12,13,14,15,16),3,2)
> B
     [,1] [,2]
[1,]   11   14
[2,]   12   15
[3,]   13   16

> suma(A,B)
     [,1] [,2]
[1,]   12   18
[2,]   14   20
[3,]   16   22
```

Con el botón *Guardar* guardamos el fichero. Como se trata de un programa de R, el fichero se guardará con extensión .R en el subdirectorio de trabajo por defecto.

Con la opción *File → Open File* se volverá a abrir el fichero en el editor, en caso de que sea necesario.

A continuación vamos a definir la función *fun1(x)=x^3-2x+cos(x)*, creando el correspondiente *fichero* de nombre *fun1.R* mediante la sintaxis que se observa en la Figura 4-5.

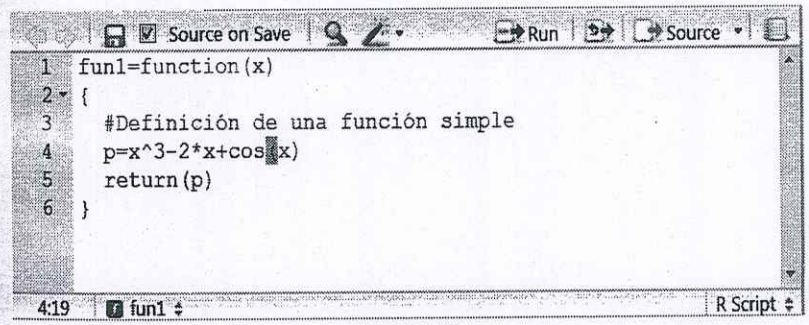


Figura 4-5

© Alfaomega - RC Libros

Ahora ya podemos ejecutar la función.

```
> fun1=function(x)
+ {
+   #Definición de una función simple
+   p=x^3-2*x+cos(x)
+   return(p)
+ }

> fun1(pi)
[1] 23.72309

> fun1(c(pi,2*pi,3*pi,4*pi))
[1]   23.72309  236.48384  817.31991 1960.26897
```

En el siguiente ejemplo (Figura 4-6) se ilustra la definición de una función que resuelve una ecuación de segundo grado.

Se observa que esta función tiene tres parámetros de entrada y dos de salida. Los tres parámetros de entrada se definen sencillamente como los tres argumentos de la función. Los dos parámetros de salida se definen como argumentos del comando return, pero en modo vector, ya que el comando return no admite más de un parámetro.

Guardamos el *fichero* con el nombre *rootquad.R*

La función, definida de esta forma, permitirá resolver en R cualquier ecuación de segundo grado con soluciones reales.

© Alfaomega - RC Libros

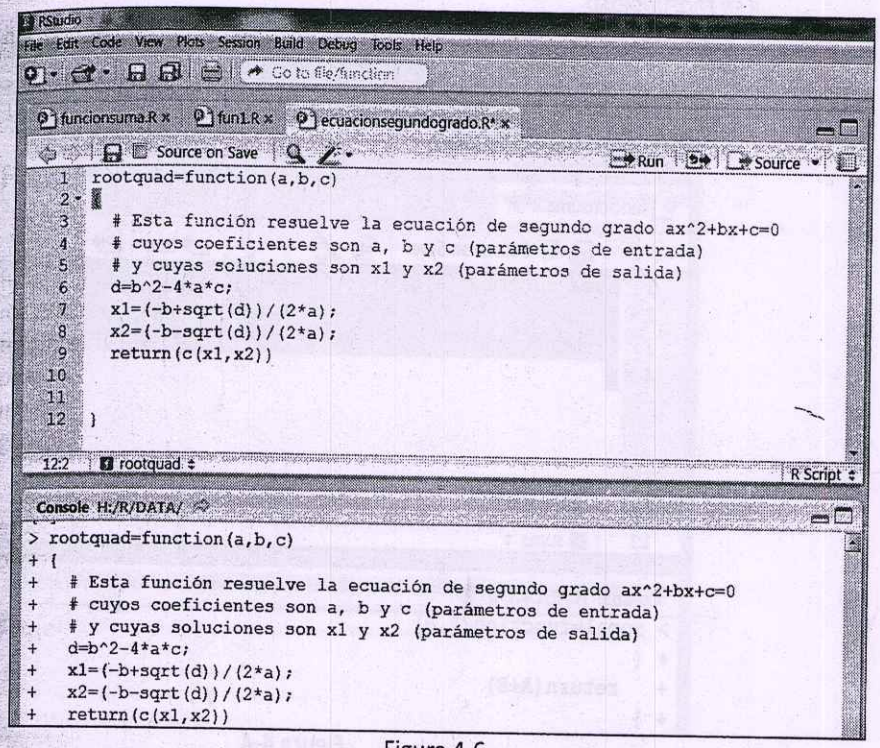


Figura 4-6

Ahora resolveremos la ecuación *x^2-6x+2=0* de la siguiente forma:

```
> rootcuad(1,-6,2)
[1] 5.6457513 0.3542487
```

Existen varios comandos en R que permiten acceder a los diferentes elementos de la estructura de una función. Los más utilizados son los siguientes:

formals	*Permite acceder a los argumentos de una función*
alist	*Permite la modificación de los argumentos de una función*
body	*Permite el acceso al cuerpo de una función*
edit	*Permite modificar funciones*
mapply(f,vec1,..,vecn	*Aplica la función f a los vectores especificados*

© Alfaomega - RC Libros

Como primer ejemplo vamos a consultar los argumentos de la función *rootquad* definida anteriormente.

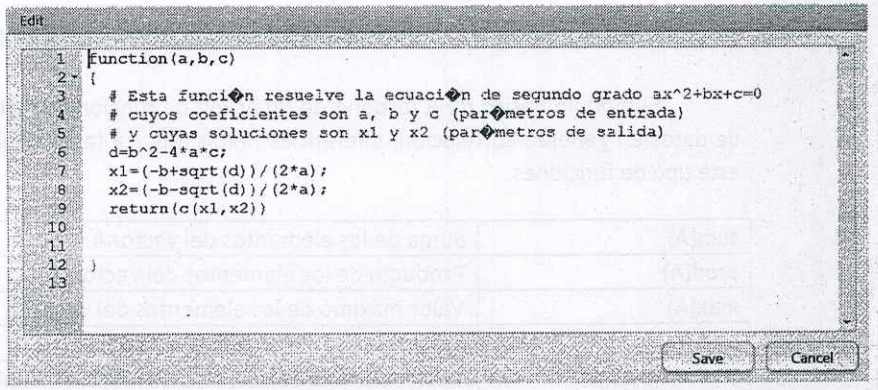
```
> formals(rootquad)
$a

$b

$c
```

Ahora accederemos al cuerpo de la función.

```
> body(rootquad)
{
    d = b^2 - 4 * a * c
    x1 = (-b + sqrt(d))/(2 * a)
    x2 = (-b - sqrt(d))/(2 * a)
    return(c(x1, x2))
}
```

Para modificar cualquier parte de la función utilizamos *edit*:

```
> edit(rootquad)
```



Podemos utilizar *mapply* para resolver la ecuación $x^2+3x+2=0$.

```
> mapply(rootquad,1,3,2)
     [,1]
[1,]   -1
```


© Alfaomega - RC Libros


```
[2,]   -2

> rootquad(1,3,2)
[1] -1 -2
```

FUNCIONES MATEMÁTICAS PREDEFINIDAS EN R

Aparte de los comandos generales de R, que a veces también se denominan funciones, el programa implementa diversos tipos de funciones estrictamente matemáticas. Algunas de ellas ya fueron estudiadas en capítulos anateriores y las restantes serán tratadas en ete capítulo y en capítulos posteriores. Para pedir ayuda al programa sobre el uso de cualquiera de las funciones, basta con teclear *"? nombre_de_función"* en la ventana de comandos de R. También se puede usar la ayuda general del programa (opción *Help* del menú general de la ventana de comandos).

Entre las funciones matemáticas de Matlab, destacan las funciones especializadas de variable real y las funciones estadísticas que se estudian en los párrafos siguientes:

Funciones matemáticas elementales de variable real

El grupo de funciones elementales del módulo básico de R que se presentan en la tabla siguiente ya fueron estudiadas en el capítulo relativo a la definición de variables (tanto escalares, como vectoriales y como matriciales). No obstante, solo a modo de resumen, se vuelven a presentar aquí.

abs	*Módulo o valor absoluto*
sin, sinh	*Seno y seno hiperbólico*
cos, cosh	*Coseno y coseno hiperbólico*
tan, tanh	*Tangente y tangente hiperbólica*
asin, asinh	*Arco seno y arco seno hiperbólico*
acos, acosh	*Arco coseno y arco coseno hiperbólico*
atan, atanh	*Arco tangente y arco tangente hiperbólico*
atan2	*Arco tangente en el cuarto cuadrante*
ceiling	*Redondeo al entero anterior*
floor	*Redondeo al entero siguiente (parte entera)*
round	*Redondeo al entero más próximo*
round(*x*, digits=*n*)	*Redondeo a n dígitos*
signif(*x*, digits=*n*)	*Redondeo a n dígitos significativos*

© Alfaomega - RC Libros

exp	*Exponencial*
expm1(x)	*exp(x)-1*
log	*Logaritmo natural (neperiano)*
log2	*Logaritmo en base 2*
log10	*Logaritmo decimal*
log1p(x)	*Logaritmo de 1+x*
round	*Redondeo al entero más cercano*
sign(x)	*Signo de x (1 si x es positivo y 0 si x es negativo)*
choose(a,b)	*Número combinatorio a sobre b con a y b enteros*
lchoose(a,b)	*Logaritmo del número combinatorio a sobre b*
factorial(x)	*Factorial del entero x*
lfactorial(x)	*Logaritmo del factorial de x*
sqrt	*Raíz cuadrada*

Como primer ejemplo calculamos las combinaciones sin repetición de 10 elementos tomados de 4 en 4.

```
> choose(10,4)
[1] 210
```

A continuación se calcula el seno y el coseno de los ángulos desde 0 a 2π de π/2 en π/2.

```
> sin(0:pi/2:2*pi)
[1]  0.000000e+00  1.000000e+00  1.224606e-16 -1.000000e+00

> round(sin(0:pi/2:2*pi))
[1]  0  1  0 -1

> round(cos(0:pi/2:2*pi))
[1]  1  0 -1  0
```

A continuación se prueban algunas propiedades de las funciones exponencial y logarítmica.

```
> exp(log(2))
[1] 2

> log(exp(2))
[1] 2
```


© Alfaomega - RC Libros


```
> log2(2^1000)
[1] 1000

> log(7^70,7)
[1] 70
```

A continuación se prueban algunas propiedades de las funciones trigonométricas e hiperbólicas.

```
> sin(pi)^2+cos(pi)^2
[1] 1
> (exp(5)+exp(-5))/2
[1] 74.20995

> cosh(5)
[1] 74.20995

> cosh(5)^2-sinh(5)^2
[1] 1


> 1+tan(pi/4)^2
[1] 2

> 1/sin(pi/4)^2
[1] 2
```

Funciones de análisis estadístico básico con argumento matricial

El módulo básico de R dispone de un grupo de funciones relativas al análisis de datos en general, correlación, diferencias finitas, etc. La tabla siguiente presenta este tipo de funciones.

sum(A)	Suma de los elementos del vector A
prod(A)	Producto de los elementos del vector A
max(A)	Valor máximo de los elementos del vector A
min(A)	Valor mínimo de los elementos del vector A
pmax(A,B,C..)	Vector de valores máximos de A,B,C.. (max(A[i],B[i],C[i],...)
pmin(A,B,C...)	Vector de valores mínimos de A,B,C.. (min(A[i],B[i],C[i],...)
cumprod(A)	Vector de productos acumulados de los elementos del vector A

© Alfaomega - RC Libros

cumsum(A)	Vector de sumas acumuladas de los elementos del vector A
cummax(A)	Vector de máximos acumulados de los elementos del vector A
cummin(A)	Vector de mínimos acumulados de los elementos del vector A
range(A)	Valores mínimo y máximo de los elementos del vector A range(A) = c(min(A), max(A))
rank(A)	Vector con los rangos de los elementos del vector A (lugar que ocupa cada elemento en A)
mean(A)	Media aritmética de las columnas de la matriz A
weighted.mean(A,W)	Media aritmética ponderada de los elementos de A con vector de ponderaciones W
median(A)	Mediana de los elementos del vector A
quantile(A,probs=0.k)	k-ésimo cuantil de los elementos del vector A
IQR(A)	Rango intercuartílico de los elementos del vector A
fivenum(A)	Mínimo, primer cuartil, mediana, tercer cuartil y máximo de los elementos del vector A
sd(A)	Cuasidesviación típica de los elementos del vector A
var(A)	Cuasivarianza de los elementos del vector A
cov(A)	Cuasivarianza de los elementos del vector A
var(A,B)	Covarianza entre los vectores A y B
cov(A,B)	Covarianza entre los vectores A y B
cor(A,B)	Coeficiente de correlación entre los vectores A y B
var(M)	Matriz de covarianzas para las columnas de las matriz M
cov(M)	Matriz de covarianzas para las columnas de las matriz M
cor(M)	Matriz de corelaciones para las columnas de las matriz M
diff(A)	Genera el vector primera diferencia del vector A
diff(A,lag=k)	Genera el vector k-ésima diferencia del vector A
scale(*x*, center=TRUE, scale=TRUE)	Estandariza el vector o matriz x

Como ejemplo, partimos de la variable formada por los 9 primeros enteros positivos y calculamos las variables de sumas acumuladas, productos acumulados, máximos acumulados, mínimos acumulados, primeras y segundas diferencias, máximo y mínimo, rangos y la variable estandarizada.

© Alfaomega - RC Libros


```
> A=c(1, 2, 3, 4, 5, 6, 7, 8, 9)
> A
[1] 1 2 3 4 5 6 7 8 9
> cumprod(A)
[1]      1      2      6     24    120    720   5040  40320
362880
> cumsum(A)
[1]  1  3  6 10 15 21 28 36 45
> cummax(A)
[1] 1 2 3 4 5 6 7 8 9
> cummin(A)
[1] 1 1 1 1 1 1 1 1 1
> diff(A)
[1] 1 1 1 1 1 1 1 1
> diff(A,lag=2)
[1] 2 2 2 2 2 2 2
> range(A)
[1] 1 9
> rank(A)
[1] 1 2 3 4 5 6 7 8 9
> scale(A, center=TRUE, scale=TRUE)
            [,1]
 [1,] -1.4605935
 [2,] -1.0954451
 [3,] -0.7302967
 [4,] -0.3651484
 [5,]  0.0000000
 [6,]  0.3651484
 [7,]  0.7302967
 [8,]  1.0954451
 [9,]  1.4605935
attr(,"scaled:center")
[1] 5
attr(,"scaled:scale")
[1] 2.738613
```

Calcularemos ahora el producto, la suma, el máximo, el mínimo, la media, la mediana, la cuasivarianza, la cuasidesviación típica, el percentil 30, los percentiles de órdenes 1, 25, 50, 75 y 100 y el rango intercuartílico de la variable A.

```
> prod(A)
[1] 362880
> sum(A)
[1] 45
> max(A)
```

© Alfaomega - RC Libros

```
[1] 9
> min(A)
[1] 1
> mean(A)
[1] 5
> median(A)
[1] 5
> var(A)
[1] 7.5
> sd(A)
[1] 2.738613

> quantile(A,probs=0.3)
30%
3.4
> fivenum(A)
[1] 1 3 5 7 9
> IQR(A)
[1] 4
```

A continuación calculamos la varianza y la desviación típica de la variable A.

```
> Varianza=(length(A)-1)/length(A)*var(A)
> Varianza
[1] 6.666667

> Desviaciontipica=sqrt((length(A)-1)/length(A)*var(A))
> Desviaciontipica
[1] 2.581989
```

A continuación calculamos el coeficiente de correlación y la covarianza entre las variables A y Sen(A).

```
> cor(A,sin(A))
[1] 0.001990017

> cov(A,sin(A))
[1] 0.00398658
```

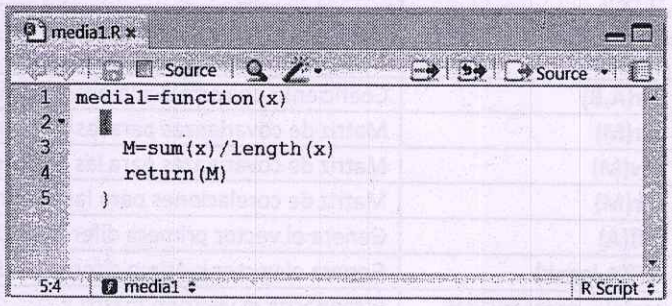
Por último consideramos la matriz cuadrada M de orden 3 cuyas columnas son tres variables con números enteros positivos consecutivos. Calcularemos la matriz de covarianzas y la matriz de correlaciones de las tres variables.

© Alfaomega - RC Libros


```
> M=matrix(c(1,2,3,4,5,6,7,8,9),3,3)

> M
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> cov(M)
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    1
[3,]    1    1    1
> cor(M)
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    1
[3,]    1    1    1
```

Ejercicio 4-1. Construir dos funciones a medida que permitan hallar la media y la varianza de cualquier variable.

Una sintaxis muy sencilla para programar la media sería la siguiente:



Para hallar la media de los 9 primeros números naturales no nulos se introduciría el siguiente comando:

```
> media1(c(1,2,3,4,5,6,7,8,9))
[1] 5
```

Una sintaxis muy sencilla para programar la varianza sería la siguiente:

© Alfaomega - RC Libros

```
Varianza1=function(x)
  {
    n=length(x)
    v=sum((x-(sum(x)/n))^2)/n
    return(v)
  }
```

Para hallar la varianza de los 9 primeros números naturales no nulos se introduciría el siguiente comando:

```
> Varianza1(c(1,2,3,4,5,6,7,8,9))
[1] 6.666667
```

Ejercicio 4-2. Construir una función a medida que permitan hallar la covarianza entre dos variables salvando el problema de los valores desaparecidos en las variables.

La sintaxis podría ser la siguiente:

```
Covarianza<-function(x,y)
  {
    x=x[!is.na(x)]
    y=y[!is.na(y)]
    n=length(x)
    cv<-sum((x-(sum(x)/n))^2*(y-(sum(y)/n))^2)/n
    return(cv)
    }
```

Calcularíamos la covarianza de dos variables como sigue:

```
> Covarianza(c(1,2,3,4),c(14,12,11,10))
[1] 4.609375
```


© Alfaomega - RC Libros

Ejercicio 4-3. El número de pétalos de 13 flores de una determinada especie es el siguiente: 8, 10, 6, 5, 8, 11, 8, 10, 7, 10, 7, 10 y 9. Calcular la media, la varianza y el coeficiente de variación.

Definimos el vector V cuyos componentes son los números de pétalos y a continuación calculamos la media varianza y coeficiente de variación con las funciones adecuadas de R.

```
> V=c(8, 10, 6, 5, 8, 11, 8, 10, 7, 10, 7, 10, 9)

> mediaV=mean(V)
> mediaV
[1] 8.384615

> varianzaV=(length(V)-1)/length(V)*var(V)
> varianzaV
[1] 3.005917

> coeficientevariacionV=sqrt(varianzaV)/media(V)
> coeficientevariacionV
[1] 0.2067785
```

El coeficiente de variación se ha calculado como cociente entre la desviación típica y la media:

Ejercicio 4-4. Dada la variable NP cuyos datos son 8, 10, 6, 5, 8, 11, 8, 10, 7, 10, 7, 10 y 9, obtener sus diferencias sucesivas de órdenes 1 y 2.

Comenzamos definiendo un vector NP que contiene los valores dados y a continuación utilizamos adecuadamente la función *diff* de MATLAB.

```
> NP=c(8, 10, 6, 5, 8, 11, 8, 10, 7, 10, 7, 10, 9)

> DS1=diff(NP)
> DS1
 [1]  2 -4 -1  3  3 -3  2 -3  3 -3  3 -1

> DS2=diff(NP,lag=2)
> DS2
 [1] -2 -5  2  6  0 -1 -1  0  0  0  2
```

© Alfaomega - RC Libros

Ejercicio 4-5. Un agente de seguros vende pólizas a 5 individuos, todos de la misma edad. De acuerdo con las tablas actuariales, la probabilidad de que un individuo con esa edad viva 30 años más es de 3/5. Determinar la probabilidad de que dentro de 30 años vivan:

a) al menos 3 individuos.
b) como mucho 2.

Como quiera que dentro de 30 años la situación de cada individuo será que viva o que no viva, y al menos una de las dos se ha de presentar, la situación para cada individuo se ajusta a una variable de Bernoulli con probabilidad de éxito (vivir 30 años más) igual a 3/5 = 0,6. Al considerar los 5 individuos, estamos ante una variable aleatoria X Binomial con $n = 5$ y $p = 0.6$, $X = B(5, 0{,}6)$. Si llamamos $F(x)$ a la función de distribución de $X = B(5, 0{,}6)$ en el punto x, los apartados del problema se calculan como sigue:

a) Habrá que calcular $P(X \geq 3)$, o lo que es lo mismo, $1\text{-}p(X < 3) = 1\text{-}F(2)$.
b) Habrá que calcular $P(X \leq 2)$, o lo que es lo mismo $F(2)$.

La primera probabilidad es equivalente a:

$$1-F(2)=1-\binom{5}{2}0{,}6^2 0{,}4^3-\binom{5}{1}0{,}6^1 0{,}4^4-\binom{5}{0}0{,}4^5$$

que puede calcularse mediante R como sigue:

```
> 1-choose(5,2)*0.6^2*0.4^3-choose(5,1)*0.6*0.4^4-
choose(5,0)*0.4^5
[1] 0.68256
```

La segunda probabilidad es equivalente a:

$$F(2)=\binom{5}{2}0{,}6^2 0{,}4^3+\binom{5}{1}0{,}6^1 0{,}4^4+\binom{5}{0}0{,}4^5$$

que puede calcularse mediante R como sigue:

```
> choose(5,2)*0.6^2*0.4^3+choose(5,1)*0.6*0.4^4+choose(5,0)*0.4^5
[1] 0.31744
```


© Alfaomega - RC Libros

Ejercicio 4-6. El número medio de automóviles que llega a una estación de suministro de gasolina es de 210 por hora. Si dicha estación puede atender a un máximo de 10 automóviles por minuto, determinar la probabilidad de que en un minuto dado lleguen a la estación de suministro más automóviles de los que puede atender.

El número aleatorio de automóviles que llegan a la estación de servicio en un minuto puede representarse por una variable X de Poisson de parámetro $m = 210/60 = 3{,}5$, ya que m es el número medio de llegadas por minuto (teníamos 210 llegadas a la hora).

La probabilidad que vamos a obtener vendrá dada por $P(X > 10)$, ya que para que lleguen a la estación más automóviles por minuto de los que se puedan atender es necesario que lleguen más de 10 por minuto. Pero $P(X > 10) = 1 - P(X \leq 10) = 1 - F(10)$ siendo F la función de distribución de una variable aleatoria de Poisson de parámetro 3,5.

Para calcular la probabilidad pedida tendremos en cuenta que:

$$1-F(10)=1-\sum_{k=0}^{10}\frac{e^{-3{,}5}3{,}5^k}{k!}$$

que puede calcularse en R como sigue:

```
> 1-exp(-3.5)*(1+3.5+3.5^2/factorial(2)+3.5^3/factorial(3)+
3.5^4/factorial(4)+3.5^5/factorial(5)+3.5^6/factorial(6)+
3.5^7/factorial(7)+3.5^8/factorial(8)+3.5^9/factorial(9)+
3.5^10/factorial(10))

[1] 0.001019394
```

Ejercicio 4-7. La proporción de individuos de una población con renta superior a los dos millones de unidades monetarias es de 0,005%. Determinar la probabilidad de que entre 5000 individuos consultados haya como mucho 2 con ese nivel de renta, supuesto que todos los consultados respondan.

Teóricamente, la variable X definida por el número de individuos, de entre los 5000, que tienen un nivel de renta superior a los dos millones de unidades monetarias es una variable binomial con $n = 5000$ y $p = 0{,}00005$. La probabilidad pedida sería $P(X \leq 2)$. Pero este cálculo con números tan extremos es molesto.

© Alfaomega - RC Libros

Ahora bien, habida cuenta de que p es muy pequeño (p = 0,00005) y de que n es muy grande (n = 5000), y además m = np = 0.25 < 5 y p < 0,1, ya podemos aproximar la variable X = B(5000, 0,00005) por una variable de Poisson de parámetro m = np = 5000* 0,00005 = 0,25. Luego P(X ≤ 2) lo podemos calcular mediante la distribución de Poisson de parámetro 0,25 como F(2), es decir:

$$F(2)=\sum_{k=0}^{2}\frac{e^{-0,25}0,25^{k}}{k!}$$

que puede calcularse en R como sigue:

```
> exp(-0.25)*(1+0.25+0.25^2/factorial(2))
[1] 0.9978385
```

Ejercicio 4-8. ¿Cuántos triángulos distintos pueden formarse con los vértices de un dodecágono?

Se trata de calcular el número de combinaciones de 12 elementos tomados de 3 en 3. Para ello se usa la siguiente sintaxis:

```
> a=choose(12,3)
> a
[1] 220
```

Ejercicio 4-9. Probar para n=100 y p=30 que: $\binom{n}{p}+\binom{n}{p+1}=\binom{n+1}{p+1}$

El problema puede resolverse comprobando la igualdad de ambos miembros mediante la siguiente sintaxis:

```
> a=choose(100,30)+choose(100,30+1)
> a
[1] 9.569698e+25

> b=choose(100+1,30+1)
> b
[1] 9.569698e+25
```


© Alfaomega - RC Libros

Ejercicio 4-10. Sean las variables X, Y y Z cuyos valores son los siguientes:

X	Y	Z
2	4	2
3	5	4
6	10	6
8	11	7
10	15	10

a) Calcular la media, la mediana, la desviación típica, la varianza y el coeficiente de variación para las tres variables, y hallar un intervalo de confianza para la media basado en cada variable con un nivel de confianza del 95%
b) Determínese la matriz de correlaciones, deduciendo de ella el grado de dependencia entre las variables
c) Determínese la matriz de covarianzas e interpretar el resultado

Comenzamos introduciendo las variables X, Y y Z como columnas de la matriz A.

```
> A=matrix(c(2,3,6,8,10,4,5,10,11,15,2,4,6,7,10),5,3)
> A
     [,1] [,2] [,3]
[1,]    2    4    2
[2,]    3    5    4
[3,]    6   10    6
[4,]    8   11    7
[5,]   10   15   10
```

Para calcular los estadísticos pedidos usamos la siguiente sintaxis R:

```
> mean(A)
[1] 6.866667
> mean(A[,1])
[1] 5.8
> mean(A[,2])
[1] 9
> mean(A[,3])
[1] 5.8

> median(A[,1])
[1] 6
> median(A[,2])
[1] 10
> median(A[,3])
[1] 6
```

© Alfaomega - RC Libros

```
> varianzaA1=4/5*var(A[,1])
> varianzaA1
[1] 8.96
> varianzaA2=4/5*var(A[,2])
> varianzaA2
[1] 16.4
> varianzaA3=4/5*var(A[,3])
> varianzaA3
[1] 7.36

> desvtipA1=sqrt(varianzaA1)
> desvtipA1
[1] 2.993326
> desvtipA2=sqrt(varianzaA2)
> desvtipA2
[1] 4.049691
> desvtipA3=sqrt(varianzaA3)
> desvtipA3
[1] 2.712932

> cvA1=desvtipA1/mean(A[,1])
> cvA1
[1] 0.5160907
> cvA2=desvtipA2/mean(A[,2])
> cvA2
[1] 0.4499657
> cvA3=desvtipA3/mean(A[,3])
> cvA3
[1] 0.4677469
```

Para calcular los intervalos de confianza usamos la sintaxis siguiente:

```
> c(mean(A[,1])-2*desvtipA1,mean(A[,1])+2*desvtipA1)
[1] -0.1866518 11.7866518

> c(mean(A[,2])-2*desvtipA2,mean(A[,2])+2*desvtipA2)
[1]  0.9006173 17.0993827

> c(mean(A[,3])-2*desvtipA3,mean(A[,3])+2*desvtipA3)
[1]  0.374136 11.225864
```

Los intervalos serán: [-0,187, 117867], [0,9006, 17,0994] y [0,3741, 11,2259]

© Alfaomega - RC Libros

Para hallar la matriz de correlaciones usamos la siguiente sintaxis R:

```
> cor(A)
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.9899319 0.9802123
[2,] 0.9899319 1.0000000 0.9830213
[3,] 0.9802123 0.9830213 1.0000000
```

En la salida anterior se observa que el coeficiente de correlación entre *X* e *Y* es 0,9899; entre *X* y *Z*, 0,9802; y entre Y y Z, 0,9830, lo que indica la fuerte dependencia lineal positiva entre cada par de variables.

Para hallar la matriz de covarianzas usamos la siguiente sintaxis R:

```
> cov(A)
      [,1] [,2]  [,3]
[1,] 11.20 15.0  9.95
[2,] 15.00 20.5 13.50
[3,]  9.95 13.5  9.20
```

A la vista de los resultados podemos decir que como todas las covarianzas son positivas, los pares de variables varían en el mismo sentido alrededor de sus medias.

Ejercicio 4-11. Para un determinado país, sean las variables R=Renta personal en unidades monetarias y H=Número de personas que se van de vacaciones al extranjero y cuyos valores son los siguientes:

R	*H*
1000	30
1100	50
1300	60
1500	70
1800	90
2000	100
2200	110
2500	120

Justificar si puede aceptarse una dependencia estadística lineal entre R y H.

© Alfaomega - RC Libros

Comenzamos introduciendo las variables R y H como vectores columna, y a continuación se calcula el coeficiente de correlación como sigue:

```
> R=c(1000,1100,1300,1500,1800,2000,2200,2500)
> H=c(30,50,60,70,90,100,110,120)
> cor(R,H)
[1] 0.986152
```

Se observa que el coeficiente de correlación entre R y H es r = 0,9862, lo que asegura la fuerte dependencia lineal positiva entre las dos variables.

© Alfaomega - RC Libros

CAPÍTULO 5

LENGUAJE DE R. CONTROL DE FLUJO, CONDICIONALES Y OTRAS ESTRUCTURAS

CONTROL DE FLUJO

El uso de funciones recursivas, condicionales y definidas a trozos es muy habitual en matemáticas. Para la definición de este tipo de funciones, es necesario el manejo de bucles. Como es natural, la definición de las funciones se hará a través de *ficheros* definidos en el editor de datos de R.

EL BUCLE FOR

R dispone de su propia versión de la sentencia FOR (definida en la sintaxis de la mayoría de los lenguajes de programación). Esta sentencia permite ejecutar de forma repetitiva un comando o grupo de comandos varias veces. Por ejemplo:

© Alfaomega - RC Libros

```
> for (i in -5:5) print(i^3)

[1] -125
[1] -64
[1] -27
[1] -8
[1] -1
[1] 0
[1] 1
[1] 8
[1] 27
[1] 64
[1] 125
```

La forma general de un bucle FOR para ejecutar una expresión repetitiva que depende de un índice *i* es la siguiente:

for(i in a:b)
{
expresión(i)
}

El bucle siempre empieza con la cláusula *for* e incluye en su interior todo un conjunto de comandos que dependen de uno o varios índices que definen el bucle.

El ejemplo anterior podría escribirse como un programa R como se indica en la Figura 5-1. El programa se ejecuta con el botón *Run* . También se puede guardar con el nombre for1.R. Posteriormente se puede cargar y ejecutar cuando se desee.

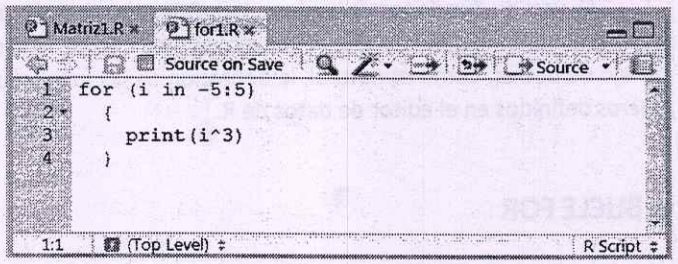


Figura 5-1

Los bucles for pueden anidarse. Veamos un ejemplo (Figura 5-2).

© Alfaomega - RC Libros

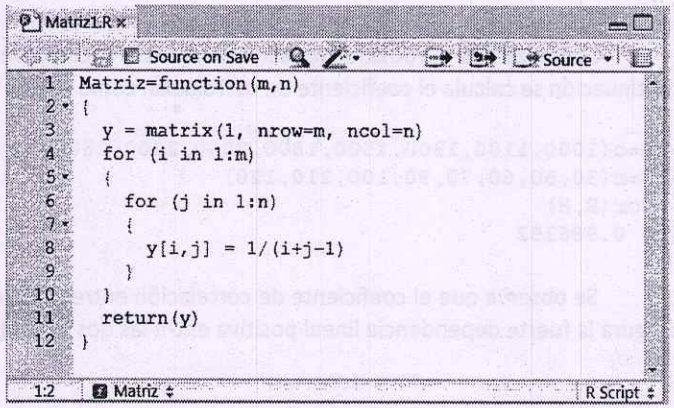


Figura 5-2

En este bucle hemos definido la matriz de Hilbert de orden *(m,n)* cuyo término general *(i,j)* es 1/(*i*+*j*-1). Si guardamos su contenido como un *fichero* de nombre *matriz1.R*, podremos construir cualquier matriz de Hilbert posteriormente cargando y ejecutando el *fichero* y especificando los valores para las variables *m* y *n* (dimensiones de la matriz) como se indica a continuación:

```
> Matriz(3,3)
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.5000000 0.3333333
[2,] 0.5000000 0.3333333 0.2500000
[3,] 0.3333333 0.2500000 0.2000000

> Matriz(4,5)
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.5000000 0.3333333 0.2500000 0.2000000
[2,] 0.5000000 0.3333333 0.2500000 0.2000000 0.1666667
[3,] 0.3333333 0.2500000 0.2000000 0.1666667 0.1428571
[4,] 0.2500000 0.2000000 0.1666667 0.1428571 0.1250000
```

EL BUCLE WHILE

R dispone de su propia versión de la sentencia WHILE definida en la sintaxis de la mayoría de los lenguajes de programación. Esta sentencia permite ejecutar de forma repetitiva un comando o grupo de comandos un número determinado de veces mientras se cumple una condición lógica especificada. Por ejemplo:

```
> i=1; while(i<=5) {print(i^3); i=i+1}
[1] 1
```

© Alfaomega - RC Libros

```
[1] 8
[1] 27
[1] 64
[1] 125
```

La sintaxis general de este bucle es la siguiente:

while(condición)
{
comandos
}

El bucle siempre empieza con la cláusula *while* seguida de una condición, e incluye en su interior todo un conjunto de comandos que se ejecutan mientras se cumple la condición. Si algún comando define una variable, se finaliza con punto y coma para evitar repeticiones en la salida.

Como ejemplo, escribimos un programa (Figura 5-3) que se guarda con el nombre *while1.m*, cuya ejecución permite calcular el mayor número cuyo factorial no excede a 10^{100}.

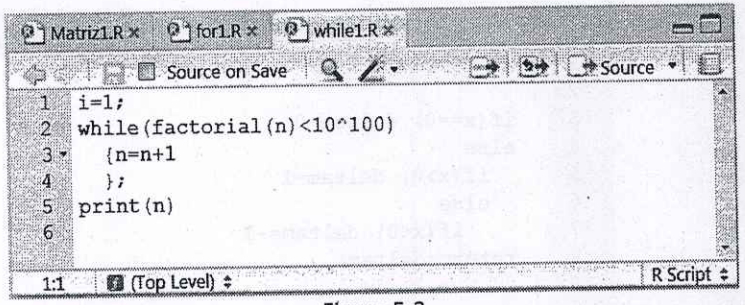


Figura 5-3

Ahora ejecutamos la sintaxis con el botón *Run*.

```
> print(n)
[1] 70
```

Podríamos haber escrito el programa como una única línea en la ventana de comandos de R.

```
> i=1; while(factorial(i)<10^100) {i=i+1}; print(i)
[1] 70
```


© Alfaomega - RC Libros

LA ESTRUCTURA CONDICIONAL IF ELSE

R, al igual que la mayoría de los lenguajes de programación estructurada, también incorpora la estructura IF-ELSE. Mediante esta estructura, se pueden ejecutar secuencias de comandos si se cumplen determinadas condiciones.

La sintaxis más sencilla del bucle es la siguiente:

{ if(condición)
comandos
}

Como ejemplo vamos a definir la función factorial mediante un bucle condicional.

La Figura 5-4 muestra una sintaxis sencilla.

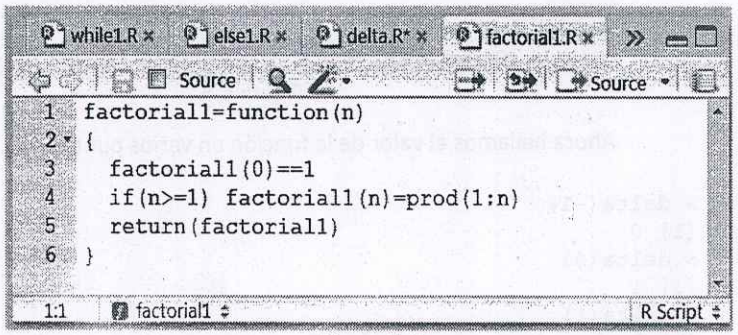


Figura 5-4

A continuación ejecutamos la función factorial para diversos valores.

```
> factorial(6)
[1] 720
> factorial(0)
[1] 1
> factorial(1)
[1] 1
> factorial(3)
[1] 6
```

En el caso anterior se ejecutan los comandos si la condición es cierta. Pero la sintaxis de este bucle puede ser más general.

© Alfaomega - RC Libros

{if condición
comandos1
else
comandos2
}

En este caso se ejecutan los *comandos1* si la condición es cierta, y se ejecutan los *comandos2* si la condición es falsa.

Como ejemplo vamos a definir la función delta de Kronecker que vale 1 si *x*=0 y vale 0 si *x* es distinto de cero. La Figura 5-5 muestra la sintaxis.

```
delta=function(x)
{
  if(x==0) delta=1
  else delta=0
  return(delta)
}
```

Figura 5-5

Ahora hallamos el valor de la función en varios puntos.

```
> delta(-1)
[1] 0
> delta(0)
[1] 1
> delta(1)
[1] 0
> delta(2)
[1] 0
> delta(-100)
[1] 0
```

Las sentencias IF, al igual que las sentencias FOR, pueden ser anidadas. Cuando se anidan varias sentencias IF, se utiliza la siguiente sintaxis:

{if(condición1)
comandos1
else

© Alfaomega - RC Libros

if(condición2)
comandos2
else
if(condición3)
comandos3
else

.
.
{

En este caso, se ejecutan los *comandos1* si *condición1* es cierta, se ejecutan los *comandos2* si *condición1* es falsa y *condición2* es cierta, se ejecutan los *comandos3* si *condición1* y *condición2* son falsas y *condición3* es cierta, y así sucesivamente.

Como ejemplo vamos a construir la función delta de Kronecker modificada que vale 0 si *x*=0, 1 si *x*>0 y -1 si *x*<0.

La Figura 5-6 muestra una sintaxis para este problema lo más sencilla posible.

```
deltam=function(x)
{
  if(x==0) deltam=0
  else
    if(x>0) deltam=1
    else
      if(x<0) deltam=-1
  return(deltam)
}
```

Figura 5-6

A continuación se ejecuta la función para diversos valores de su argumento:

```
> delta(-5)
[1] 0
> deltam(-5)
[1] -1
> deltam(-1)
```

© Alfaomega - RC Libros

```
[1] -1
> deltam(0)
[1] 0
> deltam(1)
[1] 1
> deltam(5)
[1] 1
```

LA ESTRUCTURA CONDICIONAL IFELSE

Al igual que en el caso de IF ELSE, la estructura IFELSE permite ejecutar secuencias de comandos si se cumplen una determinada condición y otra secuencia de comandos diferente si no se cumple la condición anterior.

La sintaxis más sencilla de la estructura IFELSE es la siguiente:

{ ifelse(condición, comandos1,comandos2)
}

Si se cumple la condición, se ejecuta *comandos1* y si no se cumple, se ejecuta *comandos2*. Como ejemplo vamos a definir la función factorial mediante un bucle condicional.

La Figura 5-7 muestra una sintaxis sencilla.

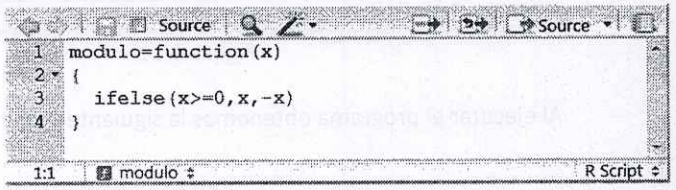


Figura 5-7

A continuación se ejecuta la función para diversos valores de su argumento:

```
> modulo(-5)
[1] 5
> modulo(-4)
[1] 4
> modulo(0)
[1] 0
> modulo(1)
[1] 1
> modulo(2)
[1] 2
```


© Alfaomega - RC Libros

LA ESTRUCTURA SWITCH

La estructura *switch* ejecuta ciertas sentencias basadas en el valor de una expresión. Su sintaxis básica es la siguiente:

{
switch (expresión, valor1=acción1, valor2=acción2,.., valork=acción)
}

Si la expresión vale *"valor1"* se ejecuta *acción1*, si vale *"valor2"* se ejecuta *accion2* y así sucesivamente. A continuación se presenta un ejemplo de una función que devuelve media o mediana de un vector (Figura 5-8).

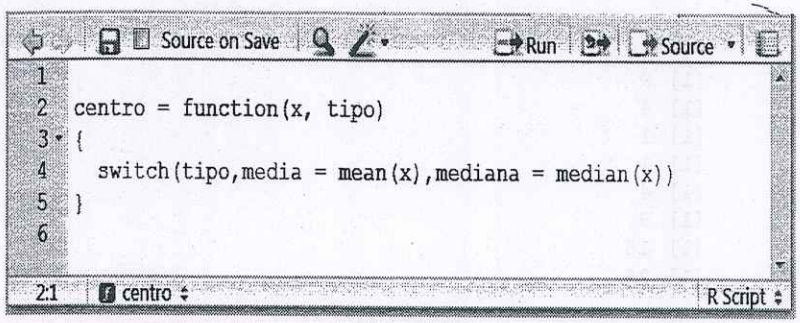


Figura 5-8

A continuación ejecutamos el ejemplo anterior.

```
> centro(c(10,11,9,8,6,1,2,3), "media")
[1] 6.25
> centro(c(10,11,9,8,6,1,2,3), "mediana")
[1] 7
```

NEXT

La instrucción *next* pasa el control a la iteración siguiente en un bucle *for* o *while* en el cual aparece, ignorando las restantes instrucciones en el cuerpo del bucle. En la Figura 5-9 se muestra un ejemplo.

© Alfaomega - RC Libros

```
for(i in -5:5)
{
  if(i==0) next;
  print(i^2)
}
```

Figura 5-9

A continuación ejecutamos el programa con el botón *Run*. La salida es la siguiente:

```
[1] 25
[1] 16
[1] 9
[1] 4
[1] 1
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
```

BREAK

La instrucción *break* finaliza la ejecución de un bucle *for* o *while* en el cual aparece continuando la ejecución en la siguiente instrucción fuera del bucle. A continuación se muestra un ejemplo (Figura 5-10).

```
for(i in -5:5)
{
  if(i==0) break;
  print(i^2)
}
```

Figura 5-10

© Alfaomega - RC Libros

A continuación ejecutamos el programa con *Run*. Se obtiene la siguiente salida:

```
[1] 25
[1] 16
[1] 9
[1] 4
[1] 1
```

LA ESTRUCTURA REPEAT

La sentencia REPEAT permite ejecutar repetidamente una expresión mientras se cumple una condición. Suele finalizar con una instrucción *break* para finalizar el proceso de repetición. En la Figura 5-11 se muestra un programa que imprime los cuadrados de los números naturales finalizando en el 9.

```
repeat{print(i^2);
  i=i+1;
  if(i==10) break
}
```

Figura 5-11

Al ejecutar el programa obtenemos la siguiente salida:

```
[1] 0
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
[1] 49
[1] 64
[1] 81
```

© Alfaomega - RC Libros

ALGORITMOS DE CÁLCULO NUMÉRICO

Los métodos de programación de R permiten abordar las técnicas de cálculo numérico mediante la implementación sencilla de los algoritmos teóricos que resuelven una casuística importante de aproximación a soluciones de problemas. Entre estos problemas juega un papel primordial la resolución de ecuaciones lineales y, sobre todo, no lineales.

Es sencillo construir programas en el lenguaje de R que desarrollen los algoritmos y que se guardan en ficheros. Ya sabemos de capítulos anteriores que estos ficheros contienen código R que simplemente ejecuta una serie de comandos o funciones que aceptan argumentos y producen una salida. Los ficheros se crean utilizando el editor de texto, tal y como ya se ha visto.

MÉTODO DE NEWTON PARA RESOLVER LA ECUACIÓN g(x)=0

El método de Newton resuelve la ecuación $f(x) = 0$, bajo determinadas condiciones exigidas a f, mediante la iteración $x_{r+1} = x_r - f(x_r)/f'(x_r)$ para un valor inicial dado x_0 lo suficientemente próximo a una solución.

Podemos elaborar el programa de la Figura 5-12 para construir el programa que resuelva ecuaciones por el método de Newton con una determinada precisión.

```
1
2  newton=function(func,dfunc,x,precis)
3  {
4    it=0
5    x0=x
6    d=func(x0)/dfunc(x0)
7    while(abs(d)>precis)
8    {
9      x1=x0-d
10     it=it+1
11     x0=x1
12     d=func(x0)/dfunc(x0)
13   }
14   return(x)
15 }
```

Figura 5-12

© Alfaomega - RC Libros

Como ejemplo resolvemos por el método de Newton la ecuación del apartado anterior:

$$x^2 - x - Sen(x + 0{,}15) = 0$$

Definimos previamente la función $f(x)$ como la función *f1* de la Figura 5-13, y su función derivada $f'(x)$ como la función *derf1* de la Figura 5-14.

```
1  f1=function(x)
2  {
3    f1=x^2-x-sin(x+0.15)
4  }
```

Figura 5-13

```
1  derf1=function(x)
2  {
3    derf1=2*x-1-cos(x+0.15)
4  }
```

Figura 5-14

Podemos ya resolver la ecuación con precisiones de una diezmilésima y una millonésima mediante la sintaxis R siguiente:

```
> x=newton(f1,derf1,1.5,0.0001)
> x
[1] 1.610069

> x=newton(f1,derf1,1.5,0.000001)
> x
[1] 1.610023
```

Se ha obtenido como solución x=1,61 en 2 iteraciones para la precisión de una diezmilésima y en 3 iteraciones para la precisión de una millonésima.

© Alfaomega - RC Libros

MÉTODO DE SCHRODER PARA RESOLVER LA ECUACIÓN g(x)=0

El método de Schroder, que es semejante al de Newton, resuelve la ecuación $f(x)=0$, bajo determinadas condiciones exigidas a f, mediante la iteración $x_{r+1} = x_r - m\, f(x_r)/f'(x_r)$ para un valor inicial dado x_0 lo suficientemente próximo a una solución, siendo m el orden de multiplicidad de la raíz buscada.

Podemos elaborar el programa de la Figura 5-15 para resolver ecuaciones por el método de Schroder con una determinada precisión.

```
schroder=function(func,dfunc,m,x,precis)
{
  it=0
  x0=x
  d=func(x0)/dfunc(x0)
  while(abs(d)>precis)
  {
    x1=x0-m*d
    it=it+1
    x0=x1
    d=func(x0)/dfunc(x0)
  }
  return(x)
}
```

Figura 5-15

Resolvemos ahora la misma ecuación ya resuelta por el método de Newton, pero buscando una solución de multiplicidad 3.

```
> x=schroder(f1,derf1,3,1.5,0.000001)
> x
[1] 1.5
```

MÉTODO DEL TRAPECIO PARA LA INTEGRACIÓN NUMÉRICA

Dada la dificultad de obtener primitivas exactas para muchas funciones, los métodos de integración numérica adquieren una importancia relevante. Existen

© Alfaomega - RC Libros

diferentes métodos para obtener resultados numéricos aproximados para integrales definidas, entre los que destacan el método del trapecio y el método de Simpson.

El método del trapecio para integración numérica tiene dos variantes: la regla compuesta del trapecio y la regla recursiva del trapecio.

La *regla compuesta del trapecio* aproxima la integral definida de la función $f(x)$ entre los puntos a y b de la siguiente forma:

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(a)+f(b)) + h\sum_{k=1}^{M-1} f(x_k)$$

calculándose $f(x)$ en los puntos equidistantes $x_k=a+kh$ $k=0,1,\ldots,M$ con $x_0=a$ y $x_M=b$.

La regla compuesta del trapecio puede implementarse mediante el programa de la Figura 5-16.

```
traprl=function(f,a,b,M)
{
  h=(b-a)/M
  s=0

  for(k in 1:(M-1))
  {
    x=a+h*k
    s=s+f(x)
  }

  s=h*(f(a)+f(b))/2+h*s
}
```

Figura 5-16

Como ejemplo calculamos mediante la regla recursiva del trapecio la integral:

$$\int_0^2 \frac{1}{x^2+\frac{1}{10}}dx$$

Comenzamos definiendo la función del integrando mediante la función *integrando1* de la Figura 5-17.

© Alfaomega - RC Libros

```
integrando1=function(x)
{
  integrando1=1/(x^2+1/10)
}
```

Figura 5-17

A continuación calculamos la integral pedida, para 14 rectángulos en el método del trapecio, se calcula como sigue:

```
> x=traprl(integrando1,0,2,14)
> x
[1] 4.471004
```

La *regla recursiva del trapecio* considera que los puntos $x_k=a+kh$ $k=0,1,\ldots,M$ con $x_0=a$ y $x_M=b$ dividen el intervalo $[a,b]$ en $2^J=2M$ subintervalos del mismo tamaño $h=(b-a)/2^J$. En esta situación se considera la siguiente fórmula recursiva:

$$T(0)=\frac{h}{2}(f(a)+f(b))$$

$$T(J)=\frac{T(J-1)}{2}+h\sum_{k=1}^{M}f(x_{2k-1}) \quad J=1,2,\cdots$$

y la integral definida de la función $f(x)$ entre los puntos a y b puede calcularse de la siguiente forma:

$$\int_a^b f(x)dx \approx \frac{h}{2}\sum_{k=1}^{2^J}(f(x_k)+f(x_{k-1}))$$

utilizando la recursión del trapecio según se aumenta el número de subintervalos de $[a,b]$ y tomando en la iteración J un conjunto de 2^J+1 puntos equidistantes.

La regla recursiva del trapecio puede implementarse mediante el programa de la Figura 5-18.

© Alfaomega - RC Libros


```
rctrap=function(f,a,b,n)
{
  M=1
  h=b-a
  T=matrix(0,1,n+1);
  T[1]=h*(f(a)+f(b))/2;
    {
      for(j in 1:n)
      M=2*M
      h=h/2
      s=0
    }
      {
        for(k in 1:M/2)
        x=a+h*(2*k-1)
        s=s+f(x)
      }

    T[j+1]=T[j]/2+h*s;
}
```

Figura 5-18

MÉTODO DE SIMPSON PARA LA INTEGRACIÓN NUMÉRICA

En el método de Simpson para integración numérica suelen considerarse dos variantes: la regla simple de Simpson y la regla compuesta de Simpson.

La *regla simple de Simpson* aproxima la integral definida de la función $f(x)$ entre los puntos a y b de la siguiente forma:

$$\int_a^b f(x)dx \approx \frac{h}{3}(f(a)+f(b)+4f(c)) \quad c=\frac{a+b}{2}$$

La regla simple de Simpson puede implementarse mediante el programa de la Figura 5-19.

© Alfaomega - RC Libros

```
simpsons=function(f,a0,b0,tolerancia0)
{
  h=(b0-a0)/2
  C=matrix(0,1,3)
  C=f(c(a0, (a0+b0)/2, b0))
  S=h*(C[1]+4*C[2]+C[3])/3
  S2=S
  tolerancia1=tolerancia0
  err=tolerancia0
  Z=c(a0, b0, S, S2, err, tolerancia1)
}
```

Figura 5-19

A continuación calculamos la integral que venimos utilizando como ejemplo mediante la regla de Simpson simple usando la siguiente sintaxis:

```
> x=simpsons(integrando1,0,2,14)
> x
[1]  0.000000  2.000000  4.626755  4.626755
[5] 14.000000 14.000000
```

Vemos que el valor de la integral es 4,626755.

La *regla compuesta de Simpson* aproxima la integral definida de la función $f(x)$ entre los puntos a y b de la siguiente forma:

$$\int_a^b f(x)dx \approx \frac{h}{3}(f(a)+f(b))+\frac{2h}{3}\sum_{k=1}^{M-1} f(x_{2k})+\frac{4h}{3}\sum_{k=1}^{M} f(x_{2k-1})$$

calculándose $f(x)$ en los puntos equidistantes $x_k=a+kh$ $k=0,1,\ldots,2M$ con $x_0=a$ y $x_{2M}=b$.

La regla compuesta de Simpson puede implementarse mediante el programa de la Figura 5-20.

© Alfaomega - RC Libros


```
simprl=function(f,a,b,M)
{
  h=(b-a)/(2*M)
  s1=0
  s2=0
  {
    for(k in 1:M)
    x=a+h*(2*k-1)
    s1=s1+f(x)
  }
    {
      for(k in 1:(M-1))
      x=a+h*2*k
      s2=s2+f(x);
    }
    s=h*(f(a)+f(b)+4*s1+2*s2)/3;
}
```

Figura 5-20

Como ejemplo calculamos mediante la regla compuesta de Simpson utilizando 2M=4 subintervalos la integral:

$$\int_0^2 \frac{1}{x^2+\frac{1}{10}}dx$$

Utilizaremos la siguiente sintaxis:

```
> x=simprl(integrando1,0,2,2)
> x
[1] 2.294035
```

Ejercicio 5-1. Realizar un programa para calcular el factorial de un número utilizando una estructura condicional IF y un bucle FOR anidado.

© Alfaomega - RC Libros

La Figura 5-21 muestra una sintaxis adecuada para la función factorial.

```
factoriala=function(n)
{
  f=1
  if(n>1)
    for(i in 1:n)
      f=f*i
  return(f)
}
```

Figura 5-21

Vamos a aplicar la función a varios argumentos.

```
> factoriala(0)
[1] 1
> factoriala(1)
[1] 1
> factoriala(2)
[1] 2
> factoriala(3)
[1] 6
> factoriala(4)
[1] 24
> factoriala(20)
[1] 2.432902e+18
```

Ejercicio 5-2. Realizar un programa para calcular el factorial de un número utilizando un bucle WHILE.

La Figura 5-22 muestra una sintaxis adecuada para la función factorial.

© Alfaomega - RC Libros


```
factorialb=function(n)
{
  f=1
  while(n>0)
  {
    f=n*f
    n=n-1
  }
  return(f)
}
```

Figura 5-22

Vamos a aplicar la función a varios argumentos.

```
> factorialb(0)
[1] 1
> factorialb(1)
[1] 1
> factorialb(3)
[1] 6
> factorialb(10)
[1] 3628800
```

Ejercicio 5-3. Realizar un programa para calcular el factorial de un número utilizando una estructura condicional IF ELSE.

La Figura 5-23 muestra una sintaxis adecuada para la función factorial.

```
factorialc=function(n)
{
  if(n>1)
    f=n*factorialc(n-1)
  else f=1
  return(f)
}
```

Figura 5-23

Vamos a aplicar la función a varios argumentos.

© Alfaomega - RC Libros

```
> factorial(0)
[1] 1
> factorial(1)
[1] 1
> factorial(2)
[1] 2
> factorial(3)
[1] 6
```

Ejercicio 5-4. Realizar un programa para calcular el factorial de un número utilizando una estructura condicional IFELSE.

La Figura 5-24 muestra una sintaxis adecuada para la función factorial.

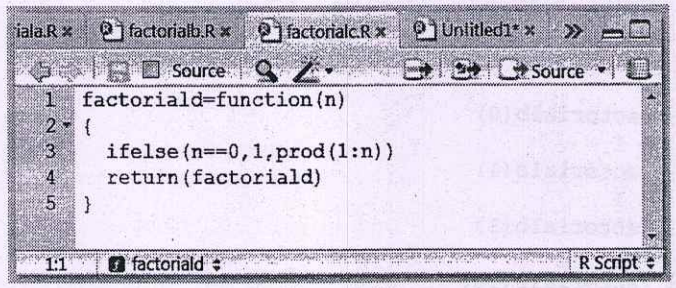


Figura 5-24

Vamos a aplicar la función a varios argumentos.

```
> factorial(0)
[1] 1
> factorial(1)
[1] 1
> factorial(2)
[1] 2
> factorial(3)
[1] 6
> factorial(20)
[1] 2.432902e+1
```

Ejercicio 5-5. Realizar un programa que genere los términos de una progresión aritmética.

© Alfaomega - RC Libros

En la Figura 5-25 se presenta la sintaxis que genera el término enésimo de una progresión aritmética cuyo primer término es *a1* y cuya razón (o diferencia) es *d*.

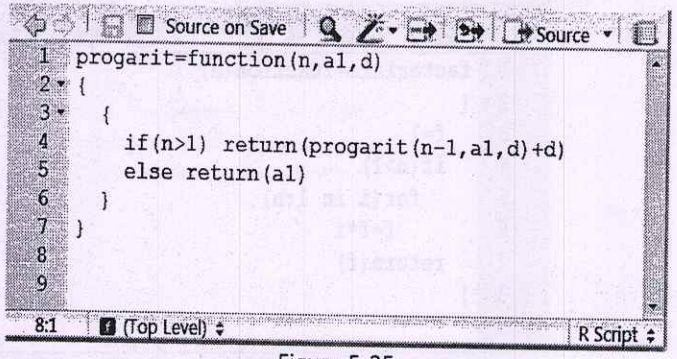


Figura 5-25

Para generar el octavo término de una progresión aritmética de razón 2 que comienza en el 3, se tecleará lo siguiente:

```
> progarit(8,3,2)
[1] 17
```

Si queremos un programa que devuelva todos los términos de la progresión hasta el enésimo, podemos utilizar la sintaxis de la Figura 5-26.

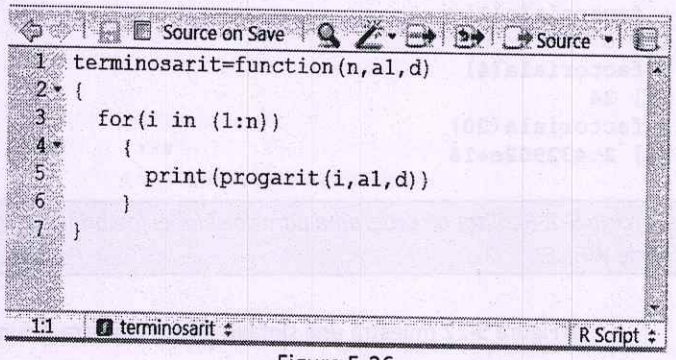


Figura 5-26

Ahora se presentan ejemplos de ejecución de la sintaxis anterior.

© Alfaomega - RC Libros

```
> terminosarit(8,3,2)
[1] 3
[1] 5
[1] 7
[1] 9
[1] 11
[1] 13
[1] 15
[1] 17

> terminosarit(5,100,50)
[1] 100
[1] 150
[1] 200
[1] 250
[1] 300
```

Los dos programas anteriores podrían escribirse en uno solo anidando las dos funciones, de modo que obtendríamos los *n* primeros términos de una progresión aritmética de razón dada y primer elemento dado con un solo programa.

La sintasis adecuada se presenta en la Figura 5-27.

```
terminosarit1=function(n,a1,d)
{
  {
    progarit=function(n,a1,d)
    {
      {
        if(n>1) return(progarit(n-1,a1,d)+d)
        else return(a1)
      }
    }
  }

  for(i in (1:n))
  {
    print(progarit(i,a1,d))
  }
}
```

Figura 5-27

Como ejemplo hallamos los 5 primeros términos de una progresión aritmética de razón 50, cuyo primer término es 100.

© Alfaomega - RC Libros


```
> terminosarit1(5,100,50)
[1] 100
[1] 150
[1] 200
[1] 250
[1] 300
```

Ejercicio 5-6. Realizar un programa que genere los términos de una progresión geométrica.

La Figura 5-28 muestra el programa que genera los *n* primeros términos de una progresión geomética de razón *r* que comienza en el término *a1*.

En esta ocasión se ha optado por realizar el programa directamente, utilizando una función anidada en otra función.

La posibilidad de anidamiento de funciones en R ofrece muchas posibilidades a la programación.

Posteriormente ejecutaremos el programa para hallar los 5 primeros términos de una progresión geométrica de razón 50, cuyo primer término es 100.

```
terminosgeom1=function(n,a1,r)
{
{
  proggeom=function(n,a1,r)
  {
{
  if(n>1) return(proggeom(n-1,a1,r)*r)
  else return(a1)
}
  }
}

for(i in (1:n))
{
  print(proggeom(i,a1,r))
}
}
```

Figura 5-28

© Alfaomega - RC Libros

```
> terminosgeom1(5,100,50)
[1] 100
[1] 5000
[1] 250000
[1] 12500000
[1] 6.25e+08
```

Ejercicio 5-7. La sucesión {an} de Fibonacci se define mediante la ley de recurrencia dada por $a_1 = 1$, $a_2 = 1$, $a_n = a_{n-1} + a_{n-2}$. Representar esta sucesión por una función recurrente y calcular a_2, a_5 y a_6.

Para representar los términos de la sucesión de Fibonacci mediante una función y recurrente, definiremos la función mediante la sintaxis de la Figura 5-29.

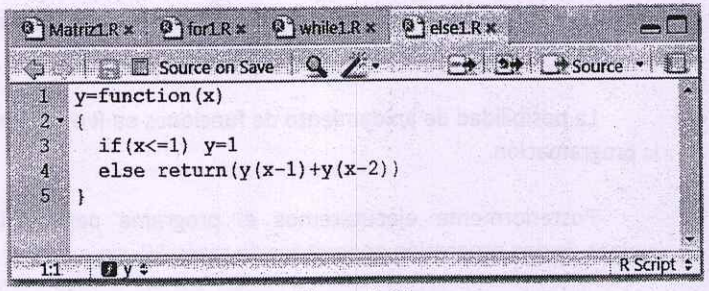


Figura 5-29

Los términos 2, 5 y 6 de la sucesión se calculan ahora mediante la sintaxis:

```
> y(2)
[1] 2
> y(5)
[1] 8
> y(6)
[1] 13
```

Ejercicio 5-8. Definir una función descriptiva(v) que devuelva la varianza y el coeficiente de variación de los elementos de un vector dado v. Como aplicación, hallar la varianza y el coeficiente de variación de los números 1, 5, 6, 7 y 9.

En la Figura 5-30 se define la sintaxis adecuada.

© Alfaomega - RC Libros

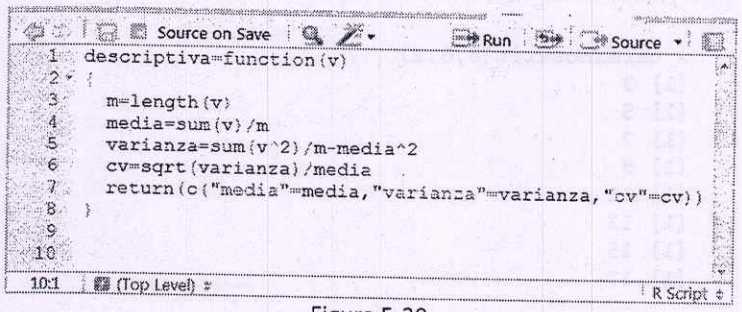


Figura 5-30

Para hallar la varianza y el coeficiente de variación del vector dado se utiliza la siguiente sintaxis:

```
> descriptiva(c(1,5,6,7,9))
    media  varianza        cv
5.6000000 7.0400000 0.4738035
```

Ejercicio 5-9. Resolver por el método de Newton con una determinada precisión y aplicarlo al cálculo de una raíz de la ecuación x3 -10x2 +29x - 20 = 0 cercana al punto x=7 con una precisión de 0,00005. Utilizar también la precisión 0,0005.

Definimos las funciones $f(x) = x^3 - 10x^2 + 29x - 20$ y su derivada a través de los ficheros de nombres *f302.R* y *f303.R* según las Figuras 5-31 y 5-32.

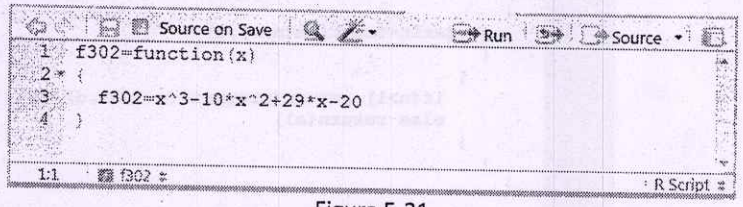


Figura 5-31

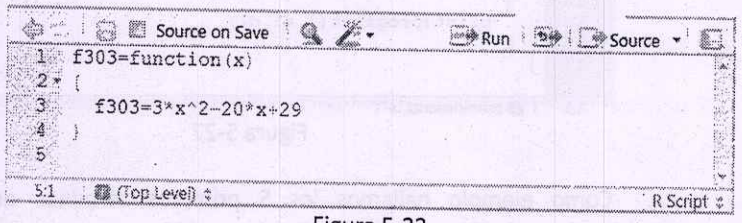


Figura 5-32

© Alfaomega - RC Libros

Para ejecutar el programa que resuelve la ecuación pedida tecleamos:

```
> print(newton(f302,f303,7,0.00005))
[1] 5
> print(newton(f302,f303,7,0.0005))
[1] 5.000212
```

Con una precisión de 0,00005 se ha obtenido la solución $x = 5$. Con una precisión de 0,0005 se obtiene la solución $x = 5{,}0002$.

Se observa que hemos utilizado el algoritmo de Newton previamente programado en este capítulo mediante la función de nombre *newton*.

Ejercicio 5-10. Aproximar la integral siguiente:

$$\int_{1}^{\frac{2\pi}{3}} Tan\left(Cos\left(\frac{\sqrt{5}+Sen(x)}{1+x^2}\right)\right)dx$$

Comenzamos definiendo la función a integrar mediante la sintaxis de la Figura 5-33.

```
funcion2=function(x)
{
  funcion2=tan(cos((sqrt(5)+sin(x))/(1+x^2)))
}
```

Figura 5-33

Podemos utilizar la regla de Simpson simple con 100 subintervalos mediante la sintaxis siguiente:

```
> print(simpsons(funcion2,1,2*pi/3,10))
[1]  1.000000  2.094395  0.685649  0.685649 10.000000
[6] 10.000000
```

Si utilizamos la regla del trapecio se obtiene el resultado siguiente:

```
> print(traprl(funcion2,1,2*pi/3,10))
[1] 0.6854009
```


© Alfaomega - RC Libros

Se observa que por ambos métodos se obtiene una solución muy similar x=0,685649 por el método de Simpson y x=0,6854009 por la regla del trapecio.

Hemos utilizado las funciones *simpsons* y *traprl* previamente definidas en este capítulo para resolver integrales aproximadas por los métodos de Simpson y la regla del trapecio, respectivamente.

Ejercicio 5-11. Realizar un programa que obtenga la transpuesta de cualquier matriz.

La Figura 5-34 muestra la sintaxis adecuada.

A continuación, para comprobar que el programa funciona correctamente, se considera la matriz cuadrada de orden 3 cuyas columnas son los 9 primeros números naturales positivos consecutivos.

Al observar los resultados que se muestran en la página siguiente, se comprueba que la transpuesta de la matriz considerada se ha calculado correctamente.

```
transpuesta = function(x)
{
  if (!is.matrix(x))
    {
    warning("el argumento no es una matriz: devolver NA")
    return(NA_real_)
  }
  y = matrix(1, nrow=ncol(x), ncol=nrow(x))
  for (i in 1:nrow(x))
    {
    for (j in 1:ncol(x))
      {
      y[j,i] = x[i,j]
    }
  }
  return(y)
}
```

Figura 5-34

© Alfaomega - RC Libros

```
> A=matrix(c(1,2,3,4,5,6,7,8,9),3,3)
> A
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> B=transpuesta(A)
> B
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```


© Alfaomega - RC Libros

CAPÍTULO 6

ANÁLISIS EXPLORATORIO DE DATOS A TRAVÉS DE R. EXPLORATORIO GRÁFICO

ANÁLISIS EXPLORATORIO DE DATOS

Antes de aplicar cualquier técnica de análisis de datos, es preciso realizar un análisis previo de la información de la que se dispone. Es necesario examinar las variables individuales y las relaciones entre ellas, así como evaluar y solucionar problemas en el diseño de la investigación y en la recogida de datos. La primera tarea que se suele abordar es el *análisis exploratorio y gráfico de los datos*. La mayoría del *software* estadístico dispone de herramientas que aportan técnicas gráficas preparadas para el examen de los datos que se ven mejoradas con medidas estadísticas más detalladas para su descripción. Estas técnicas permiten el examen de las características de la distribución de las variables implicadas en el análisis, las relaciones bivariantes (y multivariantes) entre ellas y el análisis de las diferencias entre grupos. Hay que tener presente que las representaciones gráficas nunca sustituyen a las medidas de diagnóstico formal estadístico (contrastes de ajuste de los datos a una distribución, contrastes de asimetría, contrastes de aletoriedad, etc.), pero proporcionan una forma alternativa de desarrollar una perspectiva del carácter de los datos y de las interrelaciones que existen, incluso si son multivariantes.

© Alfaomega - RC Libros

Las técnicas del análisis exploratorio de datos permiten analizar la información exhaustivamente y detectar las posibles anomalías que presentan las observaciones. J. W. Tuckey ha sido uno de los pioneros en la introducción de este tipo de análisis. Los estadísticos descriptivos más habitualmente utilizados han sido la media y la desviación típica. Sin embargo, el uso automático de estos índices no es muy aconsejable. La media y la desviación típica son índices convenientes solo cuando la distribución de datos es aproximadamente normal o, al menos, simétrica y unimodal. Pero las variables objeto de estudio no siempre cumplen estos requisitos. Por lo tanto es necesario un examen a fondo de la estructura de los datos.

Se recomienda iniciar un análisis exploratorio de datos con gráficos que permitan visualizar su estructura. Estamos ante las herramientas de ***exploración visual***. Sin embargo, para la ***exploración formal***, el uso de *estadísticos robustos* (o resistentes) es muy aconsejable cuando los datos no se ajustan a una distribución normal. Estos estadísticos son los que se ven poco afectados por valores atípicos. Suelen estar basados en la mediana y en los cuartiles y son de fácil cálculo. Fruto del análisis exploratorio, a veces es necesario realizar transformación de variables.

Para datos cuantitativos es aconsejable comenzar con el gráfico de tallo y hojas o histograma digital. El paso siguiente suele ser examinar la posible presencia de normalidad, simetría y valores atípicos (outliers) en el conjunto de datos. Para ello suelen utilizarse los gráficos de caja y bigote. No obstante, los gráficos de caja siempre deben ir acompañados de los histogramas digitales (o gráficos de tallo y hojas), ya que los primeros no detectan la presencia de distribuciones multimodales. Los gráficos de dispersión nos dan una idea de las relaciones entre variables y su ajuste.

Histograma de frecuencias

De todas formas, siempre es conveniente iniciar el análisis exploratorio de datos con la construcción del histograma de frecuencias asociado, para poder así intuir la distribución de probabilidad de los datos, su normalidad, su simetría y otras propiedades interesantes en el análisis de datos. Como ejemplo podemos considerar la variable X definida como el consumo de combustible en litros a los 1000 kilómetros de los automóviles de una determinada marca. Los valores para X son los siguientes:

© Alfaomega - RC Libros

43,1 36,1 32,8 39,4 36,1 19,9 19,4 20,2 19,2 20,5 20,2 25,1 20,5 19,4 20,6
20,8 18,6 18,1 19,2 17,7 18,1 17,5 30 27,5 27,2 30,9 21,1 23,2 23,8 23,9
20,3 17 21,6 16,2 31,5 29,5 21,5 19,8 22,3 20,2 20,6 17 17,6 16,5 18,2
16,9 15,5 19,2 18,5 31,9 34,1 35,7 27,4 25,4 23 27,2 23,9 34,2 34,5 31,8
37,3 28,4 28,8 26,8 33,5 41,5 38,1 32,1 37,2 28 26,4 24,3 19,1 34,3 29,8
31,3 37 32,2 46,6 27,9 40,8 44,3 43,4 36,4 30,4 44,6 40,9 33,8 29,8 32,7
23,7 35 23,6 32,4 27,2 26,6 25,8 23,5 30 39,1 39 35,1 32,3 37 37,7
34,1 34,7 34,4 29,9 33 34,5 33,7 32,4 32,9 31,6 28,1 30,7 25,4 24,2 22,4
26,6 20,2 17,6 28 27 34 31 29 27 24 23 36 37 31 38
36 36 36 34 38 32 38 25 38 26 22 32 36 27 27
44 32 28 31

Para explorar esta información elaboramos la tabla de frecuencias asociada a los datos y estudiamos la posible normalidad y simetría de la distribución del consumo de combustible. Como se trata de una variable cuantitativa con 154 valores comprendidos entre 13 y 49, será necesario agruparlos en intervalos o clases. Para ello tomamos 12 intervalos de igual anchura (12 es un entero que aproxima bien la raíz cuatrada de N = 154). La anchura de los intervalos será (49 – 13)/12 = 3. Se obtiene la tabla de frecuencias de la Figura 6-1.

Intervalo	Límite inferior	Límite superior	Marca de clase	n_i	$f_i = n_i/N$	N_i	$F_i = n_i/N$
1	13,0	16,0	14,5	1	0,0065	1	0,0065
2	16,0	19,0	17,5	14	0,0909	15	0,0974
3	19,0	22,0	20,5	22	0,1429	37	0,2403
4	22,0	25,0	23,5	15	0,0974	52	0,3377
5	25,0	28,0	26,5	22	0,1429	74	0,4805
6	28,0	31,0	29,5	16	0,1039	90	0,5844
7	31,0	34,0	32,5	22	0,1429	112	0,7273
8	34,0	37,0	35,5	22	0,1429	134	0,8701
9	37,0	40,0	38,5	11	0,0714	145	0,9416
10	40,0	43,0	41,5	3	0,0195	148	0,9610
11	43,0	46,0	44,5	5	0,0325	153	0,9935
12	46,0	49,0	47,5	1	0,0065	154	1,0000

Figura 6-1

Hemos observado los 154 valores sobre el consumo de los automóviles que inicialmente no aportaban mucha información. Evidentemente hay una variabilidad en el consumo de los automóviles; sin embargo, es muy difícil detectar qué patrón sigue dicha variabilidad para determinar mejor la estructura de los datos. Por ello, en primer lugar ha sido conveniente realizar una ordenación de los datos según su magnitud, es decir, una tabla de frecuencias, que aportará algo de luz sobre la distribución de frecuencias subyacente.

© Alfaomega - RC Libros

La siguiente tarea es la construcción del histograma de frecuencias, gráfico adecuado para una variable cuantitativa con sus valores agrupados en intervalos. Su representación se presenta en la Figura 6-2.

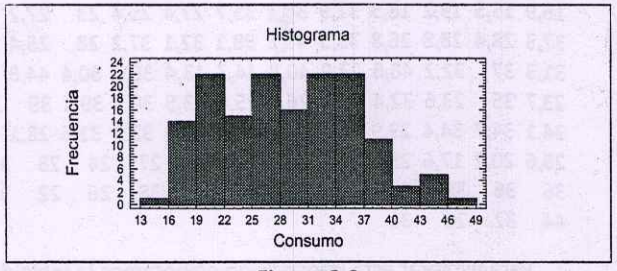


Figura 6-2

Se observa que la distribución subyacente que modela los datos sobre la variable consumo de los automóviles es aproximadamente simétrica y ajustable a forma de campana, lo que permite pensar en la existencia de normalidad y simetría en la distribución de *X*. Vemos así que el histograma da una idea clara de la distribución de la variable, incluyendo un modelo probabilístico para su modelación, en este caso la distribución normal. El simple examen de los datos tabulados inicialmente no aportaba información alguna; sin embargo, su graficación da luz al proceso.

Diagrama de tallo y hojas

El diagrama de tallo y hojas es un procedimiento semigráfico para presentar la información para variables cuantitativas, que es especialmente útil cuando el número total de datos es pequeño (menor que 50). Los principios para la realización del diagrama (debido a Tukey) son los siguientes:

- Redondear los datos a dos o tres cifras significativas.
- Disponerlos en dos columnas separadas por una línea vertical de tal forma que para los datos con dos dígitos la cifra de las decenas se encuentre a la izquierda de la línea vertical (tallo del diagrama), y a la derecha las unidades (hojas o ramas del diagrama). Por ejemplo, 87 se escribirá 8 7. Para datos con tres dígitos el tallo estará formado por los dígitos de las centenas y las decenas, que se escribirán a la izquierda de la línea vertical, y las hojas estarán formadas por el dígito de las unidades, que se escribirá a la derecha de la línea vertical.

© Alfaomega - RC Libros

- Cada tallo define una clase, y se escribe solo una vez. A su derecha se van escribiendo por orden las sucesivas hojas correspondientes a ese tallo. El número de hojas para cada tallo representa la frecuencia de cada clase.

El diagrama de tallo y hojas, también llamado *histograma digital*, es una combinación entre un histograma de barras y una tabla de frecuencias. Al mantener los valores de la variable, el diagrama de tallo y hojas resulta más informativo que el clásico histograma de barras, ya que conserva los datos originales y, al mismo tiempo, compone un perfil que ayuda a estudiar la forma y simetría de la distribución. Se trata pues de una herramienta de análisis exploratorio de datos que muestra el rango de los datos, dónde están más concentrados, su simetría y la presencia de datos atípicos. Este procedimiento no es muy aconsejable para conjuntos de datos grandes.

A continuación se presenta el diagrama de tallo y hojas (Figura 6-3) para la variable *X* relativa a la variable consumo de los automóviles definida en el apartado anterior.

```
Diagrama de Tallo y Hojas para X: unidad = 1,0   1|2 representa 12,0

     23     1|566677777788888999999999
     51     2|000000000011122233333333444
    (31)    2|5555566666777777777788888899999
     72     3|000001111111122222222222333344444444444
     34     3|555666666667777788888999
      9     4|00133444
      1     4|6
```

Figura 6-3

El rango de *X* ha sido dividido en 7 clases o intervalos llamados *tallos*, cada uno de ellos representado por una fila del diagrama. El primer número de cada fila (separado de los demás) presenta la frecuencia absoluta de la clase correspondiente. El segundo número de cada fila presenta la cifra de las decenas de cada valor de *X* en su correspondiente clase. El resto de los números de cada fila (llamados *hojas*) son las cifras de las unidades de todos los elementos de la clase definida por la fila. De esta forma, además de presentar la distribución de los elementos en forma de histograma horizontal, en el diagrama se observan los propios elementos. Las hojas permiten analizar la simetría, la normalidad y otras características de la distribución de igual forma que un histograma.

Gráfico de caja y bigotes

El gráfico de caja y bigotes permite analizar y resumir un conjunto de datos univariante dado. Esta herramienta de análisis exploratorio de datos va a permitir

© Alfaomega - RC Libros

estudiar la simetría de los datos, detectar valores atípicos y vislumbrar un ajuste de los datos a una distribución de frecuencias determinada.

El gráfico de caja y bigotes divide los datos en cuatro áreas de igual frecuencia, una caja central dividida en dos áreas por una línea vertical y otras dos áreas representadas por dos segmentos horizontales (bigotes) que parten del centro de cada lado vertical de la caja. La caja central encierra el 50 por ciento de los datos. El sistema dibuja la mediana como una línea vertical en el interior de la caja. Si esta línea está en el centro de la caja no hay asimetría en la variable. Los lados verticales de la caja están situados en los cuartiles inferior y superior de la variable. Partiendo del centro de cada lado vertical de la caja se dibujan los dos bigotes, uno hacia la izquierda y el otro hacia la derecha. El bigote de la izquierda tiene un extremo en el primer cuartil Q_1, y el otro en el valor dado por el primer cuartil menos 0,5 veces el rango intercuartílico, esto es, $Q_1-1{,}5*(Q_3-Q_1)$.

El bigote de la derecha tiene un extremo en el tercer cuartil Q_3 y el otro en el valor dado por el tercer cuartil más 1,5 veces el rango intercuartílico, esto es, $Q_3+1{,}5*(Q_3-Q_1)$. El sistema considera valores atípicos (*outliers*) los que se encuentren a la izquierda del bigote izquierdo y a la derecha del bigote derecho. El sistema separa estos datos del resto y los representa mediante puntos alineados con la línea horizontal central para que sean fáciles de detectar. En el interior de la caja central se representa la media con un signo más.

A continuación se presenta el gráfico de caja y bigotes para la variable *X* relativa a la variable consumo de los automóviles (Figura 6-4).

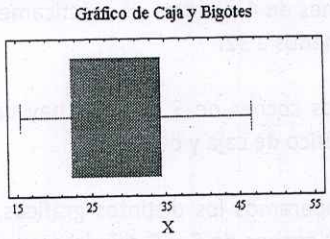


Figura 6-4

El gráfico permite afirmar que la variable *X* (consumo de los automóviles cada 1000 kilómetros) varía entre 15,5 y 46,6 y que el 50% central de los coches consume entre 22 (primer cuartil) y 34,5 (tercer cuartil) litros a los 1000 kilómetros.

© Alfaomega - RC Libros

Por otra parte, no existen valores de *X* anormalmente grandes *(outliers)*, ya que en la figura no aparecen puntos alineados con los bigotes. La distribución es ligeramente asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la figura es mayor que la de la izquierda. La mediana corresponde aproximadamente al valor 29 de *X*.

Los gráficos de caja y bigotes permiten la opción de representación *Muesca de mediana*, que sitúa dos muescas sobre los extremos de la mediana.

La anchura de la muesca representa un intervalo de confianza aproximado para la mediana con un coeficiente de confianza del 95% (α = 0,05) que viene determinado por la expresión $M \pm(1{,}25R/1{,}35\sqrt{n})\ (1+1/\sqrt{2})Z_{\alpha/2}/2$, donde *R* es el rango intercuartílico de la variable, *M* es la mediana y $Z_{\alpha/2}$ es el valor de la distribución normal (0,1) que deja a la derecha $\alpha/2$ de probabilidad.

Para nuestra variable *X*, consumo de los automóviles, tendremos el gráfico de caja y bigotes con muescas de la Figura 6-5.

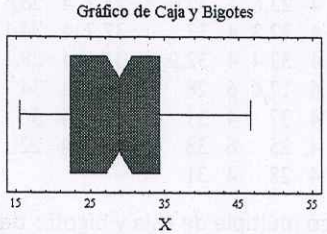


Figura 6-5

Gráfico múltiple de caja y bigotes

En estadística es típico dividir el conjunto de datos de una variable en subgrupos racionales, que pueden ser por ejemplo estratos definidos según una determinada variable de estratificación. El gráfico múltiple de caja y bigotes va a permitir analizar, resumir y comparar simultáneamente varios conjuntos de datos univariantes dados, correspondientes a los diferentes grupos en que se pueden subdividir los valores de una variable. Esta herramienta de análisis exploratorio de datos va a permitir estudiar la simetría de los datos, detectar valores atípicos y representar medias, medianas, rangos y valores extremos para todos los grupos. Al ser la representación simultánea para todos los conjuntos de datos, se podrán

© Alfaomega - RC Libros

comparar medias, medianas, rangos, valores extremos, simetrías y valores atípicos de todos los grupos. El gráfico múltiple representará horizontalmente un gráfico de caja y bigotes para cada grupo de valores de la variable en estudio.

Si clasificamos el consumo de los automóviles (variable X) según su cilindrada (variable Y), estamos haciendo subgrupos con los valores de X según la variable de estratificación Y. Los posibles valores de Y son 8, 6, 5, 4 y 3 cilindros. Los valores de X para cada valor de Y vienen dados a continuación:

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
43,1	4	36,1	4	32,8	4	39,4	4	36,1	4	19,9	8	19,4	8	20,2	8	19,2	6	20,5	6
20,2	6	25,1	4	20,5	6	19,4	6	20,6	6	20,8	6	18,6	6	18,1	6	19,2	8	17,7	6
18,1	8	17,5	8	30	4	27,5	4	27,2	4	30,9	4	21,1	4	23,2	4	23,8	4	23,9	4
20,3	5	17	6	21,6	4	16,2	6	31,5	4	29,5	4	21,5	6	19,8	6	22,3	4	20,2	6
20,6	6	17	8	17,6	8	16,5	8	18,2	8	16,9	8	15,5	8	19,2	8	18,5	8	31,9	4
34,1	4	35,7	4	27,4	4	25,4	5	23	8	27,2	4	23,9	8	34,2	4	34,5	4	31,8	4
37,3	4	28,4	4	28,8	6	26,8	6	33,5	4	41,5	4	38,1	4	32,1	4	37,2	4	28	4
26,4	4	24,3	4	19,1	6	34,3	4	29,8	4	31,3	4	37	4	32,2	4	46,6	4	27,9	4
40,8	4	44,3	4	43,4	4	36,4	5	30,4	4	44,6	4	40,9	4	33,8	4	29,8	4	32,7	6
23,7	3	35	4	23,6	4	32,4	4	27,2	4	26,6	4	25,8	4	23,5	6	30	4	39,1	4
39	4	35,1	4	32,3	4	37	4	37,7	4	34,1	4	34,7	4	34,4	4	29,9	4	33	4
34,5	4	33,7	4	32,4	4	32,9	4	31,6	4	28,1	4	30,7	6	25,4	6	24,2	6	22,4	6
26,6	8	20,2	6	17,6	6	28	4	27	4	34	4	31	4	29	4	27	4	24	4
23	4	36	4	37	4	31	4	38	4	36	4	36	4	36	4	34	4	38	4
32	4	38	4	25	6	38	6	26	4	22	6	32	4	36	4	27	4	27	4
44	4	32	4	28	4	31	4												

El gráfico múltiple de caja y bigotes del consumo de los automóviles (variable *X*) según su cilindrada (variable *Y*), se observa en la Figura 6-6.

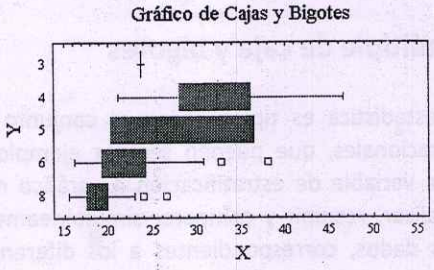


Figura 6-6

© Alfaomega - RC Libros

El gráfico permite afirmar que la variable *X* (litros consumidos a los 1000 kilómetros) *para los coches de 8 cilindros* varía entre 15,5 y 23, y que el 50% central de estos coches consume entre 17,5 (primer cuartil) y 20 (tercer cuartil) litros a los 1000 kilómetros, existiendo 2 valores de *X* anormalmente grandes *(outliers)*, ya que en la figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de *X* para los coches de 8 cilindros es ligeramente asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la figura es mayor que la de la izquierda, y la mediana corresponde aproximadamente al valor 18,5 de *X*, siendo la media 19,5 aproximadamente.

Para los coches de 6 cilindros los litros consumidos cada 1000 kilómetros (variable *X*) varían entre 16 y 31, concentrándose el 50% central de los valores de *X* entre 19 (primer cuartil) y 24 (tercer cuartil), existiendo 2 valores atípicos de *X* anormalmente grandes *(outliers)*, ya que en la figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de *X* para los coches de 6 cilindros es asimétrica hacia la derecha, la mediana de *X* se aproxima a 20,5 y la media a 22,5.

Para los coches de 5 cilindros los litros consumidos cada 1000 kilómetros (variable *X*) varían entre 20,2 y 36,5, concentrándose el 50% central de los valores de *X* entre los mismos valores, no existiendo bigotes ni *outliers*. La distribución de *X* para los coches de 5 cilindros es asimétrica hacia la derecha, la mediana de *X* se aproxima a 25,5 y la media a 27,5.

Para los coches de 4 cilindros los litros consumidos cada 1000 kilómetros (variable *X*) varían entre 21 y 47, concentrándose el 50% central de los valores de *X* entre 28 (primer cuartil) y 36 (tercer cuartil), no existiendo *outliers*. La distribución de *X* para los coches de 4 cilindros es prácticamente simétrica con valores de mediana y media aproximados a 32.

Para los coches de 3 cilindros hay un único valor de *X*, lo que no permite construir el gráfico de caja y bigotes.

Si comparamos los distintos gráficos, vemos que la asimetría de *X* es más fuerte para los coches de 5 y 6 cilindros, para los de 8 es menor y para los de 4 no existe. Valores de *X* anormalmente grandes solo aparecen para los coches de 6 y 8 cilindros. Las medias y las medianas varían bastante para los diferentes grupos de valores de *X* determinados por los valores de *Y*.

© Alfaomega - RC Libros

Gráfico de simetría y normalidad

El gráfico de simetría es una herramienta que permite analizar visualmente el grado de simetría de una variable. En el eje de abscisas se representan las distancias de los valores de la variable a la mediana que quedan por debajo de ella, y en el eje de ordenadas se representan las distancias de los valores de la variable a la mediana que quedan por encima de ella. Si la simetría fuese perfecta, el conjunto de puntos resultante sería la diagonal principal. Cuanto más se aproxime la gráfica a la diagonal, más simetría existirá en la distribución de la variable.

Para el ejemplo de la variable *X*, variable definida por el número de litros consumidos por los automóviles cada 1000 kilómetros que venimos considerando durante todo el capítulo, tenemos el gráfico de simetría de la Figura 6-7.

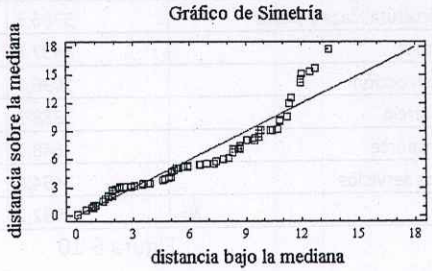


Figura 6-7

El gráfico realizado con la mediana, también puede realizarse con cuantiles y percentiles (gráficos QQ y PP respectivamente). Los gráficos PP y QQ se utilizan habitualmente para probar la normalidad de los datos. La variable será más normal cuanto mejor ajuste la nube de puntos a la diagonal del primer cuadrante.

Gráfico de dispersión

Se trata de un gráfico que permite ver la relación entre dos o más variables. Está formado por puntos cuyas coordenadas cartesianas son los pares de valores de dos variables cuya relación se quiere estudiar representada una en el eje vertical y otra en el eje horizontal.

El posicionamiento de los puntos del gráfico de dispersión define la relación entre las variables. Si se sitúan alrededor de una recta, existe correlación lineal entre

© Alfaomega - RC Libros

las variables. Si los puntos siguen una pauta no lineal, la relación entre las variables no puede definirse como lineal. Si la nube de puntos es aleatoria y dispersa, no existe relación alguna entre las variables. Un gráfico de dispersión para un análisis de varias variables que ofrece mucha información es el que representa en una estructura matricial los gráficos de dispersión de todos los pares posibles de variables (zona triangular inferior de la matriz), sus histogramas de frecuencias (diagonal de la matriz) y los coeficientes de correlación de todos los pares de variables en estudio (zona triangular superior de la matriz).

En la Figura 6-8 se muestra el gráfico matricial para las variables *Consumo*, *Potencia* y *Precio* de los automóviles.

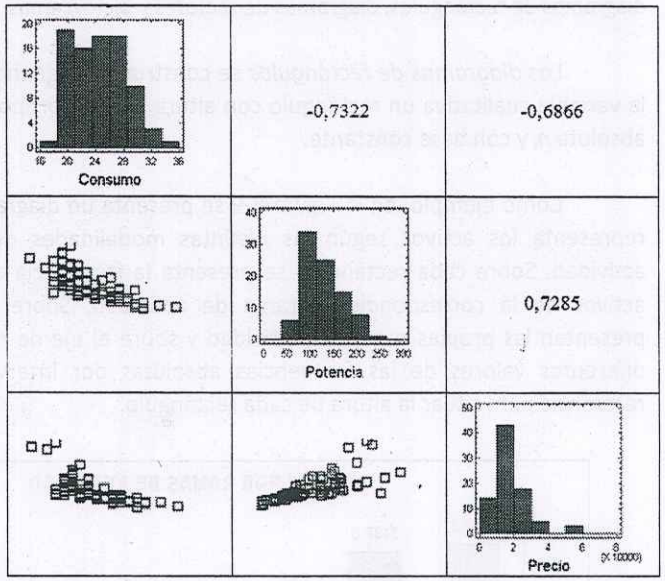


Figura 6-8

El gráfico de dispersión del consumo con la potencia situado en el elemento (2,1) de la matriz es fácilmente ajustable a una recta de pendiente negativa, lo que indica una relación fuerte y negativa entre ambas variables. Esta información la corrobora el coeficiente de correlación entre consumo y potencia (–0,7322) situado en el elemento (1,2) de la matriz. El gráfico de dispersión del consumo con el precio situado en el elemento (3,1) de la matriz es fácilmente ajustable a una recta de pendiente negativa, lo que indica una relación fuerte y negativa entre ambas variables. Esta información la

© Alfaomega - RC Libros

corrobora el coeficiente de correlación entre consumo y precio (–0,6866) situado en el elemento (1,3) de la matriz. El gráfico de dispersión de la potencia con el precio situado en el elemento (3,2) de la matriz es fácilmente ajustable a una recta de pendiente positiva, lo que indica una relación fuerte y positiva entre ambas variables. Esta información la corrobora el coeficiente de correlación entre potencia y precio (0,7285) situado en el elemento (2,3) de la matriz. Los histogramas de la diagonal principal de la matriz indican normalidad y simetría para el consumo y la potencia, y cierta asimetría para el precio.

Gráficos para variables cualitativas

La exploración visual de variables cualitativas suele llevarse a cabo mediante diagramas de rectángulos, diagramas de sectores y pictogramas.

Los *diagramas de rectángulos* se construyen asignando a cada modalidad de la variable cualitativa un rectángulo con altura igual (o proporcional) a su frecuencia absoluta n_i y con base constante.

Como ejemplo, en la Figura 6-9 se presenta un diagrama de rectángulos que representa los activos según las distintas modalidades de la variable rama de actividad. Sobre cada rectángulo se presenta la frecuencia absoluta n_i en miles de activos de la correspondiente rama de actividad. Sobre el eje de abscisas se presentan las propias ramas de actividad y sobre el eje de ordenadas se presentan diferentes valores de las frecuencias absolutas por intervalos que sirven como referencia para situar la altura de cada rectángulo.

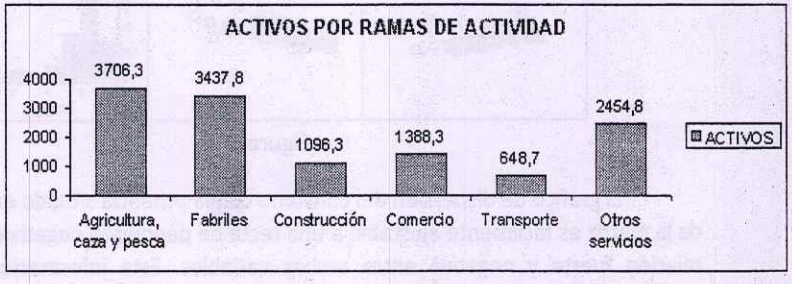


Figura 6-9

Los *diagramas de sectores* (o de pastel) constituyen el tipo de gráfico más utilizado para representar distribuciones de frecuencias de variables cualitativas. La

© Alfaomega - RC Libros

variable se representa en un círculo cuyas porciones (sectores circulares) tienen un área proporcional a las frecuencias absolutas de las modalidades de la variable. Para realizar el gráfico de sectores basta con asignar a cada modalidad de la variable un sector circular cuyo ángulo central sea proporcional a la frecuencia absoluta de la modalidad. Matemáticamente, el ángulo central α_i de la modalidad *i*-ésima cuya frecuencia absoluta es n_i puede expresarse como sigue:

$$\alpha_i = kn_i \Rightarrow \frac{\alpha_i}{n_i} = k = \frac{\sum \alpha_i}{\sum n_i} = \frac{360}{N} \Rightarrow \alpha_i = \frac{360}{N} n_i = 360\frac{n_i}{N} = 360 f_i$$

Los cálculos necesarios para nuestro ejemplo se presentan en la tabla de la Figura 6-10.

RAMA	ACTIVOS (n_i)	$f_i = n_i/N$	$\alpha_i = 360 f_i$
Agricultura, caza y pesca	3706,3	0,29	104,79
Fabriles	3437,8	0,27	97,20
Construcción	1096,3	0,09	31,00
Comercio	1388,3	0,11	39,25
Transporte	648,7	0,05	18,34
Otros servicios	2454,8	0,19	69,41
N=	12732,2		

Figura 6-10

La Figura 6-11 presenta el gráfico de sectores.

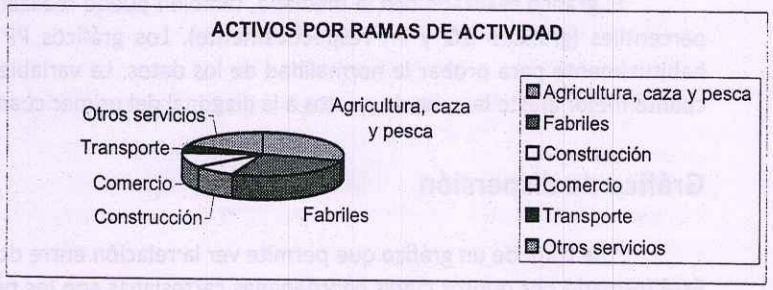


Figura 6-11

Otra forma habitual de construir gráficos de sectores consiste en asignar al sector circular relativo a la modalidad *i*-ésima un porcentaje igual al tanto por ciento que representa su frecuencia absoluta n_i sobre la frecuencia total $N = \sum n_i$.

© Alfaomega - RC Libros

Matemáticamente, la expresión del porcentaje p_i relativo a la modalidad i-ésima se expresa como sigue:

$$p_i = 100\frac{n_i}{N} = 100 f_i$$

La Figura 6-12 muestra el gráfico de sectores anterior con porcentajes.

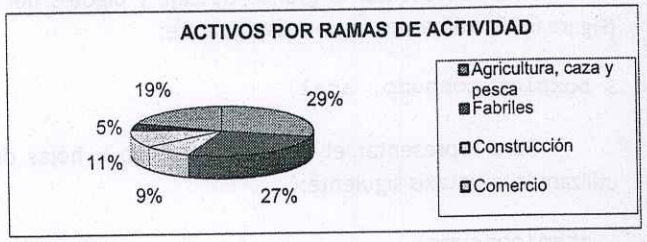


Figura 6-12

Los *pictogramas* se construyen representando de manera pictórica cada modalidad de la variable cualitativa indicando por una silueta sugestiva el significado de cada unidad de carácter. Por ejemplo, en la Figura 6-13 se representa sobre cada estado del mapa de Estados Unidos un cilindro cuya altura es proporcional a la cantidad de instalaciones con residuos peligrosos en 1997.

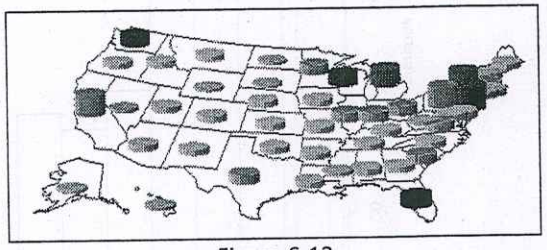


Figura 6-13

En los pictogramas puede usarse cualquier silueta para representar la frecuencia de cada modalidad de la variable cualitativa. En la Figura 6-14 se presenta el pictograma anterior con un cono sobre cada estado cuya altura indica la frecuencia de instalaciones con residuos peligrosos.

© Alfaomega - RC Libros

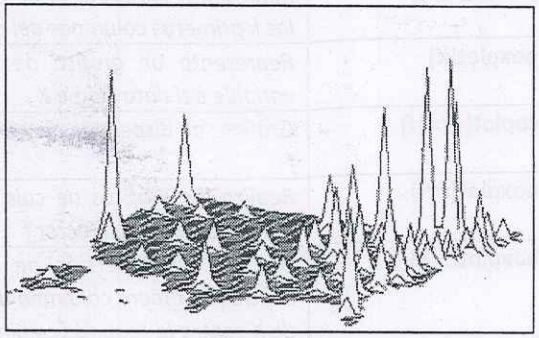


Figura 6-14

COMANDOS PARA GRÁFICOS EXPLORATORIOS EN R. SINTAXIS BÁSICA

A continuación se presenta un listado de comandos de R con su sintaxis más sencilla para realizar análisis exploratorio de datos.

hist(x)	*Representa el histograma de la variable x*
stem(x)	*Gráfico de tallo y hojas de la variable x*
barplot(x)	*Gráfico de barras con alturas dadas por la variable x*
dotchart(x)	*Gráfico de puntos de la variable x*
plot(x, y)	*Gráfico de dispersión que representa la nube de puntos* (x_i, y_i)
pie(x, etiquetas)	*Gráfico de sectores con áreas de cada sector dadas por la variable x y con las etiquetas dadas para cada área*
boxplot(x, y, z,...)	*Representa los gráficos de caja y bigotes para las variables x, y, z, ...*
qqnorm(x)	*Representa el gráfico de cuantiles para la normalidad de la variable x*
qqplot(x,y)	*Representa los cuantiles de y respecto a los cuantiles de x*
qqline(x,y)	*Añade líneas esperadas a qqplot(x)*
plot(density(x))	*Grafica la función de densidad kernel de la variable x*
pairs(X)	*Representa los gráficos de dispersión bivariantes entre las columnas del data frame X*
pairs(~x+y+z+...)	*Grafica los gráficos de dispersión bivariantes de x,y,z, ...*
plot(X)	*Representa los gráficos de dispersión bivariantes entre las columnas del data frame X*

© Alfaomega - RC Libros

plot(X[,1:k])	*Representa los gráficos de dispersión bivariantes entre las k primeras columnas del data frame X*
boxplot(X)	*Representa un gráfico de caja y bigotes para cada variable del data frame X*
coplot(y~x|f)	*Gráficos de dispersión de x e y para cada valor del factor f*
boxplot(x~f)	*Realiza los gráficos de caja y bigotes de la variable x para cada nivel del factor f*
boxplot(X,Y)	*Gráficos de dispersión de la primera columna de X contra la primera coluimna de Y, de la segunda columna de X contra la segunda coluimna de Y, etc.*
stripplot(x)	*Gráfico de los valores de x en una línea (alternatuiva a boxplot para pequeños tamaños de muestra)*
sunflowerplot(x,y)	*Semejante a plot, pero los puntos con similares coordenadas se dibujan como flores cuyo número de pétalos representan el número de puntos.*

Para ver algunos ejemplos consideramos el conjunto de datos *coches.sav* en formato SPSS que se encuentra situado en el subdirectorio *H:\R\data* del disco duro y que contiene las variables relativas al consumo (*consumo*), motorización (*motor*), potencia (*cv*), peso (*peso*), aceleración (*acel*), año de fabricación (*año*), origen geográfico (*origen*), cilindrada (*cilindr*) y un filtro indicativo de avería (*derivada*).

En primer lugar leemos el fichero y habilitamos sus variables.

```
> library(foreign)
coches=read.spss("G:/R/data/coches.sav", use.value.labels =
TRUE)
> attach(coches)
```

Para ver las variables del fichero y sus propiedades utilizamos la función *summary*.

```
> summary(coches)
        Length Class  Mode
consumo 406    -none- numeric
motor   406    -none- numeric
cv      406    -none- numeric
peso    406    -none- numeric
acel    406    -none- numeric
año     406    -none- numeric
```


© Alfaomega - RC Libros


```
origen   406    factor numeric
cilindr  406    factor numeric
derivada 406    factor numeric
```

Para representar el histograma del consumo (Figura 6-15) utilizamos la sintaxis siguiente:

```
> hist(consumo)
```

Para representar el gráfico de caja y bigotes del consumo y la aceleración (Figura 6-16) utilizamos la sintaxis siguiente:

```
> boxplot(consumo, acel)
```

Para representar el diagrama de tallo y hojas del consumo (Figura 6-17) utilizamos la sintaxis siguiente:

```
> stem(consumo)
```

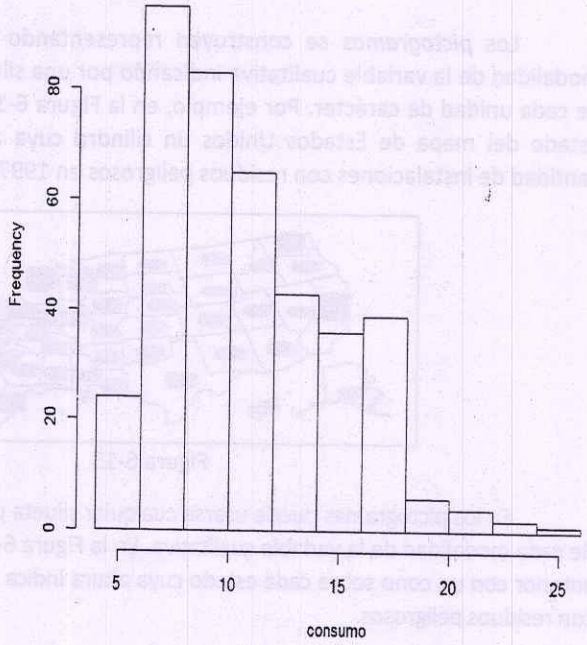


Figura 6-15

© Alfaomega - RC Libros

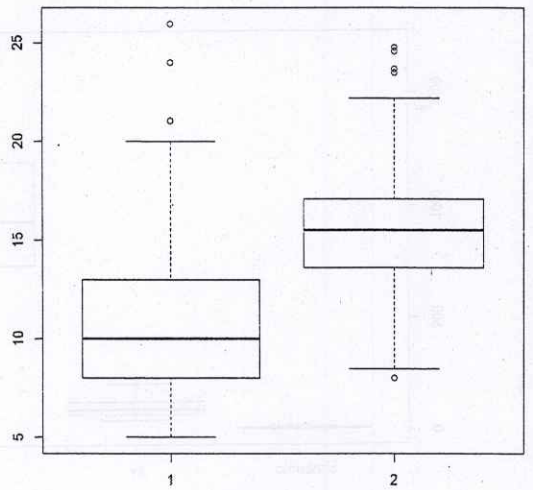


Figura 6-16

```
The decimal point is at the |

  5 | 000000
  6 | 00000000000000000
  7 | 0000000000000000000000000000000000000000000000000
  8 | 000000000000000000000000000000000000000000000000
  9 | 0000000000000000000000000000000000000000000000000000
 10 | 00000000000000000000000000000000
 11 | 000000000000000000000000000000
 12 | 00000000000000000000000000000000000
 13 | 000000000000000000000000000000000
 14 | 00000000000
 15 | 0000000000000000000
 16 | 000000000000000000
 17 | 0000000000000000000
 18 | 00000000000000000000
 19 |
 20 | 000000
 21 | 0000
 22 |
 23 |
 24 | 00
 25 |
 26 | 0
```

Figura 6-17

Para representar el diagrama de dispersión del consumo y la potencia (Figura 6-18) utilizamos la sintaxis siguiente:

```
> plot(consumo,cv)
```


© Alfaomega - RC Libros

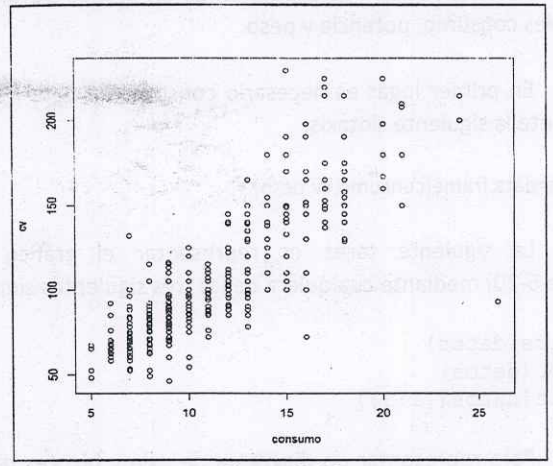


Figura 6-18

Para representar el gráfico de normalidad del peso (Figura 6-19) utilizamos la sintaxis siguiente:

```
> qqnorm(peso)
```

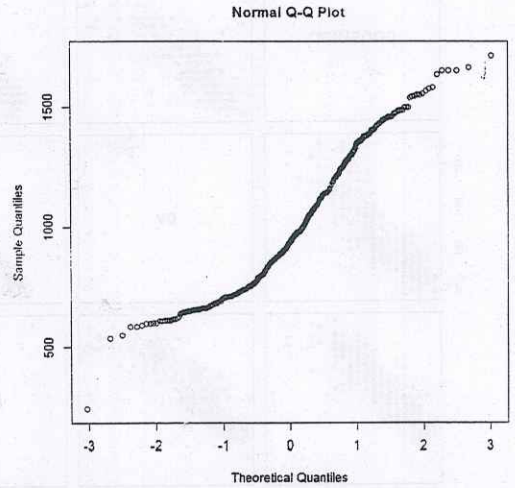


Figura 6-19

© Alfaomega - RC Libros

A continuación representaremos un gráfico de dispersión múltiple para las variables consumo, potencia y peso.

En primer lugar es necesario construir un data frame con las tres variables mediante la siguiente sintaxis:

```
> datos=data.frame(consumo,cv,peso)
```

La siguiente tarea es representar el gráfico de dispersión múltiple (Figura 6-20) mediante cualquiera de las tres siguientes sintaxis:

```
> pairs(datos)
> plot(datos)
> plot(datos[,1:3])
```

Para representar un diagrama de caja y bigotes para cada variable del data frame sobre el mismo gráfico (Figura 6-21), se utiliza la siguiente sintaxis:

```
> boxplot(datos)
```

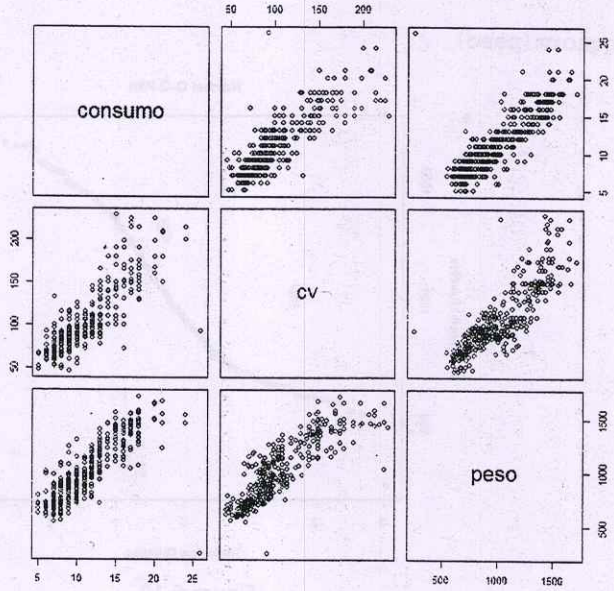


Figura 6-20

© Alfaomega - RC Libros

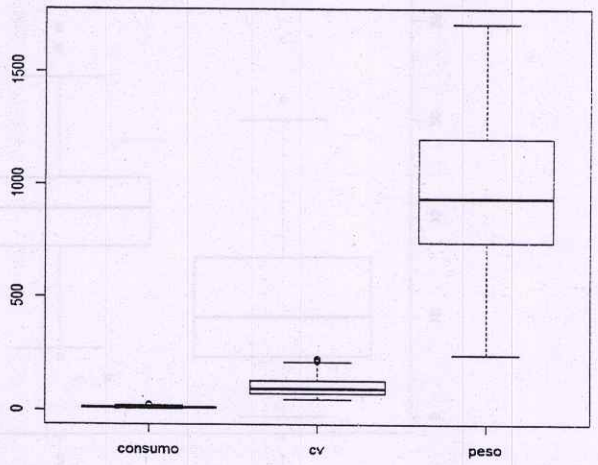


Figura 6-21

También se puede realizar un gráfico dispersión múltiple (Figura 6-22) para las dos primeras variables del dataframe mediante la sintaxis siguiente:

```
plot(datos[,1:2])
```

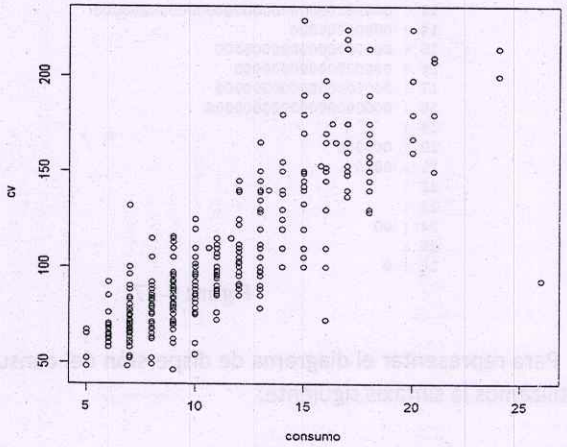


Figura 6-22

© Alfaomega - RC Libros

A continuación realizamos un gráfico de dispersión del consumo contra la potencia para todos los niveles del factor origen (Figura 6-23). La sintaxis sería la siguiente:

```
> coplot(consumo~cv|origen)
```

A continuación realizamos un gráfico de caja y bigotes del consumo para todos los niveles del factor origen (Figura 6-24). La sintaxis sería la siguiente:

```
> boxplot(consumo~origen)
```

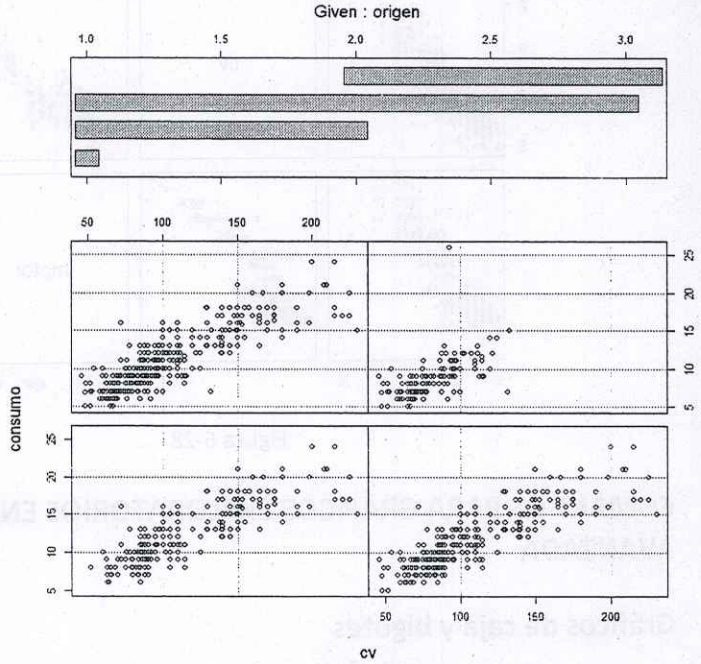


Figura 6-23

© Alfaomega - RC Libros

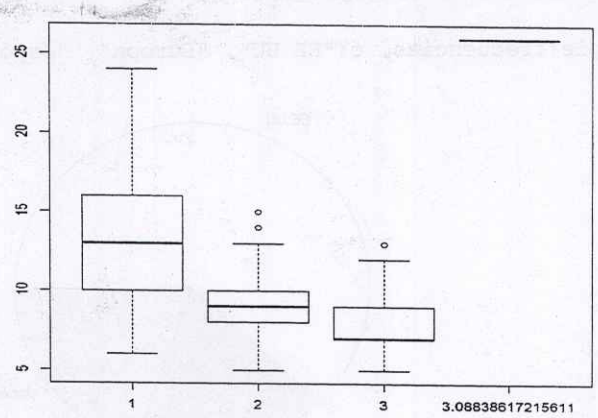


Figura 6-24

En la Figura 6-25 se presenta el gráfico de barras para la variable *origen* construido a través de la siguiente sintaxis:

```
> frecuencias=table(origen)
> frecuencias
origen
  1   2   3
253  73  80
> barplot(frecuencias)
```

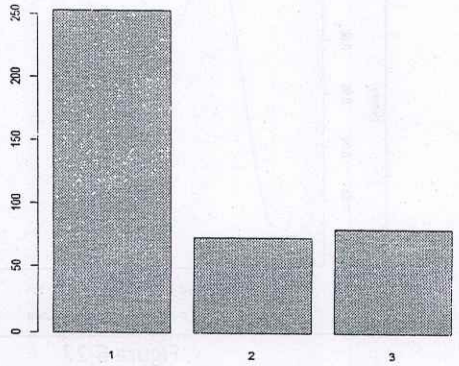


Figura 6-25

© Alfaomega - RC Libros

En la Figura 6-26 se presenta el gráfico de sectores para la variable *origen* construido a través de la siguiente sintaxis:

```
> pie(frecuencias, c("EE UU", "Europa", "Japón"))
```

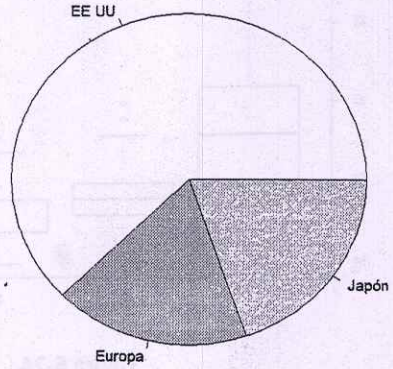


Figura 6-26

Si queremos representar la función de densidad kernel del consumo (Figura 6-27) utilizaremos la siguiente sintaxis:

```
> density(consumo)
```

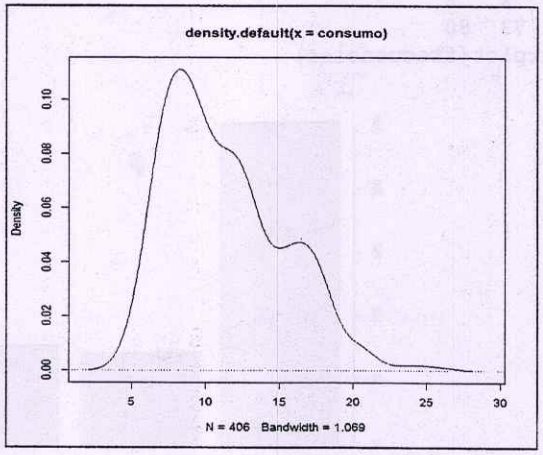


Figura 6-27

© Alfaomega - RC Libros

Si queremos representar el diagrama de dispersión múltiple de las variables *consumo, cv* y *motor* (Figura 6-28) utilizaremos la siguiente sintaxis:

```
> pairs(~consumo+cv+motor)
```

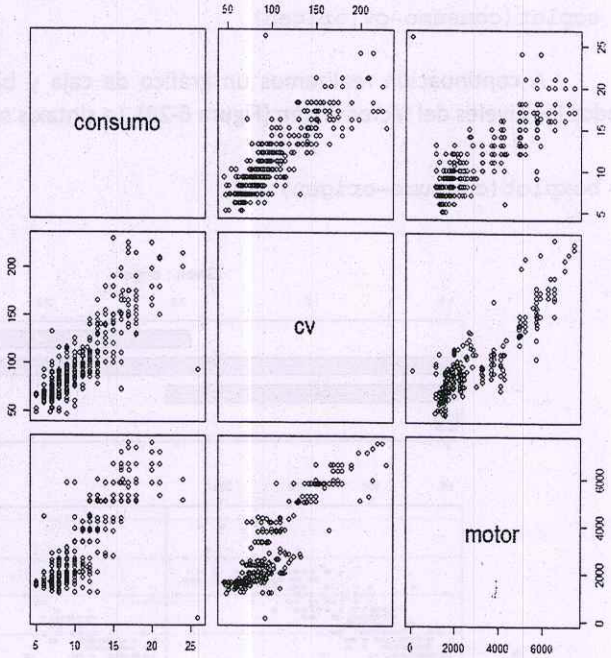


Figura 6-28

COMANDOS PARA GRÁFICOS EXPLORATORIOS EN R. SINTAXIS AVANZADA

Gráficos de caja y bigotes

La sintaxis general del comando *boxplot* es la siguiente:

boxplot(x, ..., range = 1.5, width = NULL, varwidth = FALSE, notch = FALSE, names, boxwex = 0.8, data = parent.frame(), plot = TRUE, border = par("fg"), col = NULL, log = "", pars = NULL, horizontal = FALSE, add = FALSE)

© Alfaomega - RC Libros

Sus argumentos son:

- *x, ...*: Las variables que van a ser graficadas con el boxplot. Puede ser un conjunto de vectores separados, cada uno correspondiendo a un boxplot componente o una lista que contiene a tales vectores. Alternativamente, una especifiación de la forma "x ~ g" puede ser dada para indicar que las observaciones en el vector "x" van a ser agrupadas de acuerdo a los niveles del factor "g". En este caso, el argumento "data" puede usarse para que se admitan valores NA en los datos.
- *range*: Determina la extensión de los bigotes (whiskers) de la caja. Si es positivo, los bigotes se extienden hasta el dato más extremo, el cual no es más que el rango intercuartílico de la caja. Un valor de cero hace que los bigotes se extiendan hasta los datos extremos.
- *width*: Un vector que da los anchos relativos de las cajas.
- *varwidth*: Si es "TRUE", las cajas son dibujadas con anchos proporcionales a las raíces cuadradas del número de observaciones en los grupos.
- *notch:* Si es "TRUE", una cuña es dibujada a cada lado de las cajas. Cuando las cuñas de dos gráficos de caja no se superponen, entonces las medianas son significativamente diferentes a un nivel del 5%.
- *names*: Un grupo de etiquetas a ser impresas debajo de cada boxplot.
- *boxwex*: Un factor de escala a ser aplicado a todas las cajas. Cuando solo hay unos pocos grupos, la apariencia del gráfico puede mejorarse haciendo las cajas más angostas.
- *data*: "data.frame", "list", o "environment" en el cual los nombres de las variables son evaluados cuando "x" es una fórmula.
- *plot*: Si es "TRUE" (por defecto) entonces se produce el gráfico. De lo contrario, se producen los resúmenes de los boxplots.
- *border*: Un vector opcional de colores para las líneas de las cajas y debe ser de longitud igual al número de gráficos.
- *col*: Si se especifica, los colores que contiene serán usados en el cuerpo de las cajas.

© Alfaomega - RC Libros

- *log*: Para indicar si las coordenadas "x" o "y" serán graficadas en escala logarítmica.
- *pars*: Los parámetros gráficos pueden ser pasados como argumentos para el boxplot.
- *horizontal*: Si es "FALSE" (por defecto) los boxplots son dibujados verticalmente.
- *add*: Si es "TRUE" agrega un boxplot al gráfico actual.

También se admite la siguiente sintaxis para el comando boxplot:

boxplot.formula(formula, data = NULL, subset, na.action, ...)

Los argumentos son los siguientes:

- *formula*: Una fórmula tal como "y ~ x".
- *data*: Un data.frame (o lista) del cual las variables en "fórmula" serán tomadas.
- *subset*: Un vector opcional para especificar un subconjunto de observaciones a ser usadas en el proceso de ajuste.
- *na.action*: Una función que indica lo que debería pasar cuando los datos contienen NA.

Como ejemplo, vamos a representar el gráfico de caja y bigotes (Figura 6-29) del consumo de los automóviles según cilindradas utilizando cajas con muescas de color verde, con las cajas dibujadas con anchos proporcionales a las raíces cuadradas del número de observaciones en los grupos y con títulos para el gráfico y los ejes.

```
> library(foreign)
coches=read.spss("G:/R/data/coches.sav", use.value.labels=TRUE)
> attach(coches)
```

© Alfaomega - RC Libros

```
> boxplot(consumo~cilindr,
notch=TRUE,varwidth=TRUE,col="green", main="Datos de
automóviles", xlab="Número de cilindros", ylab="consumo")
```

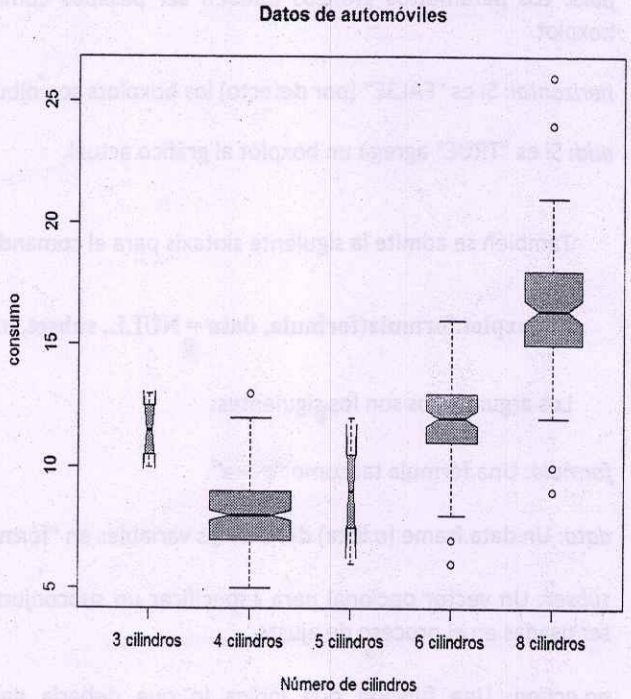


Figura 6-29

A continuación se presenta el gráfico de caja y bigotes anterior en horizontal y sin las características especiales de que se había dotado (Figura 6-30).

```
> boxplot(consumo~cilindr, horizontal=TRUE)
```


© Alfaomega - RC Libros

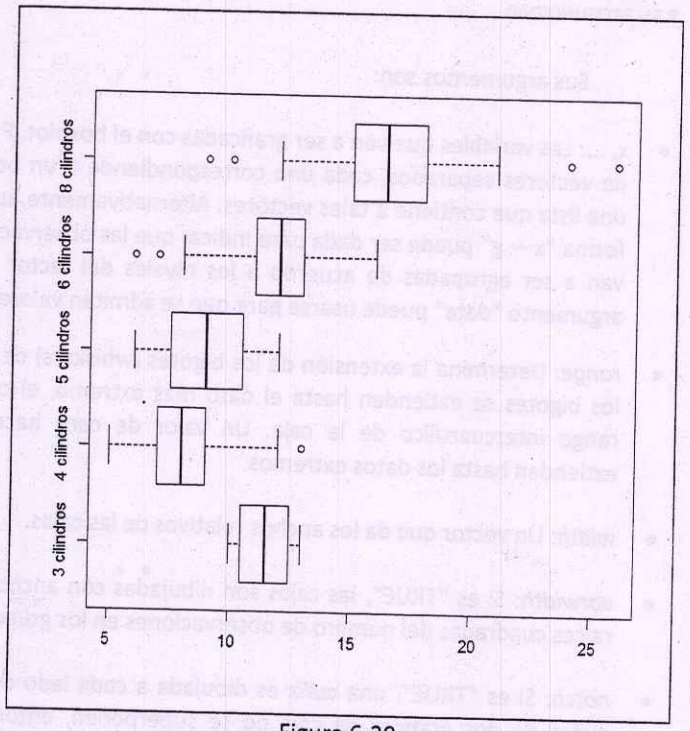


Figura 6-30

Gráficos de barras

El comando *barplot* tiene la siguiente sintaxis general:

barplot(height, width = 1, space = NULL, names.arg = NULL, legend.text = NULL, beside = FALSE, horiz = FALSE, col = heat.colors(NR), border = par("fg"), main = NULL, sub = NULL, xlab = NULL, ylab = NULL, xlim = NULL, ylim = NULL, axes = TRUE, axisnames = TRUE, inside = TRUE, plot = TRUE, ...)

Los argumentos son los siguientes:

- *height*: Vector o matriz de valores que describen las barras. Si es un vector, entonces el gráfico corresponde a un secuencia de barras rectangulares cuyas alturas corresponden a los valores del vector. Si es una matriz y "beside=FALSE" entonces cada barra del gráfico corresponde a una columna de forma que los

© Alfaomega - RC Libros

valores en cada columna serán representados por sub-barras apiladas, pero si "beside=TRUE", entonces los valores en cada columna serán yuxtapuestos en vez de apilados.

- *width*: Es opcional, para especificar mediante un vector el ancho de las barras.
- *space*: Para fijar el espacio entre barras, como una fracción del ancho promedio de estas. Puede especificarse con un solo número o un número por barra. En el caso de que el objeto representado sea una matriz y "beside=TRUE", este argumento puede darse por dos números, el primero para el espacio entre barras del mismo grupo y el segundo para el espacio entre grupos. Por defecto está ajustado a 'c(0,1)' para el caso de objeto matriz y beside=TRUE.
- *names.arg*: Un vector de nombres para colocarlos debajo de cada barra o grupo de barras. Si se omite, entonces los nombres son tomados de los atributos de nombres que estén contenidos en el objeto especificado en "height".
- *legend.text*: Un vector de texto para construir una leyenda para el gráfico. Solo es útil si "height" es una matriz, en cuyo caso las leyendas corresponderán a sus filas.
- *beside*: Un valor lógico. Si es "FALSE" las columnas de "height" serán representadas por barras apiladas y si es "TRUE" entonces serán representadas por barras yuxtapuestas.
- *horiz*: Un valor lógico. Si es "FALSE" las barras son dibujadas verticalmente.
- *col*: Para especificar un vector de colores para las barras.
- *border*: Para indicar el color a ser usado en los bordes de las barras.
- *main,sub*: Para titular y subtitular el gráfico.
- *xlab*: Para dar nombre al eje x.
- *ylab*: Para dar nombre al eje y.

© Alfaomega - RC Libros

- *xlim*: Para delimitar el rango de valores en el eje *x*.
- *ylim*: Para delimitar el rango de valores en el eje *y*.
- *axes*: Argumento lógico. Si es "TRUE", dibuja el correspondiente eje.
- *axisnames*: Argumento lógico. Si es "TRUE", y si se ha especificado "names.arg", entonces el otro eje es dibujado y etiquetado.
- *plot*: Argumento lógico. Si es "FALSE" no se produce gráfico.

Supongamos que hemos recogido una muestra de 20 datos que codificamos con valores de 1 a 4. Se trata de realizar un histograma para estos valores con leyendas adecuadas.

En primer lugar definimos la variable *x* que contiene el vector de datos.

```
> x=c(1,2,3,1,2,1,2,4,1,3,2,4,1,2,3,1,2,4,2,1)
```

A continuación definimos la tabla de frecuencias correspondiente a estos datos:

```
> datos=table(x)
> datos
x
1 2 3 4
7 7 3 3
```

Por último, creamos el diagrama de barras con leyendas (Figura 6-31):

```
> barplot(datos, main="Frecuencia relativa", xlab="Valores de
la variable")
```

© Alfaomega - RC Libros

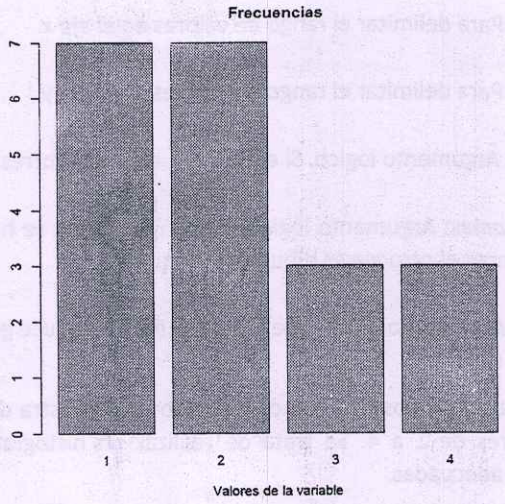


Figura 6-31

A continuación asignamos nombres a las categorías de la variable y realizamos el histograma en horizontal (Figura 6-32).

```
> barplot(datos, main="Frecuencia relativa", xlab="valores de
la variable", horiz=TRUE, names.arg=c("Menores de 18", "De 18 a
45", "De 45 a 65","Más de 65"))
```

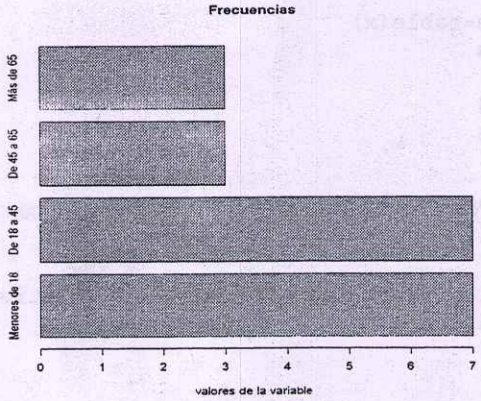


Figura 6-32

© Alfaomega - RC Libros

En el ejemplo siguiente representamos un diagrama de barras con las variables *x*, *y* y *z* agrupadas una al lado de la otra de grosores 45, 50 y 35 con los colores 1 y 2 (negro y rojo) y con la leyenda hombre y mujer (Figura 6-33).

```
> barplot(height = cbind(x = c(4, 9), y = c(8, 2), z = c(3, 7)
), beside = TRUE, width = c(45, 50, 35), col = c(1, 2),
legend.text = c("hombre", "mujer"))
```

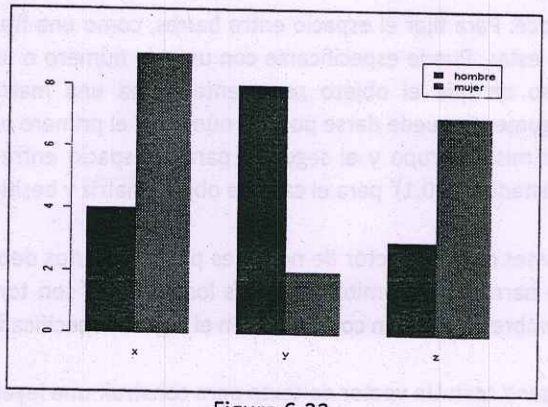


Figura 6-33

Gráficos de sectores

El comando *piechart* tiene la siguiente sintaxis general:

piechart(x, labels=names(x), shadow=FALSE, edges=200, radius=0.8, col=NULL, main=NULL, ...)

Sus argumentos son los siguientes:

- *x*: Vector de cantidades positivas a graficar como áreas de los sectores.
- *labels*: Un vector de caracteres "strings" que dan nombres a las áreas.
- *shadow*: Un vector lógico que indica si un efecto de sombrea será aplicado para el gráfico, cuando se utilizan colores de relleno.
- *edges*: Aproxima la línea exterior circular mediante un polígono con el número de lados especificado, que por defecto es 200.
- *radius*: El sector es dibujado centrado en una caja cuadrada cuyos lados se mueven de -1 a 1. Si se usan etiquetas largas puede ser necesario usar radios más pequeños.

© Alfaomega - RC Libros

- *col*: Un vector de colores, para rellenar los sectores del gráfico.
- *main*: Para dar título al gráfico.

Comenzamos realizando un gráfico de sectores sencillo para los datos del primer ejemplo del apartado anterior (Figura 6-34).

```
> x=c(1,2,3,1,2,1,2,4,1,3,2,4,1,2,3,1,2,4,2,1)
> datos=table(x)
> pie(datos)
```

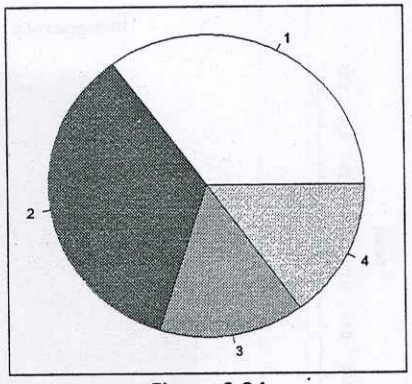


Figura 6-34

Ahora situamos etiquetas en las categorías del gráfico anterior (Figura 6-35).

```
> etiquetas=c("Menores de 18", "De 18 a 45", "De 45 a 65","Más
de 65")
> pie(datos, labels=etiquetas)
```

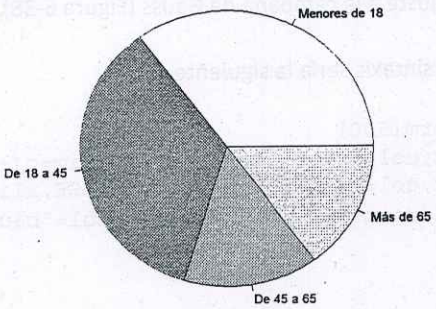


Figura 6-35

© Alfaomega - RC Libros

En el ejemplo siguiente consideramos una muestra de indicadores sintéticos de densidad de población en 5 países del mundo. Se trata de realizar un diagrama de sectores para las densidades con la etiqueta de cada país y con porcentajes indicativos de las densidades de población. Además, se coloreará cada sector con colores del arco iris.

```
> muestra <- c(15, 12, 4, 16, 8)
> paises <- c("USA", "Inglaterra", "Australia", "Alemania",
"Francia")
> pct = round(muestra/sum(muestra)*100)
> paises1=paste(paises,pct,"%")
> pie(muestra,labels = paises1, col=rainbow(length(paises)))
```

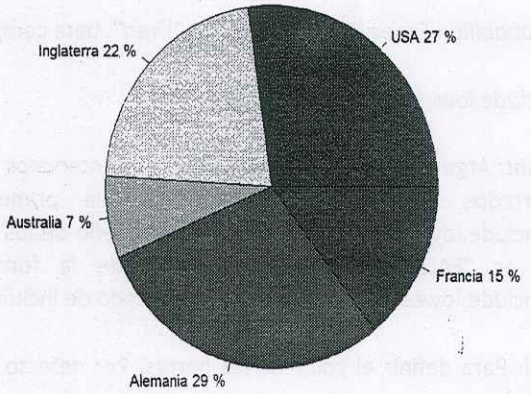


Figura 6-36

Histogramas

La sintaxis general del comando *hist* es la siguiente:

```
hist(x, breaks, freq = NULL, probability = !freq,
     include.lowest = TRUE,
     right = TRUE, col = NULL, border = par("fg"),
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, ...)
```

© Alfaomega - RC Libros

Los argumentos de este comando son:

- *x*: Vector de valores para el que se construye el histograma.
- *breaks*: Puede ser un valor con el cual se indica el número aproximado de clases o un vector cuyos elementos indican los puntos límites entre las clases o intervalos.
- *freq*: argumento lógico; si se especifica como "TRUE", el histograma representará las frecuencias absolutas o conteo de datos en cada clase; si se especifica como "FALSE", se reprentarán las frecuencias relativas. Por defecto, este argumento toma el valor de "TRUE" siempre y cuando los intervalos sean de igual ancho.
- *probability*: Especifica un alias para "!freq", para compatibilidad con S.
- *include.lowest*: Argumento lógico.
- *right:* Argumento lógico; si es "TRUE", los intervalos son abiertos a la izquierda - cerrados a la derecha (a,b]. Para la primera clase o intervalo si "include.lowest=TRUE" el valor más pequeño de los datos será incluido en este. Si es "FALSE" los intervalos serán de la forma [a,b) y el argumento "include.lowest=TRUE" tendrá el significado de incluir el "más alto".
- *col*: Para definir el color de las barras. Por defecto, "NULL" produce barras sin fondo.
- *border*: Para definir el color de los bordes de las barras.
- *plot*: Argumento lógico. Por defecto es "TRUE", y el resultado es el gráfico del histograma; si se especifica como "FALSE" el resultado es una lista de conteos por cada intervalo.
- *labels*: Argumento lógico o carácter. Si se especifica como "TRUE" coloca etiquetas arriba de cada barra.
- *nclass*: Argumento numérico (entero). Para compatibilidad con S, "nclass=n" es equivalente a "breaks=n".
- ...: Parámetros gráficos adicionales a "title" y "axis".

© Alfaomega - RC Libros

Como primer ejemplo podemos representar un histograma relativo a 500 datos de una distribución normal con 10 intervalos azules con líneas de separación de color rosa (Figura 6-37).

La sintaxis podría ser la siguiente:

```
> x<-rnorm(500)
> hist(x, freq = FALSE, breaks = 9, col="lightblue",
border="yellow")
```

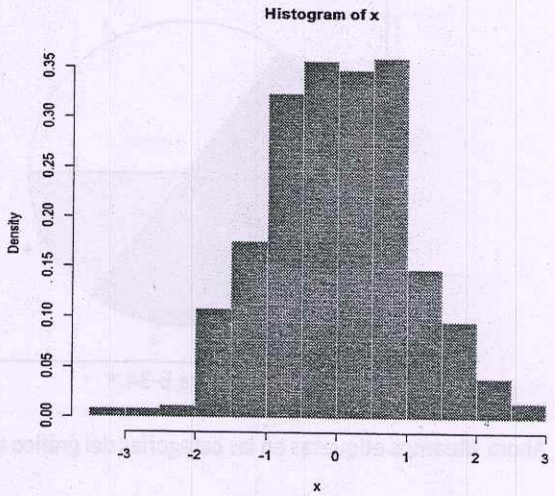


Figura 6-37

Como segundo ejemplo vamos a considerar el histograma anterior y se le añade el ajuste a la campana de Gauss (Figura 6-38).

La sintaxis sería la siguiente:

```
> x<-rnorm(500)
> hist(x,col="blue",freq=FALSE,xlim=c(-5,5))
> hist(x,col="lightblue",freq=FALSE,xlim=c(-5,5))
> curve(dnorm(x),-5,5,add=TRUE,col="blue")
```

© Alfaomega - RC Libros

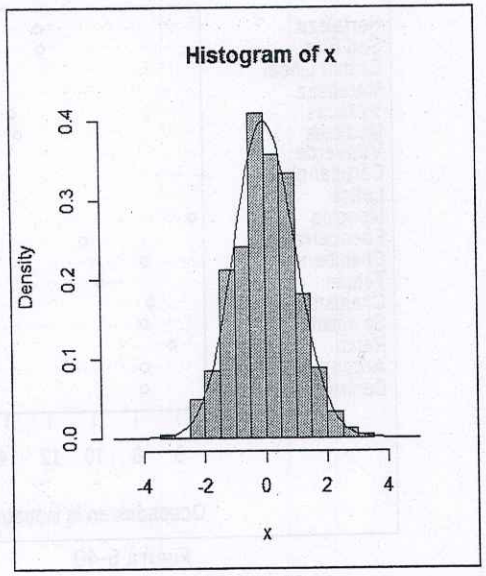


Figura 6-38

Gráficos de puntos

Los gráficos de puntos son elegantemente simples y permiten numerosas variaciones. En R se puede obtener dicho gráfico mediante la función "stripchart":

```
stripchart(x, method="overplot", jitter=0.1, offset=1/3, vertical=FALSE,
           group.names, xlim=NULL, ylim=NULL, main="", ylab="", xlab="",
           pch=0, col=par("fg"), cex=par("cex"))
```

Los argumentos de esta función son los siguientes:

- *x*: Objeto que contiene los datos a ser graficados. Puede ser un solo vector o una lista de vectores, cada uno correspondiendo a un componente del gráfico. También se puede especificar "x~g" para indicar que las observaciones en el vector "x" van a ser agrupadas según los niveles del factor "g".
- *method*: Para especificar el método a ser empleado para separar puntos coincidentes. Si se especifica como "overplot" (por defecto), los puntos

© Alfaomega - RC Libros

coincidentes son superpuestos. Si se especifica como "jitter", serán separados, y si se especifica como "stack", los puntos coincidentes son apilados. Se puede especificar el nivel de separación.

- *vertical*: Argumento lógico. Por defecto es "FALSE" y los puntos se grafican horizontalmente.
- *group.names*: Grupo de etiquetas a ser impresas al lado o debajo de cada gráfico.
- ...: Parámetros gráficos adicionales que pueden especificarse como argumentos.

Como primer ejemplo realizamos un gráfico de puntos en horizontal para los valores de la variable consumo según las cilindradas de los automóviles con los puntos coincidentes separados una décima (Figura 6-39).

```
> stripchart(consumo~cilindr, jitter=0.1, vertical=FALSE)
```

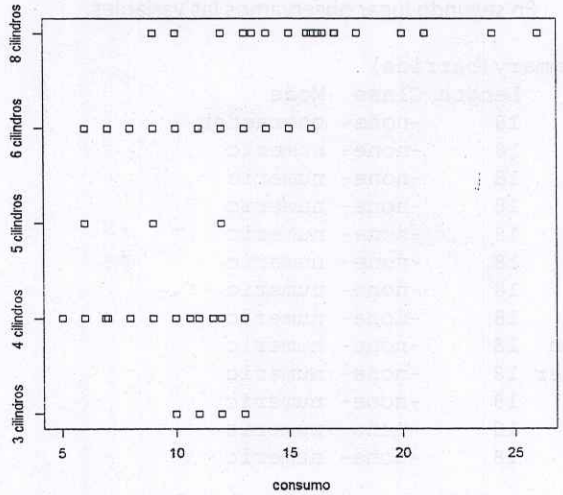


Figura 6-39

Otro comando que se suele utilizar en R para gráficos de puntos es *dotchart*, cuya sintaxis es la siguiente:

dotchart(x, labels=, cex=, gcolor= , main=, xlab=

© Alfaomega - RC Libros

Se realiza el gráfico de puntos para la variable *x*, con las etiquetas definidas en *labels* con los colores de etiquetas definidos en *gcolor* y con el tamaño de etiquetas definido en *cex*. El argumento *main* define el título del gráfico y el argumento *xlab* define la leyenda para el eje horizontal del gráfico.

En el ejemplo siguiente se realiza un gráfico de puntos para la variable del fichero *zonasmad.sav* que define los ocupados en la industria en los barrios de Madrid (*ocuin*), usando como etiqueta el barrio (variable *b*) y como tamaño de etiqueta 0.7. Situar un título adecuado en el eje horizontal (Figura 6-40).

En primer lugar leemos el fichero de formato SPSS y habilitamos sus variables.

```
> library(foreign)
> barrios=read.spss("G:/R/data/zonasmad.sav",
use.value.labels=TRUE)
> attach(barrios)
```

En segundo lugar observamos las variables.

```
> summary(barrios)
        Length Class  Mode
b       18     -none- character
pt      18     -none- numeric
p14     18     -none- numeric
p65     18     -none- numeric
p10     18     -none- numeric
anal    18     -none- numeric
nes     18     -none- numeric
ocu     18     -none- numeric
ocuin   18     -none- numeric
ocuser  18     -none- numeric
tec     18     -none- numeric
pd      18     -none- numeric
tm      18     -none- numeric
```

A continuación realizamos el gráfico de puntos.

```
> dotchart(ocuin, labels=b, xlab="Ocupados en la industria por
barrios", cex=.7)
```


© Alfaomega - RC Libros

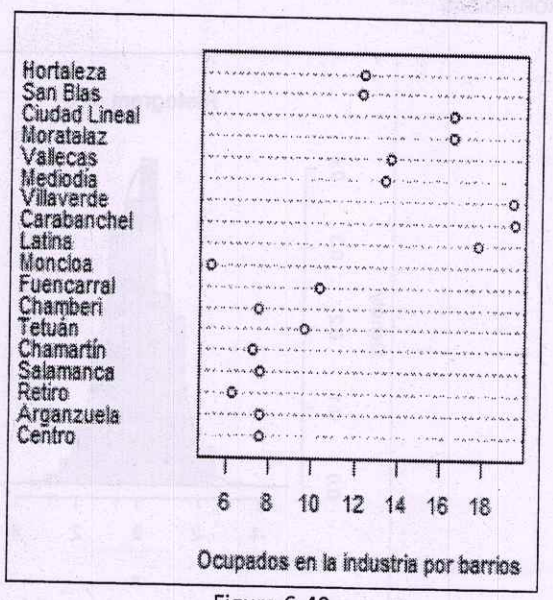


Figura 6-40

Ejercicio 6-1. Representar un gráfico de sectores y un gráfico de barras para la variable origen del fichero coches.

La sintaxis siguiente resuelve el problema y produce los gráficos de las Figuras 6-41 y 6-42. La lectura del fichero se realiza con etiquetas habilitadas para las variables cualitativas con la finalidad de obtener directamente las etiquetas de las barras y los sectores.

```
> library(foreign)
> coches=read.spss("G:/R/data/coches.sav",
use.value.labels=TRUE)
> attach(coches)
> datos=table(origen)
> windows()
> pie(datos)
> barplot(datos)
```

© Alfaomega - RC Libros

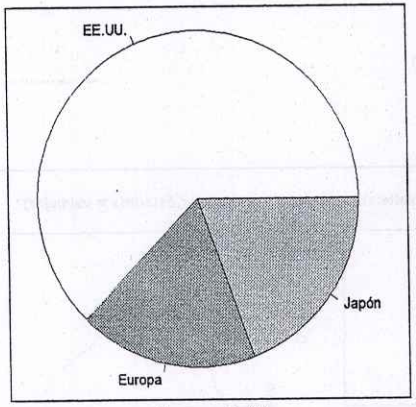


Figura 6-41

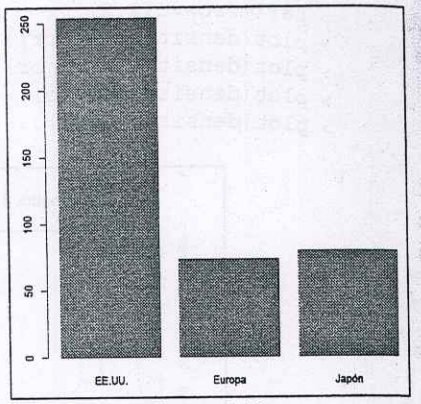


Figura 6-42

Ejercicio 6-2. Realizar un análisis exploratorio gráfico de las variables del archivo viviendas.sav que contiene información sobre barrios de una ciudad, precio de las viviendas, precio del suelo, valor de las mejoras y ratio entre precio real de la vivienda y precio tasado.

En primer lugar leemos el fichero como un data frame y habilitamos sus variables para el análisis.

```
> library(foreign)
> viviendas=read.spss("G:/R/data/viviendas.sav",
use.value.labels=TRUE)
> attach(viviendas)
```

A continuación exploramos las variables del fichero y su naturaleza.

```
> summary(viviendas)
         Length Class  Mode
SC       2440   -none- character
barrio   2440   factor numeric
valterr  2440   -none- numeric
valmejor 2440   -none- numeric
valtot   2440   -none- numeric
precio   2440   -none- numeric
tasa     2440   -none- numeric
```


© Alfaomega - RC Libros

Observamos que la variable *SC* es de tipo carácter, la variable *barrio* es un factor y el resto de las variables son numéricas.

Realizaremos un gráfico QQ de normalidad para las variables cualitativas y lo alojaremos en una matriz gráfica 2x2 (Figura 6-43).

```
> windows()
> par(mfrow=c(2,2))
> qqnorm(valterr, xlab="valor de los terrenos")
> qqnorm(valmejor, xlab="valor de las mejoras")
> qqnorm(precio, xlab="precio de la vivienda")
> qqnorm(tasa, xlab="ratio precio real sobre precio tasado")
```

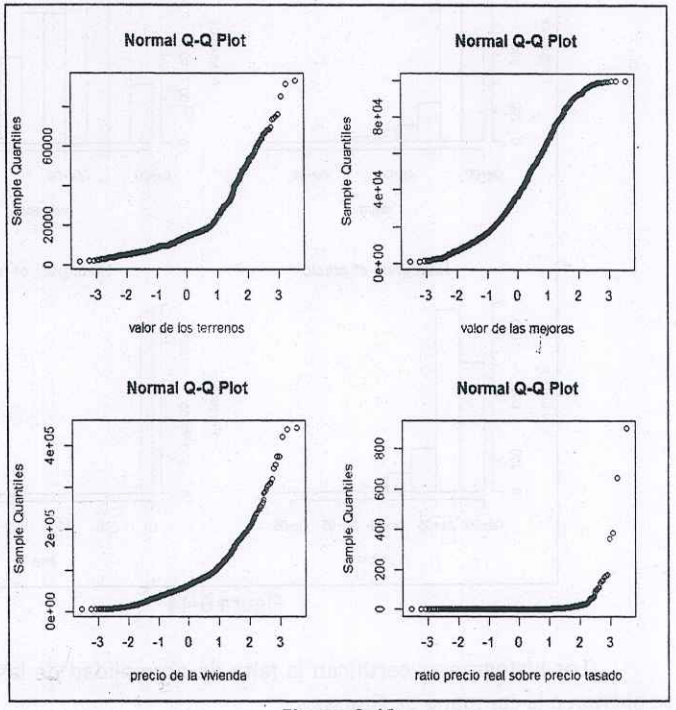


Figura 6-43

Se observa que las variables tienen dificultades para ajustarse a la normalidad ya que las nubes de puntos no se ajustan bien a la diagonal del primer cuadrante, salvo quizás el valor de las mejoras.

© Alfaomega - RC Libros

A continuación realizamos los histogramas de las cuatro variables (Figura 6-44).

```
> windows()
> par(mfrow=c(2,2))
> hist(valterr)
> hist(valmejor)
> hist(precio)
> hist(tasa)
```

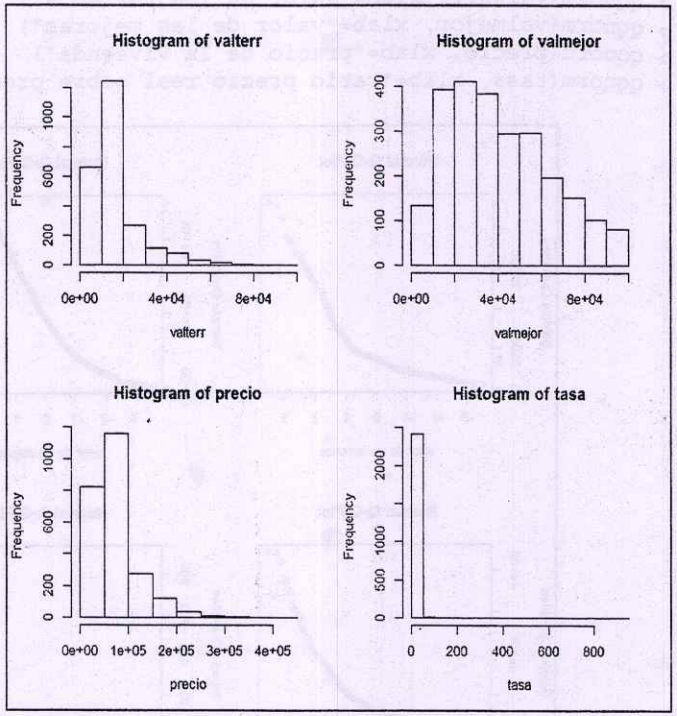


Figura 6-44

Los histogramas certifican la falta de normalidad de las variables porque no se ajustan a la campana de Gauss.

A continuación representamos las gráficas de densidad de las variables (Figura 6-45).

© Alfaomega - RC Libros


```
> windows()
> par(mfrow=c(2,2))
> plot(density(valterr))
> plot(density(valmejor))
> plot(density(precio))
> plot(density(tasa))
```

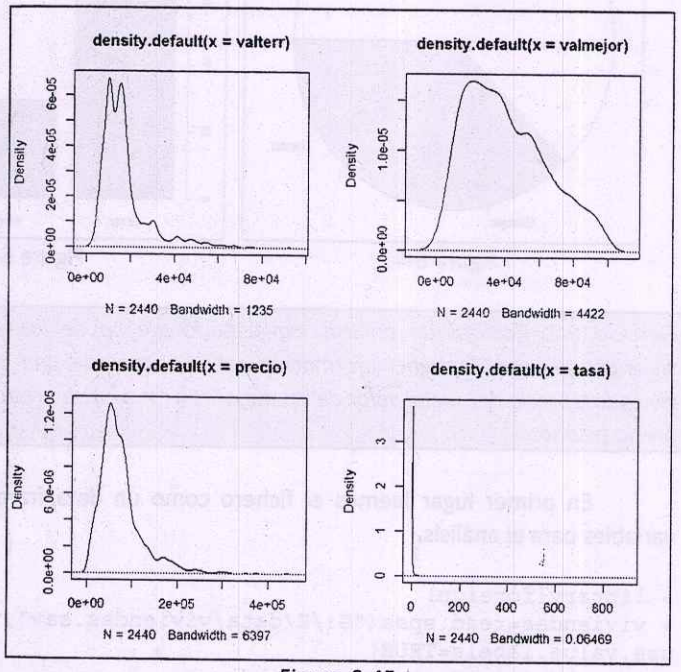


Figura 6-45

Se vuelve a certificar que las densidades no se acercan a la campana de Gauss, salvo para el valor de las mejoras.

Ahora realizamos los gráficos de caja y bigotes de las cuatro variables (Figura 6-46).

```
> windows()
> boxplot(datos)
```

Se observa la falta de simetría en las variables salvo para *valmejor*. Las variables resultan asimétricas a la izquierda, lo que provoca falta de normalidad.

© Alfaomega - RC Libros

Las variables *valor de los terrenos* y *precios de la vivienda* presentan valores atípicos superiores, resultado muy acorde con el mercado de la burbuja inmobiliaria.

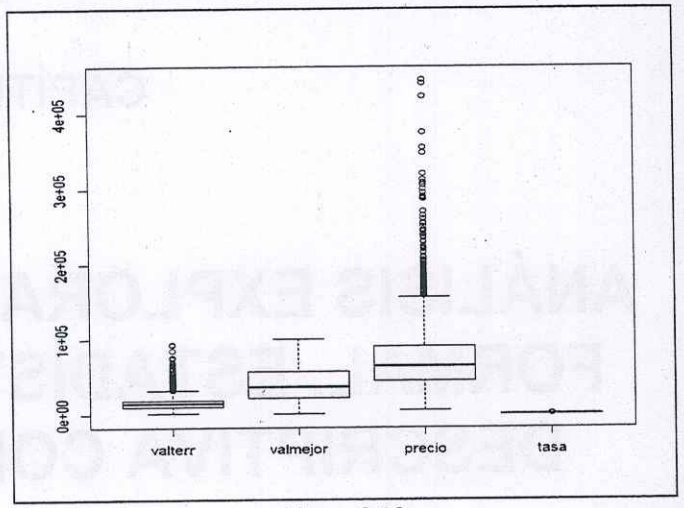


Figura 6-46

También podemos realizar el gráfico de dispersión múltiple (Figura 6-47).

```
> plot(datos)
```

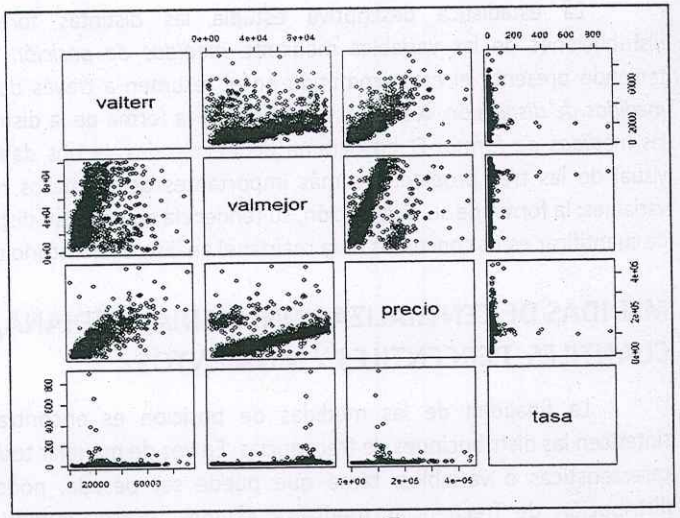


Figura 6-47

© Alfaomega - RC Libros

Ahora realizamos el gráfico de barras y el gráfico de sectores de la variable cualitativa *barrio* (Figura 6-48). Previamente construimos su tabla de frecuencias.

```
> tabla=table(barrio)
> tabla
barrio
  A   B   C   D   E   F   G
 42 319 258 467 500 372 482
> barplot(tabla)
> pie(tabla)
```

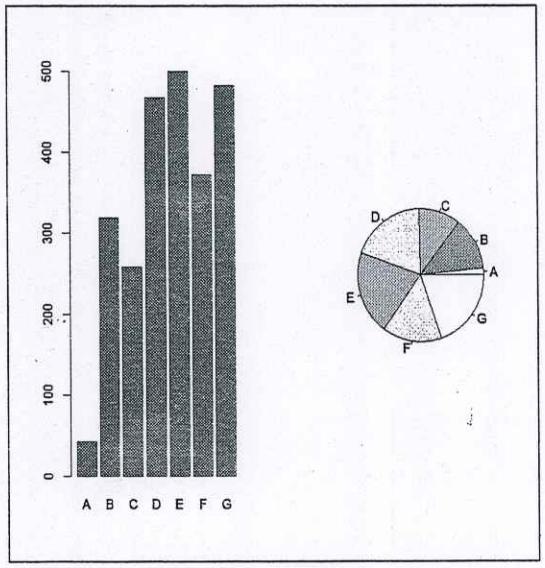


Figura 6-48

© Alfaomega - RC Libros

CAPÍTULO 7

ANÁLISIS EXPLORATORIO FORMAL. ESTADÍSTICA DESCRIPTIVA CON R

ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva estudia las distintas formas de resumir las distribuciones de las variables mediante *medidas de posición* (o de centralización), teniendo presente el error cometido en el resumen a través de las correspondientes *medidas de dispersión*. A su vez, analizaremos la forma de la distribución por medio de las *medidas de forma*. El histograma de frecuencias ya nos daba una representación visual de las tres propiedades más importantes de los datos muestrales relativos a variables: la forma de su distribución, su tendencia central y su dispersión. Ahora se trata de cuantificar estos conceptos para realizar el análisis exploratorio de datos formal.

MEDIDAS DE CENTRALIZACIÓN: MEDIA, MEDIANA, MODA, CUANTILES, PERCENTILES Y MOMENTOS

La finalidad de las medidas de posición es encontrar unos valores que sinteticen las distribuciones de frecuencias. En vez de manejar todos los datos sobre las características o variables, tarea que puede ser pesada, podemos caracterizar su distribución de frecuencias mediante algunos valores numéricos, eligiendo como resumen de los datos un valor central alrededor del cual se encuentran distribuidos los

© Alfaomega - RC Libros

valores de la variable. El valor de la variable elegido para representar a una distribución se llama promedio o medida de posición, y es un valor representativo de todos los valores que toma la variable. Debe hallarse entre el mayor y el menor valor de la variable.

Pero estas medidas de posición de una distribución de frecuencias han de cumplir determinadas condiciones para que sean verdaderamente representativas de la variable a la que resumen. Toda síntesis de una distribución se considerará como operativa si intervienen en su determinación todos y cada uno de los valores de la distribución, siendo única para cada distribución de frecuencias, y siendo siempre calculable y de fácil obtención. A continuación, se hace una relación de las medidas de posición más comunes utilizadas en estadística.

Media aritmética: se define como la suma de todos los valores de la distribución dividida por el número total de datos. La expresión matemática que representa la media aritmética coincide con el momento de primer orden respecto al origen. Pero esto solo es válido en el supuesto más sencillo, en el que los datos de la variable están sin agrupar. En el caso de que tuviésemos una distribución con datos agrupados en intervalos, los valores individuales de la variable serían desconocidos y, por tanto, no podríamos utilizar la fórmula anterior. En este supuesto, los datos estarán agrupados en clases, y se postula la hipótesis de que el punto medio del intervalo de clase (marca de clase) representa adecuadamente el valor medio de dicha clase, por lo que aplicaríamos la fórmula original de la media simple para dichos valores. En el caso de que la variable presente valores anormalmente extremos, estos pueden distorsionar la media aritmética, haciéndola incluso poco representativa. A los estadísticos que no son afectados por los valores extremos de la muestra, se les denomina estadísticos robustos. La media no es un estadístico robusto. Como veremos posteriormente, este inconveniente no lo posee la mediana. Si la distribución de frecuencias es (x_i, n_i), siendo x_i los valores de la variable o las marcas de clase, y siendo n_i las frecuencias absolutas, la media aritmética, que representaremos por $\overline{x}$, se define como sigue:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i$$

Media aritmética ponderada: caso particular de la media aritmética, que aparece cuando se otorga a cada valor de la variable x_i una ponderación o peso w_i, distinto de la frecuencia o repetición n_i. En este caso, no todos los valores de la distribución intervienen con el mismo peso en el cálculo de la media. La cuantía de dichos pesos define la importancia de cada valor de la distribución en el cálculo de la media. Si la distribución de frecuencias es (x_i, n_i, w_i), siendo x_i los valores de la variable o las marcas

© Alfaomega - RC Libros

de clase, n_i las frecuencias absolutas, y w_i los pesos o ponderaciones, la media aritmética ponderada, que representaremos como W, se define como sigue:

$$W = \frac{\sum_{i=1}^{k} x_i n_i w_i}{\sum_{i=1}^{k} n_i w_i}$$

Media geométrica: sea una distribución de frecuencias (x_i, n_i). La media geométrica, que representaremos como G, se define como la raíz N-ésima del producto de los N valores de la distribución.

$$G = \sqrt[N]{x_1^{n1} x_2^{n2} \ldots x_k^{nk}}$$

El empleo más frecuente de la media geométrica es el de promediar variables tales como porcentajes, tasas, números índices, etc.; es decir, se usa en los casos en los que se supone que la variable presenta variaciones acumulativas. Además, cuando la variable toma al menos un valor $x_i = 0$, entonces G se anula, y si la variable toma valores negativos se puede presentar una gama de casos particulares en los que tampoco queda determinada debido al problema de las raíces de índice par de números negativos.

Media armónica: sea una distribución de frecuencias (x_i, n_i). La media armónica, que representaremos como H, se define como sigue:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i} n_i}$$

Obsérvese que la inversa de la media armónica es la media aritmética de los inversos de los valores de la variable. No es aconsejable en distribuciones de variables con valores pequeños. Se suele utilizar para promediar variables tales como productividades, velocidades, tiempos, rendimientos, tipos de cambio, etc.

Media cuadrática: sea una distribución de frecuencias (x_i, n_i). La media cuadrática, que representaremos como C, se define como sigue:

© Alfaomega - RC Libros

$$C=\sqrt{\frac{n_1x_1^2+n_2x_2^2+\ldots+n_kx_k^2}{N}}$$

La media cuadrática es la raíz cuadrada de la media aritmética de los cuadrados de los valores de la variable.

Mediana: se define como mediana el valor de la distribución, suponiendo que esta esté ordenada de menor a mayor, que deja a su izquierda y a su derecha la misma frecuencia de observaciones, es decir, el valor de la variable que ocupa el lugar central, supuesto un número impar de datos. Si el número de datos fuese par, puede decirse que hay dos valores medianos, y se toma como mediana la media aritmética entre ellos. También se podría definir la mediana como aquel valor de la distribución cuya frecuencia absoluta acumulada es $N/2$ ($N=\sum n_i$). Para distribuciones agrupadas en intervalos, y suponiendo que todos los valores comprendidos dentro del intervalo mediano $[L_{i-1}, L_i]$ se encuentran distribuidos uniformemente a lo largo de él, puede calcularse la mediana, *Me*, mediante la expresión:

$$M_e=L_{i-1}+\frac{\frac{N}{2}-N_{i-1}}{n_i}c_i$$

donde $N_{i-1}<N/2<N_i$ y $[L_{i-1}, L_i]$ es el intervalo siguiente al que contiene a $N/2$, siendo c_i su amplitud.

Como ventaja de la mediana tenemos que no influyen en ella los valores extremos (estadístico robusto). La mediana tiene gran utilidad en los gráficos de control de procesos.

Moda: la moda es el valor de la variable que más veces se repite, y en consecuencia, en una distribución de frecuencias, es el valor de la variable que viene afectado por la máxima frecuencia de la distribución. En distribuciones no agrupadas en intervalos, se observa la columna de las frecuencias absolutas, y el valor de la distribución al que corresponde la mayor frecuencia será la moda. A veces aparecen distribuciones de variables con más de una moda (bimodales, trimodales, etc.), e incluso distribuciones de frecuencias que presentan una moda absoluta y una relativa.

En distribuciones agrupadas en intervalos de la misma amplitud, realmente no tendremos un valor modal, sino un intervalo modal (el intervalo que presenta la mayor

© Alfaomega - RC Libros

frecuencia $[L_{i-1}, L_i]$). Podemos tomar como valor modal el extremo inferior del intervalo ($M_0 = L_{i-1}$) o el extremo superior ($M_0 = L_i$), o bien hacer que la moda sea igual a la marca de clase del intervalo modal ($M_0 = x_i$). En caso de que todos los valores del intervalo modal estén distribuidos uniformemente dentro de él, la moda estará más cerca de aquel intervalo contiguo cuya frecuencia sea mayor, siendo las distancias de la moda, M_0, a los intervalos contiguos inversamente proporcionales a las frecuencias de dichos intervalos. La moda puede calcularse mediante la expresión:

$$M_0=L_{i-1}+\frac{n_{i+1}}{n_{i-1}+n_{i+1}}c_i$$

Para intervalos de distinta amplitud, realmente las densidades de frecuencias nos dan el número de valores que hay en cada unidad de intervalo para cada intervalo. La mayor densidad de frecuencia, ahora sí, nos determina el intervalo modal $[L_{i-1}, L_i]$, calculándose la moda mediante la expresión:

$$M_0=L_{i-1}+\frac{d_{i+1}}{d_{i-1}+d_{i+1}}c_i \text{ con } d_i=n_i/c_i$$

Cabe señalar que la moda es la medida más representativa en caso de distribuciones de variables en escala nominal.

Esto es debido a que las distribuciones de este tipo presentan datos no susceptibles de ordenación, de tal forma que para estas distribuciones no es posible realizar operaciones elementales con sus observaciones. La moda se emplea sobre todo cuando los valores de la variable presentan una gran concentración hacia un valor determinado. Solo se utilizará en distribuciones de gran frecuencia total.

Cuantiles: se definen los cuantiles de orden k como los valores de la distribución, suponiendo que esta esté ordenada de menor a mayor, que la dividen en k partes con la misma frecuencia de observaciones. Por lo tanto, existirán $r=k-1$ cuantiles de orden k. El primer cuantil de orden k deja a su izquierda la fracción $1/k$ de frecuencia de observaciones. El segundo cuantil de orden k deja a su izquierda la fracción $2/k$ de frecuencia de observaciones. El r-ésimo cuantil de orden k deja a su izquierda la fracción r/k ($100r/k$ por ciento) de frecuencia de observaciones, y a su derecha deja la fracción $1-r/k$ ($100(k-r)/k$ por ciento) de frecuencia de observaciones. El r-ésimo cuantil de orden k suele designarse como $Q_{r,k}$.

© Alfaomega - RC Libros

Para el caso más general, en que la distribución esté agrupada en intervalos, el *r*-ésimo cuantil de orden *k* tomará el valor:

$$Q_{r,k} = L_{i-1} + \frac{\frac{rN}{k} - N_{i-1}}{n_i} c_i$$

donde $N_{i-1}<rN/k<N_i$ y $[L_{i-1}, L_i]$ es el intervalo siguiente al que contiene a rN/k, siendo c_i su amplitud.

Para k=4 y r=1,2,3, tendremos los 3 cuartiles. Para k=10 y r=1,2,...,9, tendremos los 9 deciles, y para k=100 y r=1,2,...,99, tendremos los 99 percentiles. Los *cuartiles* son, por tanto, los tres valores de la distribución que la dividen en 4 partes iguales, es decir, en 4 intervalos dentro de cada cual está incluido el 25% de los valores de la distribución. Los *deciles* son los 9 puntos que dividen la distribución en 10 partes, tales que dentro de cada una está incluido el 10% de los valores de la distribución. Los *percentiles* son los 99 puntos que dividen la distribución en 100 partes, tales que dentro de cada una está incluido el 1% de los valores de la distribución.

Para distribuciones sin agrupar en intervalos, el primer cuartil, $Q_{1,4}$, es el valor de la distribución que ocupa el lugar $N/4$; el segundo cuartil, $Q_{2,4}$, es el valor de la distribución que ocupa el lugar $2N/4$; el tercer cuartil, $Q_{3,4}$, es el valor de la distribución que ocupa el lugar $3N/4$; el *r*-ésimo decil, $Q_{r,10}$, es el valor de la distribución que ocupa el lugar $rN/10$ (r=1,2,...,9); y el *r*-ésimo percentil, $Q_{r,100}$, es el valor de la distribución que ocupa el lugar $rN/100$ (r=1,2,...,99).

Momentos potenciales: entre los valores que caracterizan una distribución de frecuencias, tenemos los momentos potenciales. Dos distribuciones que tienen sus momentos iguales son iguales, y serán más parecidas cuanto más próximos sean sus momentos. Los momentos suelen considerarse respecto del origen y respecto de la media.

El momento de orden *r* respecto del origen se define como:

$$a_r = \frac{1}{N}\sum_{i=1}^{k} x_i^r n_i$$

El momento de orden *r* respecto de la media se define como:

© Alfaomega - RC Libros

$$m_r = \frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^r n_i$$

Existe una fórmula general que relaciona el momento de orden *r* centrado en la media m_r y los momentos centrados en el origen $a_r, a_{r-1}, \ldots, a_1$:

$$m_r = a_r - \binom{r}{1} a_{r-1} a_1 + \binom{r}{2} a_{r-2} a_1^2 - \cdots + (-1)^r a_1^r$$

Se observa que el momento de orden 1 centrado en el origen a_1 coincide con la media aritmética de la distribución, y el momento de orden 2 centrado en la media m_2 coincide con la varianza de la distribución.

Relación entre promedios y fórmula de Foster: se demuestra que para una misma distribución de frecuencias (x_i, n_i) se verifica que *Media armónica ≤ Media geométrica ≤ Media aritmética*, es decir:

$$H \le G \le \bar{X}$$

Para relacionar todos los promedios tenemos la fórmula de Foster, que dice que la *Media de orden m* tiene la siguiente expresión:

$$M_{(m)} = \sqrt[m]{\frac{n_1 x_1^m + n_2 x_2^m + \ldots + n_k x_k^m}{N}} = \left(\frac{1}{N}\sum x_i^m n_i\right)^{1/m}$$

de modo que como casos particulares de la fórmula de Foster se obtienen la media armónica, la media geométrica, la media aritmética y la media cuadrática haciendo m = -1, m = 0, m = 1 y m = 2 respectivamente. Podemos escribir entonces que:

$$M_{(-1)} = H,\ M_{(0)} = G,\ M_{(1)} = \bar{X},\ M_{(2)} = C$$

La relación entre promedios puede ampliarse introduciendo la media cuadrática *C* como sigue:

$$H \le G \le \bar{X} \le C$$

MEDIDAS DE DISPERSIÓN

Las medidas de dispersión permiten calcular la representatividad de una medida de posición, para lo cual será preciso cuantificar la distancia de los diferentes

© Alfaomega - RC Libros

valores de la distribución respecto a dicha medida. A tal distancia es a lo que, en términos estadísticos, denominaremos variabilidad o dispersión de la distribución. Las medidas de dispersión tienen como finalidad estudiar hasta qué punto, para una determinada distribución de frecuencias, las medidas de tendencia central o de posición son representativas como síntesis de toda la información de la distribución. Medir la representatividad de una medida de posición equivale a cuantificar la separación de los valores de la distribución respecto a dicha medida. Por ejemplo, si queremos estudiar en qué grado una media aritmética nos marca una tendencia central generalizable del comportamiento de todos los elementos del conjunto estudiado, tendremos que fijarnos en la separación o desviación de cada valor respecto a la media. Si todos los valores están cercanos al valor medio, este será representativo de ellos.

A la mayor o menor separación de los valores de una distribución respecto de otro, que se pretende que sea su síntesis, se le llama dispersión o variabilidad. Será, pues, tanto más representativa la media aritmética de una variable cuanto más agrupados en torno a ella estén los valores promediados; por el contrario, será tanto más rechazable, por no ser representativa, cuanta mayor dispersión exista de los valores de la variable respecto a la media.

Resulta pues necesario para completar la información que pueda deducirse de una medida de posición o centralización, acompañarla de uno o varios coeficientes que nos midan el grado de dispersión de la distribución de la variable respecto de esa medida de centralización. Estos coeficientes son los que llamamos medidas de dispersión. Inicialmente, se distingue entre medidas de dispersión absolutas y relativas, entendiéndose por relativas las que no dependen de las unidades de medida. Posteriormente, se clasifican las medidas absolutas y relativas según sean medidas referidas a promedios o no lo sean.

Entre las medidas de dispersión absolutas no referidas a promedios tenemos el *recorrido,* o diferencia entre el mayor valor y el menor valor de una distribución, y el *recorrido intercuartílico,* o diferencia existente entre el tercer cuartil y el primero. Entre las medidas de dispersión relativas no referidas a promedios tenemos el *coeficiente de apertura,* o cociente entre el mayor valor y el menor valor de una distribución, y el *recorrido relativo,* o cociente entre el recorrido y la media, así como el *recorrido semintercuartílico,* o cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil.

© Alfaomega - RC Libros

Entre las medidas de dispersión absolutas referidas a promedios tenemos las *desviaciones medias*, la *varianza* y la *desviación típica*. Estas medidas de dispersión involucran a los promedios, y permiten medir el error que cometemos utilizando el promedio en cuestión como resumen de los datos. Como medida de dispersión más simple relativa a la medida de posición *P*, podríamos considerar las desviaciones de cada valor al promedio y promediar estas desviaciones, es decir, considerar el valor $D = \sum (x_i - P)n_i/N$, $i=1...k$. Pero esto, que sería lo primero que se nos ocurriría, tiene como grave inconveniente las posibles compensaciones de las desviaciones positivas con las negativas al efectuar la suma, pudiendo obtenerse una medida pequeña siendo la dispersión grande. Para solucionar este inconveniente se consideran los valores absolutos de las desviaciones, o bien se elevan estas al cuadrado. A continuación, se definen las medidas de dispersión más interesantes, entre las que tenemos:

Desviaciones medias: para medir la eficacia de la media se considera la *desviación media respecto de la media aritmética*, que se define como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la media aritmética, y cuya expresión es la siguiente:

$$D_m = \frac{1}{N}\sum_{i=1}^{k} | x_j - \overline{x} | \; n_i$$

Para medir la eficacia de la mediana, M_e, suele considerarse la *desviación media respecto de la mediana*, que se define como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la mediana, y cuya expresión es la siguiente:

$$D_{Me} = \frac{1}{N}\sum_{i=1}^{k} | x_j - M_e | \; n_i$$

Varianza, cuasivarianza, desviación típica y error estándar: de todas las medidas de dispersión absolutas respecto a la media aritmética, la varianza y su raíz cuadrada (la desviación típica), son las más importantes. Si en vez de considerar los valores absolutos de las desviaciones respecto del promedio consideramos sus cuadrados, surge una nueva medida de dispersión denominada varianza, y que definimos como la media aritmética de los cuadrados de las desviaciones de los valores de la variable a la media aritmética, es decir, el momento de segundo orden respecto a la media aritmética. Se define mediante la expresión:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} (x_j - \overline{x})^2 n_i$$

© Alfaomega - RC Libros

Como propiedades más importantes de la varianza tenemos que nunca puede ser negativa, que es igual al momento de segundo orden respecto al origen menos el de primer orden elevado al cuadrado, que si en la distribución de frecuencias sumamos a todos los valores de la variable una constante la varianza no varía (un cambio de origen en la variable no afecta a la varianza), y que al multiplicar los valores de una distribución de frecuencias por una constante *k* la varianza queda multiplicada por el cuadrado de la constante.

Así como las desviaciones medias vienen expresadas en las mismas unidades de medida que la distribución, la varianza no, ya que vendrá dada en las unidades correspondientes, pero elevadas al cuadrado. Esto dificulta su interpretación, y hace necesario definir la desviación típica o desviación estándar. La desviación típica es la raíz cuadrada, con signo positivo, de la varianza, con lo que su expresion será:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^2 n_i}$$

Al ser la raíz cuadrada de la varianza, vendrá expresada en las mismas unidades de medida que la distribución, lo cual la hace más apta como medida de dispersión.

Un estadístico muy utilizado como medida de dispersión, sobre todo debido a sus propiedades muestrales, es la cuasivarianza, cuya expresión es:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}(x_j - \bar{x})^2 n_i$$

También se considera la cuasidesviación típica, cuya expresión es:

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{k}(x_j - \bar{x})^2 n_i}$$

Otro estadístico muy utilizado como medida de dispersión, también por sus propiedades muestrales, es el error estándar, cuyo valor es $e=S/\sqrt{n}$.

Siempre se cumple que $D_{Me} < D_m < \sigma$.

Entre las medidas de dispersión absolutas referentes a promedios, podríamos haber definido también la desviación media respecto a la moda y las desviaciones cuadráticas respecto a la mediana y a la moda, que vienen dadas en las mismas unidades de medida que la distribución, y que marcan la representatividad de los promedios con los que se relacionan.

© Alfaomega - RC Libros

Entre las medidas de dispersión relativas (valores adimensionales que no se ven afectados por las unidades de medida y que siempre se concretan en forma de cociente) utilizadas para comparar medidas de posición o promedios, tenemos el *índice de dispersión respecto a la mediana* y el *coeficiente de variación de Pearson*.

Coeficiente de variación de Pearson: se usa para resolver el problema de comparación de medias aritméticas de varias distribuciones que pueden venir, en general, en unidades diferentes.

Se define como la relación por cociente entre la desviación típica y la media aritmética ($V = \sigma / \bar{x}$). Evidentemente, a menor coeficiente de variación (V), mejor es la media. Observamos que al efectuar el cociente eliminamos las unidades, y por tanto V es adimensional. Por otra parte, V representa el número de veces que σ contiene a $\bar{x}$, y está claro que cuanto mayor sea V más veces contendrá σ a $\bar{x}$, luego, relativamente, a mayor valor de V menor representatividad de $\bar{x}$. Este coeficiente también se suele expresar en tantos por ciento como $V = 100\ (\sigma / \bar{x})$.

Como tanto en el cálculo de σ como en el cálculo de $\bar{x}$ han intervenido todos los valores de la distribución, V presenta la garantía, frente a otros coeficientes, de que utiliza toda la información de la distribución. La cota inferior de V es cero, y el único caso problemático se presenta cuando $\bar{x} = 0$, lo que haría que V tendiera a infinito.

Índice de dispersión respecto a la mediana: se usa para resolver el problema de comparación de medianas de varias distribuciones, que pueden venir, en general, en unidades diferentes. Se define como la relación por cociente entre la desviación media respecto de la mediana y la mediana aritmética $V_{Me} = D_{Me} / M_e$. Evidentemente, a menor índice de dispersión mejor es la mediana.

MEDIDAS DE FORMA

Una vez iniciado el análisis estadístico de sintetización de la información, para lo cual hemos estudiado las medidas de posición y dispersión de la distribución de una variable, necesitamos conocer más sobre el comportamiento de la misma. No podemos basar nuestras conclusiones únicamente en expresiones que vengan dadas en términos de medidas de posición y dispersión. Si bien intentamos globalizar el comportamiento del colectivo que sea objeto de nuestro estudio, para lo cual las medidas de posición son nuestro mejor instrumento, no debemos proceder a una interpretación que implique un comportamiento de todos los elementos del colectivo uniformemente constante e igual a la medida de posición en cuestión con un error dado por la correspondiente medida de dispersión. Este error o disparidad se hace más ostensible al analizar la representación

© Alfaomega - RC Libros

gráfica de la distribución. Pues bien, las medidas de forma de una distribución se basan en su representación gráfica, sin llegar a realizar la misma.

Las medidas de forma se clasifican en *medidas de asimetría* y *medidas de curtosis* o *apuntamiento*.

Las medidas de asimetría tienen como finalidad el elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución, sin necesidad de llevar a cabo su representación gráfica. Supongamos que hemos representado gráficamente una distribución de frecuencias. Si trazamos una perpendicular al eje de abscisas por $\bar{x}$ y tomamos esta perpendicular como eje de simetría, diremos que una distribución es simétrica si existe el mismo número de valores a ambos lados de dicho eje, equidistantes de $\bar{x}$ dos a dos, y tales que cada par de valores equidistantes de $\bar{x}$ tengan la misma frecuencia. En caso contrario, las distribuciones serán asimétricas.

Las medidas de curtosis estudian la distribución de frecuencias en la zona central de la misma. La mayor o menor concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada. Por esta razón, a las medidas de curtosis se les llama también de apuntamiento o concentración central. Las medidas de curtosis se aplican a distribuciones campaniformes, es decir, unimodales simétricas o con ligera asimetría. Para estudiar la curtosis de una distribución es necesario definir previamente una distribución tipo, que vamos a tomar como modelo de referencia. Esta distribución es la normal, que corresponde a fenómenos muy corrientes en la naturaleza, y cuya representación gráfica es una campana de Gauss.

Tomando la normal como referencia, diremos que una distribución puede ser más apuntada que la normal (es decir, *leptocúrtica*), o menos apuntada (es decir, *platicúrtica*). A la distribución normal, desde el punto de vista de la curtosis, se le llama *mesocúrtica*. Con la curtosis se estudia la deformación, en sentido vertical, respecto a la normal, de una distribución. A continuación, se definen las medidas de asimetría más comunes, entre las que destacan las siguientes:

Coeficiente de asimetría de Fisher: ahora se intenta buscar una medida que recoja la simetría o asimetría de una distribución. Si la distribución es simétrica, el eje de simetría de su representación gráfica será una recta paralela al eje de ordenadas, que pasa por el punto cuya abscisa es la media aritmética. Por ello, cuando la distribución es asimétrica, referiremos los valores de la distribución a este promedio. Si una distribución es simétrica, existe el mismo número de valores a la derecha que a la

© Alfaomega - RC Libros

izquierda de $\bar{x}$, y por tanto el mismo número de desviaciones con signo positivo que con signo negativo, siendo la suma de desviaciones positivas igual a la suma de las negativas. Podemos partir, pues, de las desviaciones ($x_i - \bar{x}$) elevadas a una potencia impar para no perder los signos de las desviaciones. Lo más sencillo sería tomar como medida de asimetría el promedio de estas desviaciones, elevadas a la potencia impar más simple (que es tres); es decir, tomaríamos como medida de asimetría el momento de orden tres centrado en la media. Pero, de hacer esto, esta medida vendría expresada en las mismas unidades que las de la variable pero elevadas al cubo, por lo que no es invariante ante un cambio de escala.

Para conseguir un indicador adimensional, debemos dividir la expresión anterior por una cantidad que venga en sus mismas unidades de medida. Esta cantidad es el cubo de la desviación típica. De este modo obtenemos el coeficiente de asimetría de R. A. Fisher, cuya expresión es:

$$g_1 = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^3 n_i}{\left(\frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^2 n_i\right)^{3/2}}$$

Si $g_1=0$, la distribución es simétrica; si $g_1>0$, la distribución es asimétrica positiva (a derechas); y si $g_1<0$, la distribución es asimétrica negativa (a izquierdas). La distribución es asimétrica a derechas o positiva cuando la suma de las desviaciones positivas de sus valores respecto de la media es mayor que la suma de las desviaciones con signo negativo (la gráfica de la distribución tiene más densidad a la derecha de la media). En caso contrario, la distribución es asimétrica a la izquierda o negativa.

Coeficiente de asimetría de Fisher estandarizado: para $N>150$, el coeficiente de asimetría es asintóticamente normal de media cero y varianza $6/N$. Este hecho nos lleva a considerar el coeficiente de asimetría estandarizado, cuya expresión es:

$$g_s = \frac{g_1}{\sqrt{\frac{6}{N}}}$$

Este coeficiente es asintóticamente normal (0,1).

© Alfaomega - RC Libros

Coeficiente de asimetría de Pearson: Karl Pearson propuso para distribuciones campaniformes, unimodales y moderadamente asimétricas, el coeficiente definido como $Ap = (\bar{x} - Mo) / \sigma$, donde Mo es la moda. Como en una distribución campaniforme simétrica $\bar{x} = Mo = Me$, si la distribución es asimétrica positiva o a derechas, $\bar{x}$ se desplaza a la derecha de la moda, y por tanto, $\bar{x} - Mo > 0$. En el caso de distribución asimétrica negativa, la media se sitúa por debajo de Mo, por lo que el valor $\bar{x} - Mo < 0$. La desviación típica que aparece en el denominador no modifica el signo de la diferencia $\bar{x} - Mo$, y sirve para eliminar las unidades de medida de dicha diferencia. Así tendremos que si $Ap = 0$ la distribución es simétrica, si $Ap>0$ la distribución es asimétrica positiva, y si $Ap<0$ la distribución es asimétrica negativa. También Pearson comprobó empíricamente para este tipo de distribuciones que se cumple $3(\bar{x} - Me) \approx \bar{x} - Mo$ (la mediana siempre se sitúa entre la media y la moda en las distribuciones moderadamente asimétricas). Por esta razón, algunos autores utilizan como coeficiente de asimetría de Pearson el valor $Ap \approx 3(\bar{x} - Me) / \sigma$.

Coeficiente de asimetría de Bowley: está basado en la posición de los cuartiles y la mediana, y viene dado por la expresión $Ab = (C3 + C1 - 2Me) / (C3 + C1)$. Se cumple que si $Ab=0$ la distribución es simétrica, si $Ab>0$ la distribución es asimétrica positiva, y si $Ab<0$ la distribución es asimétrica negativa. $C1$ y $C3$ son el primer y tercer cuartil respectivamente.

Coeficiente absoluto de asimetría: está basado también en la posición de los cuartiles y la mediana, y viene dado por la expresión:

$$A = [(C3 - C2) - (C2 - C1)] / S = (C3 + C1 - 2C2) / S = C3 + C1 - 2Me / S$$

Si $A = 0$ la distribución es simétrica, si $A>0$ la distribución es asimétrica positiva, y si $A<0$ la distribución es asimétrica negativa. $C1$, $C2$ y $C3$ son los cuartiles de la distribución.

Una vez presentadas las medidas de asimetría, a continuación se definen las medidas de curtosis más comunes, entre las que destacan las siguientes:

Coeficiente de curtosis: en la distribución normal, se verifica que $m_4 = 3\sigma^4$, siendo m_4 el momento de orden 4 respecto a la media, y σ la desviación típica. Si consideramos la expresión $g_2 = m_4/\sigma^4 - 3$, su valor será cero para la distribución normal. Por ello, como coeficiente de apuntamiento o curtosis se utiliza la expresión:

© Alfaomega - RC Libros

$$g_2 = \frac{m_4}{\sigma^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^4 n_i}{\left(\frac{1}{N}\sum_{i=1}^{k}(x_j - \bar{x})^2 n_i\right)^2} - 3$$

Una distribución es *mesocúrtica* (apuntamiento igual al de la normal) cuando $g^2 = 0$, es *leptocúrtica* (apuntamiento mayor que el de la normal) si $g^2 > 0$, y es *platicúrtica* (apuntamiento menor que el de la normal) si $g^2 < 0$.

Coeficiente de curtosis estandarizado: para $N>150$, el coeficiente de curtosis es asintóticamente normal de media cero y varianza $24/N$. Este hecho nos lleva a considerar el coeficente de curtosis estandarizado, cuya expresión es:

$$g_{ks} = \frac{g_2}{\sqrt{\frac{6}{N}}}$$

Este coeficiente es asintóticamente normal (0,1).

COEFICIENTE DE CORRELACIÓN LINEAL ENTRE DOS VARIABLES. REGRESIÓN SIMPLE MÍNIMO CUADRÁTICA

Se llama correlación al grado de dependencia mutua entre dos variables. El *coeficiente de correlación* intenta medir la intensidad con que dos variables están relacionadas. Este concepto está directamente relacionado con el concepto de curva de regresión. Mediante la regresión simple mínimo cuadrática, se expresa la estructura funcional de la relación existente entre dos variables, ajustando la nube de puntos dada por los pares de valores de las dos variables a una curva de la forma mejor posible (minimizando la varianza del error). El ajuste será de la forma $Y=f(x)+e$ o $X=f(Y)+e$, donde e denota el error cometido cuya varianza debe ser mínima. El coeficiente de correlación mide la calidad de ese ajuste.

Cuando la curva es una recta, la regresión se llama lineal, y en este caso el coeficiente de correlación se llama *coeficiente de correlación lineal,* y mide el grado de asociación lineal que existe entre las variables. El ajuste será de la forma $Y=a+bX+e$ (*recta de regresión de Y sobre X*), o $X=c+dY+e$ (*recta de regresión de X sobre Y*), donde $a = \bar{y} - \bar{x}\, S_{xy}/S_x^2$, $b = S_{xy}/S_x^2$, $c = \bar{x} - \bar{y}\, S_{xy}/S_y^2$, y $d = S_{xy}/S_y^2$.

© Alfaomega - RC Libros

A los parámetros *a* y *b* se les denomina *coeficientes de regresión de Y sobre X*, y a los parámetros *c* y *d* se les llama *coeficientes de regresión de X sobre Y*. También se pueden expresar las rectas de regresión de *Y* sobre *X* y *X* sobre *Y* respectivamente de la forma $y-\bar{y}=(x-\bar{x})S_{xy}/S_x^2$ y $(x-\bar{x})=(y-\bar{y})S_{xy}/S_y^2$.

Si suponemos el ajuste de la forma *Y=a+bX+e* (*recta de regresión de Y sobre X*), el criterio de mínimos cuadrados considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error *e*, lo que es equivalente a minimizar:

$$S(a,b)=\sum_{i,j} e_{ij}^2=\sum_{i,j}(y_j-(a+bx_i))^2$$

Derivando respecto de los parámetros *a* y *b* e igualando a cero tenemos:

$$\left.\begin{aligned}\frac{\partial S(a,b)}{\partial a}&=2\sum_{i,j}(y_j-(a+bx_i))(-1)=0\\ \frac{\partial S(a,b)}{\partial b}&=2\sum_{i,j}(y_j-(a+bx_i))(-x_i)=0\end{aligned}\right\}\Rightarrow\left\{\begin{aligned}&\sum_j y_j=na+b\sum_i x_i\\ &\sum_{i,j} y_j x_i=a\sum_i x_i+b\sum_i x_i^2\end{aligned}\right.$$

Las soluciones *a* y *b* de este *sistema de ecuaciones normales* son $a=\bar{y}-\bar{x}S_{xy}/S_x^2$ y $b=S_{xy}/S_x^2$, lo que hace que la recta de regresión de Y sobre X sea:

$$Y=a+bX=\bar{y}-\bar{x}S_{xy}/S_x^2+S_{xy}/S_x^2\,x\Rightarrow y-\bar{y}=(x-\bar{x})S_{xy}/S_x^2$$

Razonando de forma similar, se obtiene la recta de regresión de *X* sobre *Y*.

La expresión del coeficiente de correlación lineal entre las variables *X* e *Y* viene dada por la expresión:

$$r=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i-\bar{x})\,(y_j-\bar{y})\,n_{ij}}{\sqrt{\sum_{i=1}^{h}(x_i-\bar{x})^2\sum_{i=1}^{h}(y_i-\bar{y})^2}}$$

Si *r = 1* existe correlación perfecta positiva, y la relación funcional entre ambas variables es exacta y positiva, variando las dos en el mismo sentido (al aumentar una aumenta la otra, y al disminuir una disminuye la otra). Si *r = –1* existe correlación perfecta negativa, y la relación funcional entre ambas variables es exacta

© Alfaomega - RC Libros

y negativa, variando las dos en el sentido opuesto (al aumentar una disminuye la otra, y al disminuir una aumenta la otra). Si *r = 0* la correlación es nula, y las variables no están asociadas, siendo imposible encontrar una relación funcional entre ellas.

Si *0<r<1* la correlación es positiva, pero el grado de asociación entre las dos variables será mayor a medida que *r* se acerca más a 1, y será menor a medida que *r* se acerca más a cero. Si *–1<r<0* la correlación es negativa, pero el grado de asociación entre las dos variables será mayor a medida que *r* se acerca más a –1, y será menor a medida que *r* se acerca más a cero.

El cuadrado del coeficiente de correlación r^2, denotado en general por R^2, se denomina *coeficiente de determinación* y representa el porcentaje de variabilidad de la variable dependiente que es explicada por la regresión. Dada su definición, R^2 puede expresarse de forma general en función de la varianza de *Y* (σ_Y^2) y de la varianza residual σ_e^2 como sigue:

$$R^2=\frac{\textit{Varianza de la regresión}}{\textit{Varianza de Y}}=\frac{\sigma_Y^2-\sigma_e^2}{\sigma_Y^2}=1-\frac{\sigma_e^2}{\sigma_Y^2}=1-\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(y_i-\overbrace{(a-bx_i)}^{\hat{y}_i})^2 n_{ij}}{\sigma_Y^2}$$

En general, se trata de una medida de la bondad del ajuste por regresión. Si R^2 se aproxima a la unidad el ajuste es bueno y si R^2 se acerca a cero el ajuste es malo. Esta definición e interpretación de R^2 es válida para cualquier tipo de ajuste aunque no sea lineal.

En cuanto a la relación entre correlación e independencia, se observa que al definir el coeficiente de correlación lineal como $r = S_{xy}/(S_x\,S_y)$, si las variables son independientes estarán incorrelacionadas, ya que *r=0* debido a que S_{xy} es cero cuando hay independencia. Ahora bien, el recíproco no es necesariamente cierto, ya que dos variables pueden estar incorrelacionadas linealmente y ser dependientes, puesto que al ser *r=0*, lo único que podemos decir es que la asociación lineal es nula, pero esas variables pueden depender según otro tipo de asociación (parabólica, exponencial, etc.).

© Alfaomega - RC Libros

COEFICIENTE DE CORRELACIÓN POR RANGOS

Es muy típico considerar, sobre todo en variables cualitativas, el coeficiente de correlación entre los rangos de los valores de las variables. Se entiende por rango de un valor de una variable el lugar que ocupa dicho valor en el conjunto total de valores de la variable, suponiendo una ordenación de menor a mayor. Sean A_i y B_i las diferentes modalidades de dos variables cualitativas X e Y. Sean x_i e y_i los rangos o números de orden que le corresponden a A_i y B_i, supuestas ordenadas estas modalidades, con la escala que se determine, y de menor a mayor. Se define el coeficiente de correlación por rangos de Spearman para las variables cualitativas X e Y como el coeficiente de correlación lineal de las variables cuyos valores son x_i e y_i.

Este valor se utiliza para medir el grado de asociación de las variables cualitativas X e Y basándonos en la concordancia o discordancia de las clasificaciones por rangos de sus modalidades. El coeficiente de correlación por rangos también se utiliza para variables cuantitativas, con la aclaración de que el grado de asociación obtenido no es el de los valores de las variables, sino el de las clasificaciones por rangos de dichos valores. Este coeficiente viene dado por:

$$\rho = 1 - \frac{6\sum_i d_i^2}{N^3 - N}$$

siendo $d_i = x_i - y_i$. Este coeficiente también se denomina coeficiente de correlación ordinal, y por ser un coeficiente de correlación varía entre –1 y 1. Cuando la concordancia entre los rangos es perfecta, entonces $d_i = x_i - y_i = 0$ y $\rho = 1$. Cuando la discordancia es perfecta, $\rho = -1$. Cuando no hay ni concordancia ni discordancia, $\rho = 0$.

COVARIANZA

A veces es preciso introducir medidas que hagan referencia a la distribución conjunta de dos variables.

El *momento bidimensional de orden (r,s) centrado en el origen*, se define como:

$$a_{rs} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i^r y_j^s n_{ij}$$

© Alfaomega - RC Libros

El *momento bidimensional de orden (r,s) centrado en las medias*, se define como:

$$m_{rs} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} (x_i - \bar{x})^r (y_j - \bar{y})^s n_{ij}$$

Los momentos de la distribución marginal de X se calcularán como a_{r0} (los centrados en el origen) y m_{r0} (los centrados en la media). Concretamente a_{10} es la media de X, y m_{20} es la varianza de X. De la misma forma, los momentos de la distribución marginal de Y se calcularán como a_{0s} (los centrados en el origen) y m_{0s} (los centrados en la media). Concretamente, a_{01} es la media de Y, y m_{02} es la varianza de Y.

De entre las medidas relativas a la distribución conjunta, la de mayor interés en todo análisis estadístico es la covarianza σ_{xy}, que se define como m_{11} y que representa la variación conjunta de dos variables. Su fórmula será:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} (x_i - \bar{x})\ (y_j - \bar{y})\ n_{ij} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij} - \bar{X}\bar{Y} = a_{11} - a_{10}\, a_{01}$$

La covarianza es una medida de gran importancia por su signo, más que por el valor que pueda alcanzar; el signo indica el sentido de la variación conjunta de las variables que estamos considerando. De esta forma, si la covarianza es positiva, quiere decir que ambas variables varían en el mismo sentido alrededor de sus medias, mientras que si la variación de las mismas tiene lugar en sentido contrario, la covarianza tomará valores negativos.

En ocasiones, suele utilizarse la cuasicovarianza debido a sus propiedades muestrales. La expresión de la cuasicovarianza es:

$$\hat{S}_{xy} = \frac{1}{N-1}\sum_{i=1}^{h}\sum_{j=1}^{k} (x_i - \bar{x})\ (y_j - \bar{y})\ n_{ij}$$

DISTRIBUCIONES DE MÁS DE DOS DIMENSIONES: MATRIZ DE COVARIANZAS Y MATRIZ DE CORRELACIONES

Cuando se tiene una variable tridimensional (X,Y,Z) o enedimensional en general, también se puede realizar la descripción y análisis de las distribuciones de frecuencias subyacentes. La complejidad crece cuando el número de variables o

© Alfaomega - RC Libros

factores que se analizan simultáneamente aumenta, pero conocido el procedimiento para el caso tridimensional, su generalización al *n*-dimensional es inmediata. El análisis es igualmente válido cuando los caracteres analizados son de naturaleza cuantitativa o cualitativa.

Al igual que en las distribuciones bidimensionales, la forma más usual de representación de distribuciones tridimensionales son las tablas de correlación (referentes a caracteres cuantitativos) o las tablas de contingencia (referentes a caracteres cuantitativos). Como sucedía en el caso de la distribución bidimensional, se pretende que las frecuencias, tanto conjuntas y marginales como condicionadas, sean fácilmente localizables, y también sus respectivas distribuciones.

Un elemento esencial en el estudio de variables enedimensionales es la *matriz de covarianzas*, que resume las covarianzas para todos los posibles pares de variables de entre *n* dadas $X_1, X_2, \ldots, X_n$. Se define como:

$$C = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_{nn} \end{pmatrix}$$

donde cada σ_{ij} representa la covarianza entre X_i y X_j para todo *i,j*.

El signo de cada σ_{ij} indica el sentido de la variación conjunta de las dos variables X_i y X_j que estamos considerando. Si la covarianza es positiva, quiere decir que ambas variables varían en el mismo sentido, mientras que si la variación de las mismas tiene lugar en sentido contrario, la covarianza tomará valores negativos.

Con la matriz de covarianzas analizamos simultáneamente el sentido de la variación conjunta de todos los posibles pares de variables X_i y X_j para todo *i,j*.

Otro elemento esencial en el estudio de variables enedimensionales es la *matriz de correlaciones*, que resume las correlaciones para todos los posibles pares de variables de entre *n* dadas $X_1, X_2, \ldots, X_n$. Se define como:

© Alfaomega - RC Libros

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn} \end{pmatrix}$$

donde cada *rij* es el coeficiente de correlación entre X_i y X_j para todo *i,j*.

Si dada una serie de variables $X_1, X_2, \ldots, X_n$, se trata de estudiar el grado de dependencia simultánea entre todas ellas (o bien entre grupos de ellas), puede utilizarse la matriz de correlaciones. Si en base a la intensidad con que dependen se puede establecer una función que explique una variable mediante todas las demás, que se supone son sus causas influyentes, estamos ante un problema de regresión múltiple, que será estudiado en capítulos posteriores.

Mediante el *coeficiente de correlación lineal múltiple* se estudia el grado de asociación lineal simultánea entre todas las variables, mientras que mediante los coeficientes de correlación simples r_{ij} se mide el grado de asociación entre las variables X_i y X_j sin tener en cuenta a las demás variables.

FUNCIONES BÁSICAS DE R PARA ESTADÍSTICA DESCRIPTIVA

El módulo básico de R dispone de un grupo de funciones relativas a la estadística descriptiva. La tabla siguiente presenta este tipo de funciones.

sum(x)	*Suma de los elementos de la variable x*
prod(x)	*Producto de los elementos de la variable x*
max(x)	*Valor máximo de los elementos de la variable x*
min(x)	*Valor mínimo de los elementos de la variable x*
pmax(x,y,z..)	*Vector de valores máximos de las variables x, y, z...*
pmix(x,y,z..)	*Vector de valores mínimos de las variables x, y, z...*
cumprod(x)	*Vector de productos acumulados de los elementos de la variable x*
cumsum(x)	*Vector de sumas acumuladas de los elementos de la variable x*
cummax(x)	*Vector de máximos acumulados de los elementos de la variable x*

© Alfaomega - RC Libros

cummin(x)	*Vector de mínimos acumulados de los elementos del vector A*
range(x)	*Valores mínimo y máximo de los elementos de la variable x. range(x) = c(min(x), max(x))*
rank(x)	*Vector con los rangos de los elementos de la variable x (lugar que ocupa cada elemento en x)*
mean(x)	*Media aritmética de la variable x*
weighted.mean(x,y)	*Media aritmética ponderada de la variable x con vector de ponderaciones y*
median(x)	*Mediana de la variable x*
quantile(x,probs=0.k)	*k-ésimo cuantil de la variable x*
IQR(x)	*Rango intercuartílico de la variable x*
fivenum(x)	*Mínimo, primer cuartil, mediana, tercer cuartil y máximo de la variable x*
sd(x)	*Cuasidesviación típica de la variable x*
var(x)	*Cuasivarianza de la variable x*
cov(x)	*Cuasivarianza de la variable x*
var(x,y)	*Covarianza entre las variables x e y*
cov(x,y)	*Covarianza entre las variables x e y*
cor(x,y)	*Coeficiente de correlación entre las variables x e y*
var(M)	*Matriz de covarianzas para las columnas de las matrices de variables o data frame M*
cov(M)	*Matriz de covarianzas para las columnas de las matrices de variables o data frame M*
cor(M)	*Matriz de correlaciones para las columnas de las matrices de variables o data frame M*
cor(M, use=, method=)	*Matriz de correlaciones para las columnas de las matrices de variables o data frame M usando los métodos de Pearson, Spearman o Kendall en el argumento method. El argumento use se utiliza para el tratamiento de valores perdidos. Su valor all.obs produce un error si hay missing, su valor everything calcula las correlaciones en presencia de valores missing, su valor complete.obs elimina observaciones por lista y su valor pairwise.complete elimina observaciones según pareja.*
diff(x)	*Genera la variable primera diferencia de la variable x*
diff(x,lag=k)	*Genera la variable k-ésima diferencia de la variable x*

© Alfaomega - RC Libros

scale(x, center=TRUE, scale=TRUE)	*Estandariza el vector o matriz x*
summary(D)	*Estadísticos descriptivos de las variables del data frame D*
apply(D, función)	*Aplica la función a todas las variables del data frame D*
table(x)	*Tabla de frecuencias de la variable x*
table(x,y,z..)	*Tabla de contingencia de los factores especificados*
xtabs(~x+y+z+...)	*Tabla de contingencia de los factores especificados*

Como primer ejemplo, vamos a leer el fichero *coches.sav*, habilitamos sus variables y construimos un data frame con las variables cuantitativas para aplicarle las funciones de estadística descriptiva que engloban a todo el data frame.

```
> library(foreign)
> coches=read.spss("G:/R/data/coches.sav",use.value.labels=FALSE)
> attach(coches)
> variables=data.frame(consumo, cv, peso, motor, acel)
> summary(variables)

   consumo          cv              peso            motor           acel
 Min.   : 5.00   Min.   : 46.00   Min.   : 244.0   Min.   :  66   Min.   : 8.00
 1st Qu.: 8.00   1st Qu.: 75.21   1st Qu.: 741.2   1st Qu.:1708   1st Qu.:13.62
 Median :10.31   Median : 94.50   Median : 936.5   Median :2434   Median :15.50
 Mean   :11.27   Mean   :104.45   Mean   : 989.5   Mean   :3180   Mean   :15.50
 3rd Qu.:14.00   3rd Qu.:128.00   3rd Qu.:1203.8   3rd Qu.:4806   3rd Qu.:17.07
 Max.   :26.00   Max.   :230.00   Max.   :1713.0   Max.   :7456   Max.   :24.80
```

A continuación calculamos la media de todas las variables del data frame.

```
> sapply(variables,mean)

  consumo        cv      peso     motor      acel
 11.27362  104.44688  989.51232 3179.73399  15.49507
```

A continuación aplicamos la function *fivenum* a todas las variables del data frame, es decir, calculamos mínimo (1), primer cuartil (2), mediana (3), tercer cuartil (4) y máximo (5).

```
> sapply(variables,fivenum)
     consumo    cv   peso  motor acel
[1,]  5.0000  46.0  244.0   66.0  8.0
[2,]  8.0000  75.0  741.0 1704.0 13.6
[3,] 10.3061  94.5  936.5 2433.5 15.5
[4,] 14.0000 129.0 1204.0 4949.0 17.1
[5,] 26.0000 230.0 1713.0 7456.0 24.8
```

© Alfaomega - RC Libros

A continuación calculamos la matriz de correlaciones y la matriz de covarianzas de las variables del data frame.

```
> cor(variables)
          consumo        cv      peso     motor     acel
consumo 1.0000000 0.8399707 0.8404614 0.8396855 -0.5004536
cv      0.8399707 1.0000000 0.8597706 0.8973697 -0.7022883
peso    0.8404614 0.8597706 1.0000000 0.9333371 -0.4151784
motor   0.8396855 0.8973697 0.9333371 1.0000000 -0.5446238
acel   -0.5004536 -0.7022883 -0.4151784 -0.5446238  1.0000000
```

```
> cov(variables)
          consumo        cv      peso     motor     acel
consumo 15.587618 127.38870 939.9813  5715.405 -5.573838
cv     127.388701 1475.54956 9355.5868 59427.683 -76.101453
peso   939.981275 9355.58682 80245.9492 455817.077 -331.777470
motor 5715.404955 59427.68267 455817.0774 2972220.092 -2648.730449
acel    -5.573838  -76.10145  -331.7775  -2648.730  7.957951
```

También podíamos haber utilizado una estructura matricial para calcular las matrices de varianzas y covarianzas anteriores.

Se comienza definiendo la matriz cuyas columnas son las variables y a continuación se aplican las funciones para calcular las matrices de correlaciones y covarianzas.

```
> variables1=cbind(consumo, cv,peso,motor,acel)
> cor(variables1)
          consumo        cv      peso     motor     acel
consumo 1.0000000 0.8399707 0.8404614 0.8396855 -0.5004536
cv      0.8399707 1.0000000 0.8597706 0.8973697 -0.7022883
peso    0.8404614 0.8597706 1.0000000 0.9333371 -0.4151784
motor   0.8396855 0.8973697 0.9333371 1.0000000 -0.5446238
acel   -0.5004536 -0.7022883 -0.4151784 -0.5446238  1.0000000
```

```
> cov(variables1)
          consumo        cv      peso     motor     acel
consumo 15.587618 127.38870 939.9813  5715.405 -5.573838
cv     127.388701 1475.54956 9355.5868 59427.683 -76.101453
peso   939.981275 9355.58682 80245.9492 455817.077 -331.777470
motor 5715.404955 59427.68267 455817.0774 2972220.092 -2648.730449
acel    -5.573838  -76.10145  -331.7775  -2648.730  7.957951
```


© Alfaomega - RC Libros

A continuación calculamos la tabla de frecuencias de la variable *cilinder* y la utilizamos para construir un diagrama de barras y un gráfico de sectores (Figuras 7-1 y 7-2).

```
> tabla1=table(cilindr)
> tabla1
cilindr
  3   4   5   6   8
  4 207   3  84 108
> pie(tabla1)
> barplot(tabla1)
```

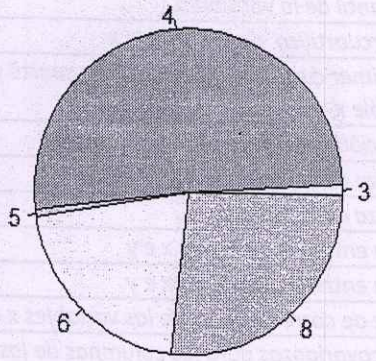


Figura 7-1

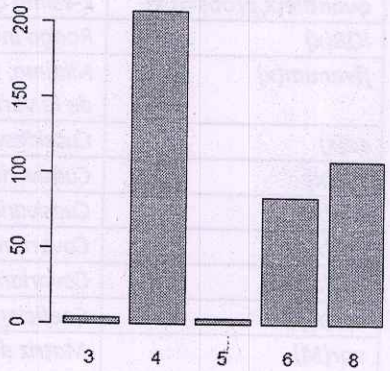


Figura 7-2

También podemos construir la tabla de contingencia de las variables *origen* y *cilinder* utilizando los comandos *table* y *xtabs*.

```
> table(origen,cilindr)
      cilindr
origen   3   4   5   6   8
     1   0  72   0  74 108
     2   0  66   3   4   0
     3   4  69   0   6   0
```

```
> xtabs(~origen+cilindr)
      cilindr
origen   3   4   5   6   8
```

© Alfaomega - RC Libros

```
1   0  72   0  74 108
2   0  66   3   4   0
3   4  69   0   6   0
```

Ejercicio 7-1. Las calificaciones en Álgebra y Cálculo de 7 alumnos son las siguientes:

Álgebra	1	2	3	4	5	6	7
Cálculo	3	4	1	2	7	5	6

Determinar el grado de asociación o concordancia de las calificaciones en Álgebra y Cálculo de los alumnos.

Calculamos el coeficiente de correlación y la covarianza de las dos variables (declaradas como vectores).

```
> algebra=c(1,2,3,4,5,6,7)
> calculo=c(3,4,1,2,7,5,6)
> cor(algebra,calculo)
[1] 0.6071429
> cov(algebra,calculo)
[1] 2.833333
```

Se observa que están correlacionadas positivamente con un coeficiente de correlación de 0,6 que puede considerarse significativo. Además la covarianza es positiva, lo que indica que ambas variables varían en el mismo sentido.

Ejercicio 7-2. Diez autores concurren a un premio de novela, siendo clasificados por tres jueces distintos X, Y y Z en los siguientes órdenes:

J1	1	6	5	10	3	2	8	7	9	4
J2	2	4	3	9	7	10	6	5	1	8
J3	5	3	9	6	4	2	8	7	10	1

Analizar cuál es el par de jueces que más próximos están en cuanto a sus preferencias.

Al tratarse de puntuaciones de ordenación será necesario utilizar la correlación de Spearman a la matriz de variables.

© Alfaomega - RC Libros


```
> J1=c(1,6,5,10,3,2,8,7,9,4)
> J2=c(2,4,3,9,7,10,6,5,1,8)
> J3=c(5,3,9,6,4,2,8,7,10,1)
> M=cbind(J1,J2,J3)
> M
      J1 J2 J3
 [1,]  1  2  5
 [2,]  6  4  3
 [3,]  5  3  9
 [4,] 10  9  6
 [5,]  3  7  4
 [6,]  2 10  2
 [7,]  8  6  8
 [8,]  7  5  7
 [9,]  9  1 10
[10,]  4  8  1

> cor(M,method="spearman")
           J1          J2         J3
J1  1.00000000 -0.07878788  0.5878788
J2 -0.07878788  1.00000000 -0.6121212
J3  0.58787879 -0.61212121  1.0000000
```

Observamos que los jueces más próximos en sus preferencias son el 2 y el 3 porque tiene la mayor correlación posible n la matriz de correlaciones de Spearman.

Ejercicio 7-3. Consideramos las variables X, Y, Z cuyos valores son los siguientes:

X	*Y*	*Z*
2	4	2
3	5	4
6	10	6
8	11	7
10	15	10

Hallar la matriz de correlaciones deduciendo el grado de dependencia de las variables. Hallar la matriz de covarianzas e interpretar el resultado.

Comenzamos definiendo una matriz con las variables del ejercicio, para posteriormente aplicar los comandos que calculan la matriz de correlaciones y la matriz de covarianzas.

© Alfaomega - RC Libros

```
> X=c(2,3,6,8,10)
> Y=c(4,5,10,11,15)
> Z=c(2,4,6,7,10)
> matriz=cbind(X,Y,Z)
> matriz
      X  Y  Z
[1,]  2  4  2
[2,]  3  5  4
[3,]  6 10  6
[4,]  8 11  7
[5,] 10 15 10
> cor(matriz)
          X         Y         Z
X 1.0000000 0.9899319 0.9802123
Y 0.9899319 1.0000000 0.9830213
Z 0.9802123 0.9830213 1.0000000
```

En la salida anterior, se observa que el coeficiente de correlación entre las variables *X* e *Y* es 0,9899, entre *X* y *Z* es 0,9802, y entre *Y* y *Z* es 0,9830, lo que indica la fuerte dependencia lineal positiva entre cada par de variables. La matriz de correlaciones tiene la forma siguiente:

$$\begin{pmatrix} r_{xx} & r_{xy} & r_{xz} \\ r_{xy} & r_{yy} & r_{yz} \\ r_{xz} & r_{yz} & r_{zz} \end{pmatrix} = \begin{pmatrix} 1 & 0{,}9899 & 0{,}9802 \\ 0{,}9899 & 1 & 0{,}9830 \\ 0{,}9802 & 0{,}9830 & 1 \end{pmatrix}$$

Para hallar la matriz de covarianzas utilizamos la siguiente sintaxis:

```
> cov(matriz)
      X    Y     Z
X 11.20 15.0  9.95
Y 15.00 20.5 13.50
Z  9.95 13.5  9.20
```

A la vista de los resultados, podemos decir que como todas las covarianzas son positivas, los pares de variables varían en el mismo sentido alrededor de sus medias.

Por otro lado, como la covarianza entre *X* e *Y* es 12, entre *X* y *Z* es 7,96 y entre *Y* y *Z* es 10,8 y además las varianzas de X, Y y Z son respectivamente 8,96, 16,4 y 7,36 la matriz de covarianzas tiene la forma siguiente:

© Alfaomega - RC Libros

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix} = \begin{pmatrix} 8{,}96 & 12 & 7{,}96 \\ 12 & 16{,}4 & 10{,}8 \\ 7{,}96 & 10{,}8 & 7{,}36 \end{pmatrix}$$

Ejercicio 7-4. Consideramos el archivo coches.sav y queremos construir la tabla de contingencia múltiple entre las variables categóricas origen, cilindr y derivada.

Para construir la tabla de contingencia de las variables *origen, cilindr* y *derivada* utilizamos directamente el comando *xtabs*.

```
> xtabs(~origen+cilindr)
      cilindr
origen   3   4   5   6   8
     1   0  72   0  74 108
     2   0  66   3   4   0
     3   4  69   0   6   0
> xtabs(~origen+cilindr+derivada)
, , derivada = 0

      cilindr
origen   3   4   5   6   8
     1   0   0   0   0 108
     2   0   0   0   0   0
     3   0   0   0   0   0
, , derivada = 1

      cilindr
origen   3   4   5   6   8
     1   0  72   0  74   0
     2   0  66   3   4   0
     3   4  69   0   6   0
```

Ejercicio 7-5. Construir dos funciones que calculen la varianza y la cuasivarianza de dos variables. Calcular la varianza y la cuasivarianza de los 10 primeros números naturales.

Definimos la varianza de una variable *x* como:

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

© Alfaomega - RC Libros

Ahora construimos una función en R para la varianza de una variable *x*.

```
varianza=function(x)
  {
    n<-length(x)
    v<-sum((x-(sum(x)/n))^2)/n
    return(v)
    }
```

A continuación calculamos la varianza de los 10 primeros números naturales.

```
> x=c(1,2,3,4,5,6,7,8,9,10)
> Varianza(x)
[1] 8.25
```

Ahora construimos una función en R para la cuasivarianza de una variable *x*.

```
cuasivarianza=function(x)
  {
    n=length(x)
    v=sum((x-(sum(x)/n))^2)/n
    return(v)
    }
```

A continuación calculamos la cuasivarianza de los 10 primeros números naturales.

```
> cuasivarianza(x)
[1] 9.166667
```


© Alfaomega - RC Libros

CAPÍTULO 8

PROBABILIDAD. DISTRIBUCIONES DISCRETAS Y CONTINUAS

CONCEPTO DE PROBABILIDAD: REGLA DE LAPLACE

Puede definirse la probabilidad de un suceso como una constante en torno a la cual tiende a estabilizarse la frecuencia relativa de ese suceso cuando se repite indefinidas veces el experimento que lo genera. De esta definición se deduce la ***regla de Laplace*** para el cálculo de probabilidades. Dicha regla dice que la probabilidad de un suceso es el cociente entre el número de casos favorables a que ocurra el suceso y el número total de casos posibles. Esta definición se justifica a continuación.

Cuando se realiza un experimento se origina un espacio muestral. Sean S_1, S_2, ..., S_n los sucesos elementales asociados a un fenómeno o experimento aleatorio dado, entendiendo por *sucesos elementales* los más simples posibles, es decir, aquellos que no pueden ser descompuestos en otros sucesos. El conjunto $\{S_1, S_2, ..., S_n\}$ se denomina *espacio muestral* asociado al fenómeno o experimento. El suceso del que se quiere hallar la probabilidad será normalmente un suceso compuesto de varios sucesos elementales del espacio muestral, con lo que su probabilidad será el cociente entre el número de sucesos elementales que lo componen (casos favorables) y el número total

© Alfaomega - RC Libros

PROBABILIDAD CONDICIONADA E INDEPENDENCIA: TEOREMA DE BAYES

Hay ocasiones en que es preciso calcular la probabilidad de un suceso *B* cuando se sabe que se ha realizado previamente un suceso *A* que influye en *B* de alguna forma. Dicha probabilidad se llama *probabilidad del suceso B condicionada al suceso A*, y se representa como $p(B/A)$. En este caso, los casos favorables a que se verifique *B* cuando es seguro que se verifica *A*, serán los casos favorables a la verificación de *A* y *B* simultáneamente (suceso $A \cap B$). Los casos posibles son los correspondientes a la verificación de *A*, ya que este es el suceso seguro, y por tanto, universo (espacio muestral). Luego aplicando la regla de Laplace, podemos escribir:

$$p(B/A) = n(A \cap B)/n(A) = [n(A \cap B)/n]/[n(A)/n] = p(A \cap B)/p(A) \Rightarrow p(A \cap B) = p(B/A)p(A)$$

$$p(A/B) = n(A \cap B)/n(B) = [n(A \cap B)/n]/[n(B)/n] = p(A \cap B)/p(B) \Rightarrow p(A \cap B) = p(A/B)p(B)$$

Dos sucesos A y B son independientes si $p(B/A)=p(B)$ o $p(A/B)=p(A)$, es decir, cuando las condiciones no intervienen en el cálculo de las probabilidades. Esta condición de independencia de los sucesos *A* y B es evidentemente equivalente a $p(A \cap B) = p(A)*p(B)$. Por otro lado, se tiene que $p(A \cup B) = p(A)+p(B)- p(A \cap B)$ y $p(\overline{A}) = 1-P(A)$ ($\overline{A}$ = complementario de A).

En el trabajo con probabilidad condicionada juega un papel importante el *teorema de Bayes*. Supongamos que tenemos un espacio muestral relativo a un experimento cuyos sucesos elementales son $S_1, S_2, \ldots, S_n$, y sea *A* un suceso cualquiera. Cuando calculamos la probabilidad condicionada $p(A/S_i)$, el suceso S_i ocurre antes que el suceso *A*. Inicialmente, parece absurdo intentar calcular la probabilidad $p(S_i/A)$, ya que el suceso *A* es posterior al suceso S_i. Sin embargo, el teorema de Bayes va a permitir el cálculo de la probabilidad condicionada, $p(S_i/A)$, a partir del conocimiento *a priori* de las probabilidades $p(S_i)$ y $p(A/S_i)$, mediante la expresión siguiente:

$$p(S_i/A) = \frac{p(Si) * p(A/S_i)}{\sum_{i=1}^{n} p(Si) * p(A/S_i)}$$

El resultado es inmediato, ya que, por una parte, $p(A) = p(A \cap E) = p(A \cap (\cup S_i)) = p(\cup(A \cap S_i)) = \sum p(A \cap S_i)$, y por otra parte, $p(S_i/A) = p(A \cap S_i)/p(A) = p(A \cap S_i)/\sum p(A \cap S_i) = [p(S_i)*p(A/S_i)] / \sum [p(S_i)*p(A/S_i)]$.

© Alfaomega - RC Libros

Un *índice complejo* ha de ser sencillo, pero a la vez ha de reunir la mayor cantidad posible de información. Así, podemos considerar dos tipos de índices complejos, en los que va a primar, en mayor o menor medida, cada una de las características anteriores. Si se prefiere sencillez, tendremos los índices complejos no ponderados, y si, por el contrario, lo que se desea es que contengan la mayor cantidad de información posible, se utilizarán los índices complejos ponderados.

VARIABLE ALEATORIA DISCRETA Y DISTRIBUCIÓN DE PROBABILIDAD DE TIPO DISCRETO

Los métodos estadísticos se aplican al estudio de fenómenos en cuya toma de decisiones prevalecen condiciones de incertidumbre. Esta incertidumbre plantea al investigador la necesidad de medirla, siendo la medida de la incertidumbre la probabilidad.

A cada uno de los posibles resultados de un experimento se le atribuye una posibilidad de ocurrir, que, siguiendo los axiomas de Kolmogorov, será un número comprendido entre 0 y 1. Cualquier suceso S asociado al experimento tendrá mayor probabilidad de suceder cuanto más próxima esté a 1 su probabilidad, y menor cuanto más próxima esté a 0.

Asociado al concepto de probabilidad está el concepto de variable aleatoria. Una *variable aleatoria* representa el conjunto de valores que pueden observarse en un fenómeno aleatorio, valores que dependen del azar y sobre los cuales es posible establecer una medida de su probabilidad. Según el número de valores que pueda tomar, la variable aleatoria puede ser de dos tipos:

- *Variable aleatoria discreta*: si únicamente toma un número finito o numerable de valores.
- *Variable aleatoria continua*: cuando puede tomar los infinitos valores de un intervalo.

Se denomina *distribución de probabilidad de una variable aleatoria* a la función que asigna probabilidad a los valores que puede tomar la variable. Cuando se especifican los posibles valores de la variable aleatoria y sus probabilidades respectivas, tenemos construido el modelo de distribución de probabilidad.

© Alfaomega - RC Libros

La clasificación de las variables aleatorias en discretas y continuas origina la presencia de dos grandes grupos de distribuciones de probabilidad, que se asocian a estos dos tipos de variables. De esta forma, tendremos las distribuciones discretas y las distribuciones continuas.

Una distribución de probabilidad es de tipo discreto cuando la cuantía total de su probabilidad se encuentra distribuida entre un número finito, o a lo sumo numerable, de puntos. Esto quiere decir que si la variable aleatoria es discreta, puede tomar los valores $x_1, x_2, \ldots$ para cada uno de los cuales existirá una probabilidad $p_1, p_2, \ldots$

Se denomina *ley de probabilidad* o *función de cuantía* de la variable aleatoria discreta X a la función $P_i=P(x_i)=P(X=x_i)$. La función de cuantía nos proporciona la cantidad de probabilidad contenida en cada uno de los puntos, o valores, de la variable aleatoria. De esta forma, el modelo de distribución de probabilidad queda especificado de la forma siguiente:

$X=x_i$	$x_1, x_2, \cdots, x_n$
$P(X=x_i)$	$p_1, p_2, \cdots, p_n$

La ley de probabilidad de la variable X verifica que $P(x)\geq 0$ para cualquier valor x de la variable, y también verifica que $\sum p(x)=1$, extendida la suma a todos los valores x de la variable aleatoria X.

Se denomina *función de distribución* de una variable aleatoria X en el punto x a la función $F(x)$ definida como:

$$F(x)=P(X\leq x)=\sum_{x_i\leq x} P(X=x_i)$$

De esta forma, la función de distribución de la variable aleatoria X en el punto x es la suma de todas las probabilidades contenidas en los puntos del intervalo $(-\infty, x]$.

La función de distribución de una variable aleatoria discreta es escalonada con saltos de valor $p(x_i)$ en los puntos x_i. Además es siempre no decreciente y verifica que $F(-\infty)=0$ y $F(+\infty)=1$.

Al igual que la ley de probabilidad (función de cuantía), la función de distribución caracteriza completamente a una variable aleatoria.

© Alfaomega - RC Libros

Las distribuciones de probabilidad discretas y las distribuciones de frecuencias presentan gran analogía. La ley de probabilidad o función de cuantía de una distribución de probabilidad discreta, es equivalente a las frecuencias relativas sin acumular de una distribución de frecuencias. La función de distribución de una distribución de probabilidad discreta es equivalente a las frecuencias relativas acumuladas de una distribución de frecuencias.

DISTRIBUCIONES DISCRETAS DE UNA Y DOS DIMENSIONES: CARACTERÍSTICAS

Suelen considerarse características esenciales de una distribución de frecuencias la esperanza matemática, la varianza, la desviación típica y los momentos. Estos conceptos ya no son nuevos para nosotros, pues fueron estudiados al tratar las distribuciones de frecuencias.

La esperanza matemática de una variable aleatoria es un concepto similar al de media de una distribución de frecuencias. Si la variable X es discreta, se define su esperanza como:

$$E(x)=\mu=\sum_i x_i P(X=x_i)=\sum_i x_i P_i$$

La varianza de una variable aleatoria se define como:

$$V(X)=\sigma^2=E[(X-\mu)]^2$$

Para variables discretas, tendremos:

$$V(X)=\sigma^2=\sum_i (x_i-\mu)^2 P_i$$

La desviación típica se define como la raíz cuadrada positiva de la varianza.

Al igual que en las distribuciones de frecuencias, se define el momento de orden k centrado en el origen como $m_k = E(X^k)$, que para variables discretas será:

$$m_k=E(X^k)=\sum_i x_i^k P_i$$

© Alfaomega - RC Libros

Se define el momento de orden k centrado en la media como $\mu_k = E((X-\mu)^k)$, que para variables discretas será:

$$\mu_k = E((X-\mu)^k) = \sum_i (x_i - \mu)^k P_i$$

Las *distribuciones bidimensionales de probabilidad* surgen cuando representamos los posibles resultados de un experimento aleatorio por medio de una variable aleatoria (X,Y) cuyos posibles valores son pares del tipo (x_1,y_1), (x_2,y_2), ... , (x_n,y_n), de forma que se puedan definir las probabilidades de las regiones $[-\infty<X\leq x)x[-\infty<Y\leq y)$ para todo par de valores (x,y).

Se define la *ley de probabilidad conjunta* para la variable aleatoria discreta bidimensional (X,Y) como:

$$P(X=x_i,\ Y=y_j) = P_{ij}$$

La *función de distribución conjunta* para la variable (X,Y), se define como:

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P(X = x_i, Y = y_j) = \sum_{x_i \leq x} \sum_{y_j \leq y} P_{ij}$$

Las *leyes de probabilidad marginales* para X e Y se definen como:

$$P(X = x_i) = P(X = x_i, Y < \infty) = \sum_j P_{ij} = P_{i.}$$

$$P(Y = y_i) = P(X < \infty, Y = y) = \sum_i P_{ij} = P_{.j}$$

Las *funciones de distribución marginales* para X e Y se definen como:

$$F_1(x) = P(X \leq x, Y < \infty) = \sum_{x_i \leq x} P_{i.}$$

$$F_2(y) = P(X < \infty, Y \leq y) = \sum_{y_j \leq y} P_{.j}$$

Las *leyes de probabilidad condicionadas* de X a Y e Y a X se definen como:

$$P\left(X = x_i / Y = y_j\right) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{P_{ij}}{P_{.j}}$$

© Alfaomega - RC Libros

$$P\left(Y = y_j / X = x_i\right) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{P_{ij}}{P_{i.}}$$

Por último, las *funciones de distribución condicionadas* de X a Y e Y a X se definen como:

$$F\left(x / Y = y_j\right) = P\left(X \leq x / Y = y_j\right) = \frac{P(X \leq x, Y = y_j)}{P(Y = y_j)} = \sum_{x_i \leq x} \frac{P_{ij}}{P_{.j}}$$

$$F\left(y / X = x_i\right) = P\left(Y \leq y / X = x_i\right) = \frac{P(X = x_i, Y \leq y)}{P(X = x_i)} = \sum_{y_j \leq y} \frac{P_{ij}}{P_{i.}}$$

Otro concepto esencial es el de independencia. *Dos variables aleatorias X e Y son independientes* cuando se cumple que la función de distribución conjunta de ambas es igual al producto de las respectivas funciones de distribución marginales es decir:

$$F(x,y) = F_1(x)\ F_2(y)$$

Esta condición general de independencia puede traducirse para variables discretas en que la ley de probabilidad conjunta sea igual al producto de las leyes de probabilidad marginales, es decir:

$$P_{ij} = P_{i.}\ P_{.j}$$

DISTRIBUCIONES BINOMIAL, DE POISSON, GEOMÉTRICA E HIPERGEOMÉTRICA

Existen determinados experimentos tales que la probabilidad de sus sucesos puede modelarse según una determinada distribución de tipo discreto. Un modelo de probabilidad discreto define la ley de probabilidad para los sucesos fruto de un determinado tipo de experimento o fenómeno aleatorio. A continuación se exponen algunos de los modelos de probabilidad discretos más habituales.

© Alfaomega - RC Libros

Distribución	*Ley de probabilidad*	*Parámetros*	*Media*	*Varianza*
Bernoulli	$p^x(1-p)^{1-x}$ $x=0,1$	p $0<p<1$	p	$p(1-p)$
Binomial	$\binom{n}{x}p^x(1-p)^{1-x}$ $x=0,1,\ldots,n$	n,p $n=1,2,\ldots$ $0<p<1$	np	$np(1-p)$
Poisson	$\frac{\lambda^x}{x!}e^{-\lambda}$	$\lambda>0$	λ	λ
Geométrica	$p(1-p)^{x-1}$ $x=1,2,\ldots$	$0\le p\le 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial negativa	$\binom{n+x-1}{n-1}p^n(1-p)^x$ $x=0,1,\ldots$	$n>0$ $0\le p\le 1$	$\frac{n(1-p)}{p}$	$\frac{n(1-p)}{p^2}$
Hiper-geométrica	$\frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}$ $x=0,1,\ldots,n$	$N=1,2,\ldots$ $n=1,2,\ldots,N$ $p=0,\frac{1}{N},\ldots,1$	np	$np(1-p)$ $*\frac{N-n}{N-1}$

En cuanto a la *distribución de Bernoulli*, los ensayos o pruebas repetidas e independientes se llaman de Bernoulli cuando en cada ensayo solo hay dos resultados posibles (éxito o fracaso) que tienen la misma probabilidad en todas las repeticiones. Estas dos probabilidades posibles suelen denotarse por p y q ($p + q = 1$), siendo p la probabilidad de éxito y q la probabilidad de fracaso. Suele utilizarse esta distribución en modelos físicos 0-1 (fracaso-éxito o fallo-funcionamiento).

En cuanto a la *distribución binomial*, supongamos el experimento aleatorio consistente en la realización de n ensayos o pruebas de Bernoulli. Queremos calcular la probabilidad de que entre estos n ensayos se obtengan k éxitos. La variable aleatoria X que define esta probabilidad se denomina binomial. Suele usarse en control de calidad en diagramas de control por atributos y muestreo con reemplazo.

Es el modelo probabilístico apropiado para muestrear una población infinitamente grande, donde p representa la fracción o proporción de éxitos o fracasos en la población. X suele representar el número de artículos disconformes encontrados en una muestra aleatoria de tamaño n. Una variable aleatoria que aparece frecuentemente en el control estadístico de la calidad es X/n, donde X es una binomial $B(n,p)$. Esta variable suele asociarse al cociente del número de artículos defectuosos en la muestra entre el tamaño muestral, y suele denominarse en control de calidad *fracción muestral de defectuosos*. La esperanza de esta variable es p, y la varianza es pq/n.

© Alfaomega - RC Libros

En cuanto a la *distribución de Poisson*, cuando en la realización de las pruebas de Bernoulli la probabilidad de éxito es muy pequeña con respecto a np, y nq muy pequeño con respecto a n, el cálculo de los valores de la ley de probabilidad resulta muy complicado y laborioso, por lo que dichos cálculos se omiten, utilizando una distribución que se conoce con el nombre de distribución de Poisson, que se aproxima a la binomial. La aproximación de las probabilidades binomiales por las de Poisson son buenas cuando $np < 5$ y $p < 0{,}1$. Por ser esta distribución una buena aproximación de la binomial cuando p es muy pequeño con respecto a n, es por lo que a veces se denomina distribución de los sucesos raros. Haciendo tender n a infinito y p hacia 0 (manteniendo $np = \mu$ constante), se demuestra matemáticamente que la ley de probabilidad de la binomial tiende a la ley de probabilidad de la variable de Poisson de parámetro μ. La distribución de Poisson se utiliza como distribución de las ocurrencias de un fenómeno en un intervalo de tiempo. También se utiliza como modelo del número de defectos o disconformidades que ocurren en una unidad de producto. En realidad, cualquier fenómeno aleatorio que ocurre por unidad (de área, de volumen, de tiempo, etc.) se puede aproximar bien en la mayoría de los casos por la ley de Poisson. También se utiliza en el diseño de los límites de control de los diagramas p (proporción de unidades defectuosas) y u (número de defectos por unidad) en control de calidad.

En cuanto a la *distribución geométrica*, la variable aleatoria X, definida como el número de pruebas independientes necesarias para que un suceso A de probabilidad p aparezca por primera vez, recibe el nombre de geométrica. La variable geométrica también suele caracterizarse como el tiempo de espera hasta el primer éxito, y se utiliza en estudios de fiabilidad y en situaciones cíclicas donde se alternan éxitos y fracasos.

En cuanto a la *distribución binomial negativa*, consideremos una serie de pruebas independientes y sea n el número exacto de veces que aparece el suceso A de probabilidad p. La variable aleatoria X, que denota el número de fallos antes del éxito n-ésimo, recibe el nombre de binomial negativa. La variable binomial negativa también suele caracterizarse como el tiempo de espera hasta el n-ésimo éxito, y tiene campo de aplicación similar a la geométrica (fiabilidad, procesos de alternancia entre éxitos y fallos, etc.). También es útil como modelo estadístico fundamental para varios tipos de datos de conteo, como la ocurrencia de disconformidades en una unidad de un producto.

En cuanto a la *distribución hipergeométrica*, es el modelo probabilístico apropiado para seleccionar una muestra aleatoria de n artículos, sin reposición, de un lote de N de ellos, de los cuales $D = Np$ son disconformes o defectuosos. En este caso, si

© Alfaomega - RC Libros

la razón *n/N*, llamada a menudo fracción de muestreo, es pequeña (menor que 0,1), entonces la distribución binomial *B(n,p)* con *p = D/N* es una adecuada aproximación a la hipergeométrica. La aproximación es mejor cuanto más pequeños sean los valores de la fracción de muestreo *n/N*. La distribución hipergeométrica se utiliza con asiduidad en planes de muestreo para aceptación.

En cuanto a la relación entre distribuciones, se cumple que si $X_1, X_2, \ldots, X_n$ son variables aleatorias de Bernoulli independientes, entonces su suma es una binomial de parámetros *n* y *p* (*B(n,p)*). También se cumple que si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes geométricas de parámetro *p*, $\sum X_i$ es una variable aleatoria binomial negativa de parámetros (*n,p*). Por otra parte, si $n \geq 30$ y $min(np,nq) < 10$, la distribución binomial *B(n,p)* se puede aproximar por una distribución de Poisson de parámetro $\lambda = np$. Además, si $n \geq 30$ y $min(np,nq) \geq 10$, la distribución binomial *B(n,p)* se puede aproximar por una distribución normal con $\mu = np$ y $\sigma^2 = npq$.

DISTRIBUCIÓN MULTINOMIAL

Consideramos un vector aleatorio enedimensional ($e_1, e_2, \ldots, e_N$) cuyas marginales e_i son binomiales con los siguientes parámetros:

$$e_i \rightarrow binomial(n, P_i), \quad E[e_i] = nP_i, \quad V[e_i] = nP_iQ_i = nP_i(1-P_i)$$

La ley de probabilidad de ($e_1, e_2, \ldots, e_N$) seguirá el modelo multinomial (conjunta de *N* binomiales e_i).

$$P(e_1 = t_1, e_2 = t_2, \cdots, e_N = t_N) = \frac{n!}{t_1!t_2!\cdots t_N!} P_1^{t_1} P_2^{t_2} \cdots P_N^{t_N} \qquad n! = (t_1 + t_2 + \cdots + t_N)! \qquad \sum_{i=1}^{N} t_i = n$$

Supongamos que se lanzan *n* bolas sobre *N* compartimentos, de forma que en un compartimento puedan caer 0,1,2,...,*n* bolas. El modelo multinomial viene definido por la probabilidad de que al lanzar las *n* bolas sobre los *N* compartimentos t_1 de ellas caigan en el compartimento 1, t_2 en el compartimento 2, ..., t_n en el compartimento *N*. P_i sería la probabilidad de que una bola caiga en el compartimento *i* en cada lanzamiento.

Se demuestra que el vector esperanza matemática y la matriz de varianzas covarianzas $\sum$ para la ley de probabilidad multinomial son los siguientes:

© Alfaomega - RC Libros

$$E(e_1, \cdots, e_N) = (nP_1, \cdots, nP_N)$$

$$\sum_{(e_1,\cdots,e_N)} = \begin{pmatrix} nP_1(1-P_1) & -nP_1P_2 & \cdots & -nP_1P_N \\ -nP_2P_1 & nP_2(1-P_2) & \cdots & -nP_2P_N \\ \vdots & \vdots & & \vdots \\ -nP_NP_1 & -nP_NP_2 & \cdots & nP_N(1-P_N) \end{pmatrix}$$

DISTRIBUCIÓN HIPERGEOMÉTRICA GENERALIZADA

Consideramos un vector aleatorio enedimensional ($e_1, e_2, \ldots, e_N$) cuyas marginales e_i son hipergeométricas con los siguientes parámetros:

$$e_i \rightarrow Hipergeométrica(M, n, P_i)$$
$$E[e_i] = nP_i$$
$$V[e_i] = \frac{M-n}{M-1} nP_iQ_i = \frac{M-n}{M-1} nP_i(1-P_i)$$

La ley de probabilidad de ($e_1, e_2, \ldots, e_N$) seguirá el modelo multinomial (conjunta de *N* hipergeométricas e_i).

$$P(e_1 = t_1, e_2 = t_2, \cdots, e_N = t_N) =$$

$$= \frac{\binom{M_1}{t_1}\binom{M_2}{t_2}\cdots\binom{M_N}{t_N}}{\binom{M_1 + M_2 + \cdots M_N}{t_1 + t_2 + \cdots t_N}} = \frac{\binom{M \cdot P_1}{t_1}\binom{M \cdot P_2}{t_2}\cdots\binom{M \cdot P_N}{t_N}}{\binom{M}{n}} \quad con \quad \sum_{i=1}^{N} t_i = n$$

Se considera un esquema de urna en el que la unidad u_i viene representada por M_i bolas. En este esquema de selección al seleccionar la unidad u_i se retira una bola de entre las M_i que representan a u_i y no se vuelve a reponer a la urna para la siguiente extracción. Se podrá extraer la unidad u_i las veces que corresponda mientras no se acaben las M_i bolas que la representan o mientras no se cubra el número total de bolas *n* que pueden ser extraídas, por lo que la unidad u_i puede extraerse un máximo de veces igual a $Mín(M_i,n)$ $i = 1,..,N$. Si ahora definimos la variable aleatoria e_i como el número de veces que la unidad i-ésima u_i de la urna pertenece al conjunto total de *n* bolas a extraer, evidentemente e_i se distribuye según una hipergeométrica de parámetro *M*, *n* y P_i, siendo P_i la probabilidad de selección de la unidad i-ésima. Ya sabemos que la variable aleatoria e_i puede tomar

© Alfaomega - RC Libros

los valores 0,1,2,.., *Mín*(M_i,n) i=1,..,N. La selección del conjunto de la n bolas (t_1 de entre M_1, t_2 de entre M_2, ..., t_N de entre M_N) seguirá el modelo hipergeométrico generalizado (conjunta de n hipergeométricas e_i).

Se demuestra que el vector esperanza matemática y la matriz de varianzas covarianzas Σ para nuestra ley de probabilidad hipergeométrica generalizada son los siguientes:

$$E(e_1,\cdots,e_N)=(nP_1,\cdots,nP_N)$$

$$\sum_{(e_1,\cdots,e_N)}=\begin{pmatrix} \frac{M-n}{M-1}nP_1(1-P_1) & -\frac{M-n}{M-1}nP_1P_2 & \cdots & -\frac{M-n}{M-1}nP_1P_N \\ -\frac{M-n}{M-1}nP_2P_1 & \frac{M-n}{M-1}nP_2(1-P_2) & \cdots & -\frac{M-n}{M-1}nP_2P_N \\ \vdots & \vdots & & \vdots \\ -\frac{M-n}{M-1}nP_NP_1 & -\frac{M-n}{M-1}nP_NP_2 & \cdots & \frac{M-n}{M-1}nP_N(1-P_N) \end{pmatrix}$$

DISTRIBUCIÓN UNIFORME DISCRETA

Una variable aleatoria X se dice que tiene una distribución uniforme sobre n puntos (x_1, x_2, ..., x_n) si su ley de probabilidad es la siguiente:

$$P(X=x)=\frac{1}{n} \quad para\ x=x_1,x_2,\cdots,x_n$$

Las características de esta distribución son las siguientes:

$$E(X)=\sum_{i=1}^{n}x_iP(X=x_i)=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$m_k=E(X^k)=\sum_{i=1}^{n}x_i^kP(X=x_i)=\frac{1}{n}\sum_{i=1}^{n}x_i^k$$

$$V(X)=E(X^2)-(E(X))^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2$$

© Alfaomega - RC Libros

R Y LAS DISTRIBUCIONES DISCRETAS

R dispone de varias funciones que permiten trabajar con distribuciones discretas. En general, R permite trabajar con la ley de probabilidad, la función de distribución y los cuartiles de cualquier distribución discreta. Asimismo,. permite generar números aleatorios según cualquier distribución discreta.

Las funciones esenciales de R para el trabajo con distribuciones discretas se presentan a continuación:

nchosek(n,k)	*Devuelve el número combinatorio n sobre k, esto es, el número de combinaciones sin repetición de n elementos tomados de k en k. Su valor es:* $\binom{n}{k}=\frac{P_{k,n}}{k!}=\frac{n!}{k!(n-k)!} \quad P_{k,n}=\frac{n!}{(n-k)!}$
factorial(n)	*Factorial de n (n!=n(n-1)(n-2)...(3)(2)(1).*
dbinom(x,n,p)	*Halla el valor de la ley de probabilidad (función de masa) en el punto x de una variable aleatoria binomial de parámetros n y p* $dbinom(x,n,p)=\binom{n}{x}p^x(1-p)^{n-x}$
pbinom(x,n,p)	*Halla el valor de la función de distribución en el punto x de una variable aleatoria binomial de párametros n y p* $pbinom(x,n,p)=\sum_{k=1}^{x}\binom{n}{k}p^k(1-p)^{n-k}$
qbinom(q,n,p)	*Devuelve el cuantil de orden q de una distribución binomial (n,p). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una binomial (n.p)*
rbinom(m,n,p)	*Devuelve una muestra aleatoria de tamaño m de una distribución binomial (n,p)*
dpois(x,λ)	*Halla el valor de la ley de probabilidad (función de masa) en el punto x de una variable aleatoria de Poisson de parámetro λ.* $dpois(x,\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}$

© Alfaomega - RC Libros

ppois(x,λ)	*Halla el valor de la función de distribución en el punto x de una variable aleatoria de Poisson de parámetro λ..* $$ppois(x,\lambda)=\sum_{k=1}^{x}\frac{e^{-\lambda}\lambda^{k}}{k}$$
qpois(q,λ)	*Devuelve el cuantil de orden q de una distribución de Poisson de parámetro λ. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Poisson de parámetro λ*
rpois(m,λ)	*Devuelve una muestra aleatoria de tamaño m de una distribución de Poisson de parámetro λ.*
dgeom(x,p)	*Halla el valor de la ley de probabilidad (función de masa) en el punto x de una variable aleatoria geométrica de parámetro p* $$dgeom(x,p)=p(1-p)^{x-1}$$
pgeom(x,p)	*Halla el valor de la función de distribución en el punto x de una variable aleatoria geométrica de parámetro p.* $$pgeom(x,\lambda)=\sum_{k=1}^{x}\frac{e^{-\lambda}\lambda^{k}}{k!}$$
qgeom(q,p)	*Devuelve el cuantil de orden q de una distribución goemétrica de parámetro p. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una geométrica de parámetro p*
rgeom(m,p)	*Devuelve una muestra aleatoria de tamaño m de una distribución geométrica de parámetro p*
dnbinom(x,n,p)	*Halla el valor de la ley de probabilidad (función de masa) en el punto x de una variable aleatoria binomial negativa de parámetros n y p* $$dnbinom(x,n,p)=\binom{x+n-1}{n-1}p^{n}(1-p)^{x}$$
pnbinom(x,n,p)	*Halla el valor de la función de distribución en el punto x de una variable aleatoria binomial negativa de parámetros n y p* $$pnbinom(x,n,p)=\sum_{k=1}^{x}\binom{k+n-1}{n-1}p^{n}(1-p)^{k}$$

© Alfaomega - RC Libros

qnbinom(q,n,p)	*Devuelve el cuantil de orden q de una distribución binomial (n,p). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una binomial negativa(n.p)*
rqnbinom(m,n,p)	*Devuelve una muestra aleatoria de tamaño m de una distribución binomial negativa (n,p)*
dhyper(x,D,N-D,n)	*Halla el valor de la ley de probabilidad (función de masa) en el punto x de una variable aleatoria hipergeométrica de parámetro (D,N,n)* $$dhyper(x,D,N-D,n)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$
phyper(x,D,N-D,n)	*Halla el valor de la función de distribución en el punto x de una variable aleatoria hipergeométrica de parámetros (D,N,n)* $$phyper(x,D,N-D,n)=\sum_{k=1}^{x}\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$
qhyper(q,D,N-D,n)	*Halla el valor del cuantil de orden q de una hipergeométrica de parámetros (D,N,n). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una hipergeométrica de parámetros (D,N,n)*
rhyper(m,D,N-D,n)	*Halla m números aleatorios de una distribución hipergeométrica de parámetros (D,N,n). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una hipergeométrica de parámetros (D,N,n)*
dmultinom(x,n,p)	*Función de masa en el punto x de la multinomial(n,p), x es el vector de tamaños a seleccionar de las clases, p es el vector de probabilidades de las clases y n es el número de clases*
rmultinom(m,n,p)	*Muestra de tamaño m de una multinomial(n,p)*

© Alfaomega - RC Libros

VARIABLES ALEATORIAS CONTINUAS

La distribución de probabilidad de una variable es continua cuando la variable aleatoria puede tomar los infinitos valores de un intervalo. En este caso, la masa o cuantía de probabilidad se distribuye en el intervalo dado. Como características destacables de las distribuciones continuas tenemos por un lado que no es posible conocer su valor exacto, ya que medir el valor de una variable de este tipo consiste únicamente en clasificarlo dentro de un intervalo de mayor o menor amplitud, y por otro lado que las probabilidades se asignan a intervalos, no a puntos. De hecho, la probabilidad de un punto es cero para cualquier variable continua.

Dado que la probabilidad de un punto es cero, es necesario introducir un concepto que aproxime tal probabilidad puntual. Dicho concepto es el de elemento diferencial de probabilidad, que viene dado como:

$$P(x < X \leq x + dx) = f(x)dx$$

donde $f(x)$ es la *densidad de probabilidad* en un punto, o *función de densidad* de la variable aleatoria continua X en el punto x.

De esta forma, podríamos definir la función de densidad de una variable aleatoria continua X en el punto x como el límite de la densidad media de probabilidad de un intervalo que contiene al punto x, cuando la amplitud de dicho intervalo tiende a cero. De esta forma, la probabilidad que un modelo de distribución de probabilidad continua asigna a un punto concreto es cero, mientras que la probabilidad de cualquier intervalo, por pequeño que este sea, vendrá dada por el área bajo la función de densidad, en dicho intervalo.

A partir de la noción de probabilidad elemental, la *función de distribución* de la variable aleatoria continua X en el punto x puede expresarse de la forma:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

donde debe verificarse:

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} f(u)du = 1$$

La distribución de probabilidad de una variable aleatoria continua vendrá caracterizada por su función de densidad, lo mismo que la distribución de probabilidad de una variable aleatoria discreta venía caracterizada por su función de cuantía o ley de probabilidad.

© Alfaomega - RC Libros

Lo mismo que para variables discretas se verificaba que $P(x) \geq 0$ para cualquier valor x de la variable, y también $\sum p(x) = 1$, extendida la suma a todos los valores x de la variable aleatoria X, para variables continuas se verifica:

- $f(x) \geq 0$ para todo x
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a < x < b) = \int_{a}^{b} f(x)dx$

Las distribuciones de probabilidad continuas y las distribuciones de frecuencias también presentan analogía. Dada una distribución de frecuencias por clases, si cada vez se hacen clases de longitud más pequeña, el histograma de frecuencias tenderá a una curva límite suave, que describirá el comportamiento de todos los valores posibles de la variable estudiada. Esta función límite se llama función de densidad.

Al igual que la función de distribución, la función de densidad caracteriza completamente a una variable aleatoria.

La relación entre la función de densidad y la función de distribución de una variable aleatoria continua es evidente, ya que:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \Rightarrow F'(x) = f(x)$$

Otro concepto muy utilizado es el de *función complementaria de la función de distribución*, que también se denomina *función de supervivencia* o *función de fiabilidad* (para variables no negativas), y que viene definida como:

$$R(x)=P(X>x)=1-F(x)$$

Evidentemente, se cumple que $R'(x)= -f(x)$

Otro concepto muy útil cuando se estudia la fiabilidad o el mantenimiento de un sistema es el de *función de riesgo* o *tasa de fallos*. Se trata de una función no negativa, $\lambda(t)$, definida como $\lambda(t) = f(t)/R(t) = -R'(t)/R(t)$. Esta función da una medida de la probabilidad de fallo o avería de un equipo en cada momento, y, en general, mide el riesgo o propensión a que la variable aleatoria X tome un valor de un intervalo próximo a x, sabiendo que no ha tomado un valor del intervalo definido por los puntos anteriores a x.

© Alfaomega - RC Libros

También se define la *función de riesgo acumulada* como $H(t) = \int_0^t \lambda(s)ds$.

Adicionalmente, se define la *función de riesgo promedio* como $h(t) = H(t)/t$.

Como $\lambda(t) = -R'(t)/R(t)$, integrando tenemos que $R(t) = e^{-H(t)}$.

CARACTERÍSTICAS DE LAS DISTRIBUCIONES CONTINUAS

Consideraremos ahora características esenciales de una distribución de frecuencias, como son la esperanza matemática, la varianza, la desviación típica y los momentos en el caso continuo.

Si la variable X es continua, con variación en toda la recta real, se define su esperanza matemática como:

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

La varianza de una variable aleatoria se define como:

$$V(X) = \sigma^2 = E[(X-\mu)^2]$$

Para variables continuas, tendremos $V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

La desviación típica se define como la raíz cuadrada positiva de la varianza.

Al igual que en las distribuciones de frecuencias, se define el momento de orden k centrado en el origen como $m_k = E(X^k)$.

Si la variable X es continua, con variación en toda la recta real, se define su momento de orden k centrado en el origen como:

$$m_k = E(X^k) = \int_{-\infty}^{\infty} x^k f(x)dx$$

Se define el momento de orden k centrado en la media como $\mu_k = E((X-\mu)^k)$.

Si la variable X es continua, con variación en toda la recta real, se define su momento de orden k centrado en la media como:

$$\mu_k = E((X-\mu)^k) = \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx$$

© Alfaomega - RC Libros

VARIABLES ALEATORIAS CONTINUAS BIDIMENSIONALES. INDEPENDENCIA

Ya sabemos que las *distribuciones bidimensionales de probabilidad* surgen cuando representamos los posibles resultados de un experimento aleatorio por medio de una variable aleatoria (X,Y) cuyos posibles valores son pares del tipo (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), de forma que se puedan definir las probabilidades de las regiones $[-\infty < X \leq x) \times [-\infty < Y \leq y)$ para todo par de valores (x,y).

La *función de distribución conjunta* para la variable (X,Y) se define como:

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$$

donde $f(x,y)$ es la función de densidad conjunta de la variable bidimensional (X,Y).

Las *funciones de densidad marginales* para X e Y se definen como:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,v)dv$$

$$f_2(x) = \int_{-\infty}^{\infty} f(u,y)du$$

Las *funciones de distribución marginales* para X e Y se definen como:

$$F_1(x) = P(X \leq x, Y < \infty) = F(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dudv = \int_{-\infty}^{x} f_1(u)du$$

$$F_2(y) = P(X < \infty, Y \leq y) = F(\infty,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(u,v)dudv = \int_{-\infty}^{y} f_2(v)dv$$

Las *funciones de densidad condicionadas* se definen como:

$$f(y \mid x) = f(x,y) / f_1(x)$$

$$f(x \mid y) = f(x,y) / f_2(y)$$

Por último, las *funciones de distribución condicionadas* se definen como:

$$F(x \mid Y = y) = \frac{\int_{-\infty}^{x} f(x,y)dx}{f_2(y)} \qquad F(y \mid X = x) = \frac{\int_{-\infty}^{y} f(x,y)dy}{f_1(x)}$$

© Alfaomega - RC Libros

Dos variables aleatorias X e Y son independientes cuando se cumple que la función de distribución conjunta de ambas es igual al producto de las respectivas funciones de distribución marginales, es decir:

$$F(x,y) = F_1(x)\, F_2(y)$$

La condición general de independencia también puede traducirse para variables continuas en las que la función de densidad conjunta sea igual al producto de las funciones de densidad marginales, es decir:

$$f(x,y) = f_1(x)\, f_2(y)$$

También se puede expresar la independencia de dos variables continuas en función de las funciones de densidad condicionadas, de tal forma que las variables *X* e *Y* son independientes si se verifica:

$$f(y \mid x) = f_2(y)$$

$$f(x \mid y) = f_1(x)$$

DISTRIBUCIONES CONTINUAS RELEVANTES, NORMAL, EXPONENCIAL, $\chi 2$, T, F, ETC.

De la misma forma que existen modelos de probabilidad que caracterizan variables de tipo discreto, también existe la posibilidad de modelar variables aleatorias de tipo continuo. De entre todos los modelos continuos, la distribución normal es la más utilizada, sobre todo mediante la convergencia hacia ella de otras distribuciones en determinadas condiciones no demasiado fuertes. Por ejemplo, el *teorema de De Moivre* permite aproximar la distribución binomial $B(n,p)$ por una distribución normal $N(np, (npq)^{1/2})$, para valores grandes de *n*. Además, el *teorema central del límite* permite aproximar la variable $X_1 + X_2 + \ldots + X_n$ por una normal $N[nE(X_i), nV(X_i)^{1/2}]$ para *n* grande, siendo las X_i variables independientes e idénticamente distribuidas.

A continuación se presenta una tabla que contiene la función de densidad, la media y la varianza para las distribuciones continuas más representativas y utilizadas en la práctica.

© Alfaomega - RC Libros

Distribuc.	*Función de densidad*		*Media*	*Varianza*
Uniforme	$\frac{1}{b-a}$ $a < x < b$	$a \in R$ $b \in R$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mu \in R$ $\sigma > 0$	μ	σ^2
Exponencial	$\lambda e^{-\lambda x}$	$\lambda, x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	$\frac{p^a}{\Gamma(a)} e^{-px} x^{a-1}$ $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$	$a, p > 0$ $x > 0$	$\frac{a}{p}$	$\frac{a}{p^2}$
Beta	$\frac{1}{B(p,q)} x^{p-1} (1-x)^{q-1}$ $B(p,q) = \frac{\Gamma(p)\Gamma(p)}{\Gamma(p+q)}$	$p, q > 0$ $0 < x < 1$	$\frac{p}{p+q}$	$\frac{pq}{\frac{p+q+1}{(p+q)^2}}$
Chi - cuadrado	$\frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-x/2}$	$x \geq 0$ $n = 1,\ldots$	n	$2n$
T de Student	$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	$n = 1,\ldots$	0	$\frac{\Gamma(\frac{3}{2})\Gamma(\frac{n-2}{2})n}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}$
F de Fisher	$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{mx}{n}\right)^{-\frac{m+n}{2}}$	$x > 0$ $m = 1,\ldots$ $n = 1,\ldots$	$\frac{n}{n-2}$	$\frac{n^2(2m+2n-4)}{m(n-2)^2(n-4)}$
Weibull	$\frac{\beta}{\eta^\beta}(x-\gamma)^{\beta-1} e^{-\left(\frac{x-\gamma}{\eta}\right)^\beta}$	$x \geq \gamma$ $\beta, \eta > 0$ $\gamma \in R$	$\gamma + \eta\Gamma(1 + \frac{1}{\beta})$	$\frac{\beta^2}{\alpha}[2\Gamma(\frac{2}{\alpha}) - \frac{1}{\alpha}(\Gamma(\frac{1}{\alpha}))^2]$
Log - normal	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-(Lnx-\mu)^2/2\sigma^2}$	$\mu \in R$ $x > 0$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$
Laplace y *Cauchy*	$\frac{\lambda}{2} e^{-\lambda\lvert x-\alpha\rvert}$ $\left(\pi\beta\left[1 + \left(\frac{x-\alpha}{\beta}\right)^2\right]\right)^{-1}$	$\alpha \in R$ $\lambda > 0$ $\beta > 0$	α	$\frac{2}{\lambda^2}$

© Alfaomega - RC Libros

La función de distribución de *la variable gamma (z,p)* cumple:

$$F(x) = 1 - \int_x^{\infty} \frac{1}{\Gamma(z)} p^z e^{-pt} t^{z-1} dt = 1 - \sum_{k=0}^{z-1} e^{-px} \frac{(px)^k}{k!} \text{ si } z \text{ es entero}$$

Por lo tanto, la *distribución gamma* puede considerarse como una suma de *z* variables de Poisson de parámetro *px*. De esta forma puede considerarse la distribución de Poisson como un modelo del número de ocurrencias de algún evento en un intervalo fijo, y a la distribución gamma como el modelo de la proporción de intervalos necesaria para obtener un número específico de ocurrencias. Esta distribución se usa en teoría de la fiabilidad, fenómenos de espera y teoría de colas.

En cuanto a la *distribución exponencial*, observamos que la suma de *z* variables independientes y exponenciales de parámetro *p* será una gamma *(z,p)*, y que la distribución exponencial de parámetro *p* es una gamma *(1,p)*.

La exponencial se utiliza como modelo adecuado para la duración de la vida de los equipos que no envejecen. Representa en el campo continuo un papel similar al de la geométrica en el campo discreto. La distribución exponencial se utiliza ampliamente en el campo de la ingeniería de la confiabilidad, como un modelo para el tiempo hasta el fallo de un componente o de un sistema. En estas aplicaciones el parámetro *p* se llama índice de fallo del sistema, y la media de la distribución *1/p* se denomina tiempo medio hasta el fallo. La exponencial se usa como modelo de fiabilidad cuando el índice de fallos es constante. También se utiliza la exponencial en modelos que definen el número de ocurrencias de determinados eventos en intervalos de tiempo.

Existe una relación importante entre las distribuciones exponencial y de Poisson, ya que si se considera esta última como un modelo del número de ocurrencias de algún evento en el intervalo *(0,t)*, entonces $P(x)= e^{-pt}(pt)^x/x!$. Suponer *x=0* es suponer que el evento no ocurre en el intervalo *(0,t]*, y $P(x=0) = P(0) = e^{-pt}$. Es entonces posible considerar a *P(0)* como la probabilidad de que el intervalo hasta la primera ocurrencia del evento sea mayor que *t*, esto es, $P(Y>t) = P(0) = e^{-pt}$ donde *Y* es la variable aleatoria que denota el intervalo hasta la primera ocurrencia del evento considerado. La función de distribución de *Y* será $F(t) = P(Y\leq t) = 1-e^{-pt}$. La función de densidad de *Y* será $f(y)=F'(y)$, es decir, $f(y) = pe^{-py}$, con lo que *Y* es una exponencial de parámetro *p*. Esto es, la distribución exponencial es la distribución del intervalo hasta la primera ocurrencia de un determinado evento. Por tanto, si el número de ocurrencias de un evento tiene distribución de Poisson con parámetro *p*, entonces la distribución del intervalo entre dos ocurrencias será una exponencial de parámetro *p*.

© Alfaomega - RC Libros

En cuanto a la *distribución beta*, suele utilizarse como distribución de una proporción aleatoria, como modelo aproximado en la ausencia de datos, y como ajuste a distribuciones empíricas. Si F_b es la distribución de una binomial *(n,p)* y F_B es la distribución de una beta (α,β), entonces $F_b(x) = 1-F_B(p)$, con $\alpha = x$, $\beta = n - x - 1$.

En cuanto a la *distribución chi-cuadrado*, se tiene $\chi_n^2 = X_1^2 + X_2^2 + \ldots + X_n^2$ con $X_i \rightarrow N(0,1)$ e independientes. La variable chi-cuadrado con *n* grados de libertad equivale a una gamma *(n/2,1/2)*, y si *n>30* $\sqrt{2\chi_n^2} \rightarrow N(\sqrt{2n-1},1)$.

Una propiedad importante de la distribución chi-cuadrado es la siguiente: dada una muestra aleatoria de tamaño n, $X_1, X_2, \ldots, X_n$ de una distribución normal $N(\mu,\sigma)$, el estadístico $(n-1)S^2$, con S^2 = cuasivarianza muestral, es una chi-cuadrado con $n-1$ grados de libertad, y $\bar{X}$ y S^2 son independientes. Este resultado, conocido como teorema de Cochran, es de aplicación en control de calidad para la obtención de límites de control en los diagramas de S^2. La chi-cuadrado es una distribución esencial en inferencia estadística (fiabilidad, tests de independencia, ajuste, homogeneidad, etc.).

En cuanto a la *distribución T de Student*, sea *Z* una variable aleatoria *N*(0,1) y sea χ_n^2 una variable aleatoria chi-cuadrado con *n* grados de libertad. Entonces, si *Z* y χ_n^2 son independientes, la variable

$$T = \frac{Z}{\sqrt{\chi_n^2 / n}}$$

es una T de Student con *n* grados de libertad. Se trata de una distribución esencial en inferencia estadística, sobre todo al tratar con muestras pequeñas y con poblaciones normales de varianza desconocida. También se usa en contrastes de la significatividad individual de los parámetros estimados en un modelo econométrico.

En cuanto a la *distribución F de Fisher Snedecor*, sean $\chi_1^2(m)$ y $\chi_2^2(n)$ dos variables aleatorias chi-cuadrado con *m* y *n* grados de libertad respectivamente, e independientes. Entonces, la variable aleatoria:

$$F = \frac{\chi_1^2(m)/m}{\chi_2^2(n)/n}$$

© Alfaomega - RC Libros

tiene una distribución *F* con *m* grados de libertad del numerador y *n* grados de libertad del denominador. Una propiedad importante de la *F* es que el cuadrado de una *T*(*n*) de Student es una *F*(*1,n*). Otra propiedad importante de la *F* es que si *X* es una *F*(*m,n*), entonces la variable 1/*X* es una *F*(*n,m*), y la variable

$$\frac{1}{1+\frac{m}{n}X}$$

es una beta (*n/2,m/2*). Se trata de una distribución fundamental en las técnicas del análisis de la varianza, el diseño de experimentos, en inferencias de la razón de varianzas y en contrastes de la significatividad conjunta de los parámetros estimados en un modelo econométrico.

En cuanto a la *distribución log-normal*, se cumple que si *X* es $N(\mu,\sigma)$, entonces *Ln*(*X*) es log-normal (μ,σ). Esta distribución se utiliza en la teoría de la supervivencia como modelo de ajuste de vida, y en la teoría de la fiabilidad para analizar la ley de vida de equipos cuando los fallos se producen por desgaste, y como ley de duración de reparaciones en los modelos de mantenimiento. También se usa en la teoría de errores de medida.

En cuanto a las distribuciones de Weibull y Erlang, se utilizan en la teoría de la supervivencia como modelo de ajuste de vida, así como en la teoría de la fiabilidad y mantenimiento. La distribución de Weibull se usa como modelo preferido de ajuste de datos a una distribución de frecuencias cuando no hay normalidad, y la de Erlang es la distribución del tiempo transcurrido entre las ocurrencias *k* y *k+1* de un proceso de Poisson de media β. La exponencial *Exp*(β) es caso particular de ambas distribuciones.

DESIGUALDAD DE TCHEBICHEV

El teorema de Tchebichev dice que la probabilidad de que cualquier variable aleatoria *X* tome un valor dentro de *k* desviaciones típicas de la media es al menos $1-1/k^2$. Esto se puede expresar de la siguiente forma:

$$P(\mu-k\sigma<X<\mu+k\sigma)\geq 1-\frac{1}{k^2}\Leftrightarrow P(-k\sigma<X-\mu<k\sigma)\geq 1-\frac{1}{k^2}$$

© Alfaomega - RC Libros

La expresión del teorema también puede escribirse como sigue:

$$P(|X-\mu|<k\sigma)\geq 1-\frac{1}{k^2}\Leftrightarrow P(|X-\mu|<k)\geq 1-\frac{\sigma^2}{k^2}\Leftrightarrow P(|X-\mu|>k)\leq\frac{\sigma^2}{k^2}$$

Para el caso de una muestra aleatoria de tamaño *n* de la variable *X* podemos escribir la desigualdad de Tchebichev como sigue:

$$P(|\overline{X}-\mu|<k\frac{\sigma}{\sqrt{n}})\geq 1-\frac{1}{k^2}\Leftrightarrow P(|\overline{X}-\mu|<k)\geq 1-\frac{\sigma^2}{nk^2}\Leftrightarrow P(|\overline{X}-\mu|>k)\leq\frac{\sigma^2}{nk^2}$$

TEOREMA CENTRAL DEL LÍMITE Y TEOREMA DE MOIVRE

Sea $(x_1,x_2,\ldots,x_n)$ una muestra de tamaño *n* de una variable aleatoria X con esperanza μ finita y varianza σ^2 también finita. Se demuestra lo siguiente:

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\rightarrow N(0,1)$$

En la práctica, suele darse validez a esta aproximación cuando n>30.

La expresión anterior también puede escribirse como sigue:

$$\frac{\overline{X}_n-E(\overline{X}_n)}{\sigma^2(\overline{X}_n)}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu}{\sigma/\sqrt{n}}=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sigma\sqrt{n}}\rightarrow N(0,1)$$

Si aplicamos el teorema central del límite a una muestra de *n* variables de Bernouilli de parámetro *p* ($\mu=p$, $\sigma^2=pq$) tenemos lo siguiente:

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu}{\sigma/\sqrt{n}}=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sigma\sqrt{n}}=\frac{Y-np}{\sqrt{npq}}\rightarrow N(0,1)$$

© Alfaomega - RC Libros

siendo $Y = \sum_{i=1}^{n} X_i$ una binomial (*n,p*) por ser suma de *n* variables de Bernouilli de parámetro *p* independientes.

Esta última variante del teorema central del límite se denomina *teorema de Moivre* y permite aproximar una distribución binomial por una normal cuando *n* es grande ($n > 30$).

R Y LAS DISTRIBUCIONES CONTINUAS

R dispone de varias funciones que permiten trabajar con distribuciones continuas. En general, R permite trabajar con la ley de probabilidad, la función de distribución y los cuartiles de cualquier distribución continua. Asimismo, permite generar números aleatorios según cualquier distribución continua.

Las funciones esenciales de R para el trabajo con distribuciones continuas se presentan a continuación:

dnorm (x,a,b)	*Halla la función de densidad en x de una Normal(a,b)*
pnorm(x,a,b)	*Halla la función de distribución en x de una Normal(a,b)*
qnorm(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Normal(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Normal(a,b)*
rnorm(m,a,b)	*Halla m números aleatorios de una Normal(a,b)*
dunif(x,a,b)	*Halla la función de densidad en x de una Uniforme(a,b)*
punif(x,a,b)	*Halla la función de distribución en x de una Uniforme(a,b)*
qunif(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Uniforme(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Uniforme(a,b)*
runif(m,a,b)	*Halla m números aleatorios de una Uniforme(a,b)*
dexp(x,a)	*Halla la función de densidad en x de una Exponencial de parámetro a*

© Alfaomega - RC Libros

pexp(x,a)	*Halla la función de distribución en x de una Exponencial de parámetro a*
qexp(q,a)	*Halla el valor del cuantil de orden q de una distribución exponencial de parámetro a. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Exponencial de parámetro a*
rexp(m,a)	*Halla m números aleatorios de una Exponencial de parámetro a*
dgamma(x,a)	*Halla la función de densidad en x de una Gamma(a,b)*
pgamma(x,a,b)	*Halla la función de distribución en x de una Gamma(a,b)*
qgamma(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Gamma(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Gamma(a,b)*
rgamma(m,a,b)	*Halla m números aleatorios de una Gamma(a,b)*
dbeta(x,a,b)	*Halla la función de densidad en x de una Beta(a,b)*
pbeta(x,a,b)	*Halla la función de distribución en x de una Beta(a,b)*
qbeta(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Beta(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Beta(a,b)*
rbeta(m,a,b)	*Halla m números aleatorios de una Beta(a,b)*
dchisqf(x,a)	*Halla la función de densidad en x de una Chi-cuadrado con a grados de libertad*
pchisq(x,a)	*Halla la función de distribución en x de una Chi-cuadrado con a grados de libertad*
qchisq(q,a)	*Halla el valor del cuantil de orden q de una distribución Chi-cuadrado con a grados de libertad. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Chi-cuadrado con a grados de libertad*
rchisq(m,a)	*Halla m números aleatorios de una Chi-cuadrado con a grados de libertad*

© Alfaomega - RC Libros

df(x,a,b)	*Halla la función de densidad en x de una F de Fisher con (a,b) grados de libertad*
pf(x,a,b)	*Halla la función de distribución en x de una F de Fisher con (a,b) grados de libertad*
qf(q,a,b)	*Halla el valor del cuantil de orden q de una distribución F de Fisher con (a,b) grados de libertad. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una F de Fisher con (a,b) grados de libertad*
rf(m,a,b)	*Halla m números aleatorios de una F de Fisher con (a,b) grados de libertad*
dt(x,a)	*Halla la función de densidad en x de una T de Student con a grados de libertad*
pt(x,a)	*Halla la función de distribución en x de una T de Student con a grados de libertad*
qt(q,a)	*Halla el valor del cuantil de orden q de una distribución T de Student con a grados de libertad. Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una T de Student con a grados de libertad*
rt(m,a)	*Halla m números aleatorios de una T de Student con a grados de libertad*
dcauchy(x,a,b)	*Halla la función de densidad en x de una Cauchy(a,b)*
pcauchy(x,a,b)	*Halla la función de distribución en x de una Cauchy(a,b)*
qcauchy(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Cauchy(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Cauchy(a,b)*
rcauchi(m,a,b)	*Halla m números aleatorios de una Cauchy(a,b)*
dweibull(x,a,b)	*Halla la función de densidad en x de una Weibull(a,b)*
pweibull(x,a,b)	*Halla la función de distribución en x de una Weibull(a,b)*

© Alfaomega - RC Libros

qweibull(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Weibull(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Weibull(a,b)*
rweibull(m,a,b)	*Halla m números aleatorios de una Weibull(a,b)*
dlnorm(x,a,b)	*Halla la función de densidad en x de una Lognormal(a,b)*
plnorm(x,a,b)	*Halla la función de distribución en x de una Lognormal(a,b)*
qlnorm(q,a,b)	*Halla el valor del cuantil de orden q de una distribución Uniforme(a,b). Es decir, calcula $F^{-1}(q)$ siendo F la función de distribución de una Lognormal(a,b)*
rlnorm(m,a,b)	*Halla m números aleatorios de una Lognormal(a,b)*

Ejercicio 8-1. Un agente de seguros vende pólizas a 5 individuos, todos de la misma edad. De acuerdo con las tablas actuariales, la probabilidad de que un individuo con esa edad viva 30 años más es de 3/5. Determinar la probabilidad de que dentro de 30 años vivan:

- Los 5 individuos.
- Al menos 3.
- Solo 2.
- Al menos 1.

Como quiera que dentro de 30 años la situación de cada individuo será que viva o que no viva, y al menos una de las dos se ha de presentar, la situación para cada individuo se ajusta a una variable de Bernoulli con probabilidad de éxito (vivir 30 años más) igual a 3/5=0,6. Al considerar los 5 individuos, estamos ante una variable aleatoria X binomial con n=5 y p = 0,6, $X = B(5, 0{,}6)$.

1. Habrá que calcular $P(X = 5)$.
2. Habrá que calcular $P(X \geq 3)$, o lo que es lo mismo, $1 - p(X < 3) = 1 - P(X \leq 2)$
3. Habrá que calcular $P(X = 2)$.
4. Habrá que calcular $P(X \geq 1)$, o lo que es lo mismo, $1 - P(X=0)$.

© Alfaomega - RC Libros

Ahora ya podemos calcular las probabilidades del problema, que serán:

1. $P(X = 5) = 0{,}07776$

```
> dbinom(5,5,0.6)
[1] 0.07776
```

2. $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 0{,}68256$

```
> 1-pbinom(2,5,0.6)
[1] 0.68256
```

3. $P(X = 2) = 0{,}2304$

```
> dbinom(2,5,0.6)
[1] 0.2304
```

4. $P(X \geq 1) = 1 - P(X = 0) = 1 - 0{,}01024 = 0{,}98976$

```
> 1-dbinom(0,5,0.6)
[1] 0.98976
```

Ejercicio 8-2. La proporción de alumnos de un distrito universitario con calificación media de sobresaliente es de 0,005%. Determinar la probabilidad de que entre 5000 alumnos seleccionados al azar haya dos con calificación media de sobresaliente.

Teóricamente, la variable *X* definida por el número de alumnos (de entre los 5000) que tienen una calificación media de sobresaliente es una variable binomial con $n = 5000$ y $p = 0{,}00005$. La probabilidad pedida sería $P(X=2)$. Pero este cálculo con números tan extremos es molesto. Ahora bien, habida cuenta de que *p* es muy pequeño ($p = 0{,}00005$) y de que *n* es muy grande ($n = 5000$), y además $m = np = 0{,}25 < 5$ y $p < 0{,}1$, ya podemos aproximar la variable $X=B(5000, 0{,}00005)$ por una variable de Poisson de parámetro $m = np = 5000*0{,}00005 = 0{,}25$.

Luego, $P(X=2)$ lo podemos calcular mediante la distribución de Poisson de parámetro 0,25 como sigue:

$$P(X = 2) = e^{-0{,}25}\frac{0{,}25^2}{2!} = 0{,}024337524$$

```
> dpois(2,0.25)
[1] 0.02433752
```


© Alfaomega - RC Libros

También podríamos haber calculado $P(X=2)$ directamente mediante la distribución binomial $B(5000, 0{,}00005)$ como sigue:

$$P(X = 2) = \binom{5000}{2} 0{,}00005^2 (1 - 0{,}00005)^{5000-2} = 0{,}024334938$$

```
> dbinom(2,5000, 0.00005)
[1] 0.02433494
```

Ejercicio 8-3. Los datos provenientes de un censo clasifican la proporción de adultos según la siguiente tabla:

Edad	*Proporción*
18 – 24	0,18
25 – 30	0,23
35 – 44	0,16
45 – 64	0,27
65 –	0,16

Se seleccionan al azar cinco adultos de esa población y se pide calcular la probabilidad de que la muestra contenga una persona entre 18 y 24 años, dos personas entre 25 y 34 años, y dos personas entre las edades de 45 y 64 años.

Si numeramos las cinco clases de edades como 1, 2, 3, 4 y 5 desde arriba hacia abajo y suponemos que las proporciones dadas son las probabilidades de cada clase, podemos ajustar el problema a una distribución multinomial con la siguiente ley de probabilidad:

$$P(y_1, y_2, y_3, y_4, y_5) = \frac{n!}{y_1! y_2! y_3! y_4! y_5!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5}$$

Nuestro problema se resuelve tomando:

$$n = 5, y_1 = 1, y_2 = 2, y_3 = 0, y_4 = 2, y_5 = 0$$

Tendremos:

© Alfaomega - RC Libros

$$P(1,2,0,2,0)=\frac{5!}{1!2!0!2!0!}(0{,}18)^1(0{,}23)^2(0{,}16)^0(0{,}27)^2(0{,}16)^0=0{,}0208$$

```
> x=c(1,2,0,2.0)
> prob=c(0.18,0.23,0.16,0.27,0.16)
> dmultinom(x,size=5,prob)
[1] 0.02082461
```

Ejercicio 8-4. De un grupo de tres republicanos, dos demócratas y un independiente se selecciona al azar un comité de tres personas. Sea y_1 el número de republicanos, y_2 el número de demócratas e y_3 el número de independientes en el comité. Calcular la probabilidad de que en el comité haya exactamente una persona de cada afiliación política.

Este problema puede ajustarse a una distribución hipergeométrica generalizada con la siguiente distribución de probabilidad:

$$P(y_1=a,y_2=b,y_3=c)=\frac{\binom{3}{a}\binom{2}{b}\binom{1}{c}}{\binom{5}{a+b+c}}$$

En el problema nos piden lo siguiente:

$$P(y_1=1,y_2=1,y_3=1)=\frac{\binom{3}{1}\binom{2}{1}\binom{1}{1}}{\binom{5}{1+1+1}}=\frac{3}{5}$$

Ejercicio 8-5. Un coleccionista desea capturar un ejemplar de una clase de mariposas que se encuentra con un porcentaje del 15%. Hallar la probabilidad de que tenga que cazar 10 mariposas de la clase no deseada antes de encontrar:

- Un ejemplar de la clase deseada.
- Tres ejemplares de la clase deseada.

El número de ejemplares X capturados antes de encontrar uno de la clase deseada sigue una distribución geométrica de parámetro p = 0,15 ya que se trata de

© Alfaomega - RC Libros

un tiempo de espera hasta el primer éxito (encontrar una mariposa de la clase deseada).

Luego, la probabilidad pedida en el primer apartado se calculará así:

$$P(X=10)=(0{,}15)(0{,}85^{10})=0{,}02953$$

```
> dgeom(10,0.15)
[1] 0.02953116
```

En el segundo apartado nos encontramos ante una binomial negativa, ya que se trata de calcular la probabilidad de un tiempo de espera hasta el *r*-ésimo éxito (tercer éxito en nuestro caso). La probabilidad pedida se calcula como sigue:

$$P(X=10,3)=\binom{10+3-1}{3-1}(0{,}15)^3(0{,}85)^{10}=0{,}04385$$

```
> dnbinom(10,3,0.15)
[1] 0.04385377
```

Ejercicio 8-6. El número medio de automóviles que llega a una estación de suministro de gasolina es de 210 por hora. Si dicha estación puede atender a un máximo de 10 automóviles por minuto, determinar la probabilidad de que en un minuto dado lleguen a la estación de suministro más automóviles de los que puede atender. Hallar y representar gráficamente la función de distribución de la variable aleatoria número aleatorio de automóviles que llegan a la estación de servicio en un minuto. Hallar también la ley de probabilidad y representarla gráficamente.

El número aleatorio de automóviles que llegan a la estación de servicio en un minuto puede representarse por una variable *X* de Poisson de parámetro *m* = 210/60 = 3,5, ya que *m* es el número medio de llegadas por minuto (teníamos 210 llegadas a la hora).

La probabilidad que vamos a obtener vendrá dada por *P(X >10)*, ya que para que lleguen a la estación más automóviles por minuto de los que se pueden atender es necesario que lleguen más de 10 por minuto.

Para calcular dicha probabilidad, utilizamos que *P(X>10)=1– p(X≤10)*, siendo *p(X≤10)* la función de distribución de una variable de Poisson de parámetro 3,5.

© Alfaomega - RC Libros

```
> 1-ppois(10,3.5)
[1] 0.001019394
```

Para representar la función de distribución, tendremos presente que es una función escalonada (Figura 8-1). Por lo tanto, consideraremos varios puntos de su campo de definición para conseguir una representación más fina.

```
> x=0:10
> plot(ppois(x,3.5), type="s")
```

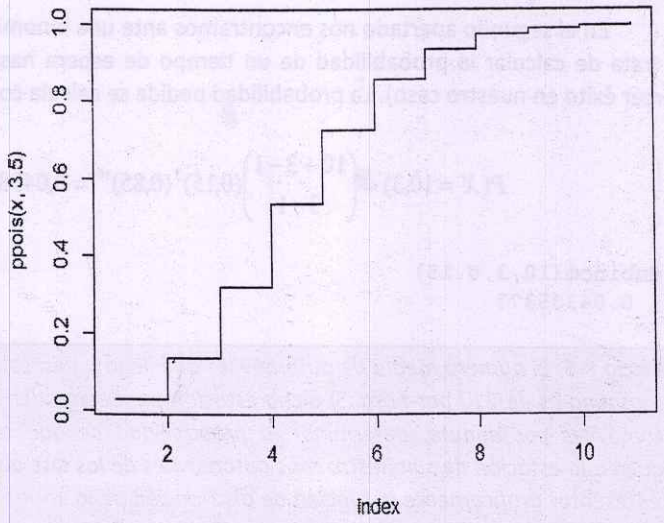


Figura 8-1

Para representar la ley de probabilidad (Figura 8-2), seleccionamos su rango de valores *(x,f(x))* y utilizamos las fórmulas adecuadas.

```
> windows()
> x=0:10
> plot(dpois(x,3.5), type="l")
```


© Alfaomega - RC Libros

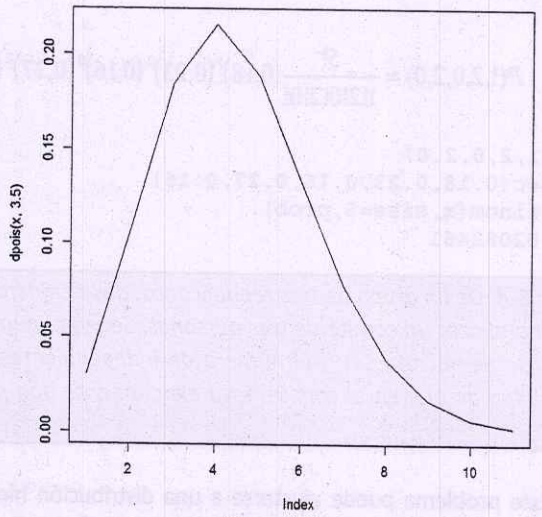


Figura 8-2

Ejercicio 8-7 En un proceso de fabricación, se sabe que el número aleatorio de unidades defectuosas producidas diariamente viene dado por una variable de Poisson de parámetro 10. Determinar la probabilidad de que en 150 días el número de unidades defectuosas producidas supere las 1480 unidades.

Sea X_i la variable definida por el número de piezas defectuosas producidas en el día *i*. X_i es una variable de Poisson de parámetro 10. El número de piezas defectuosas producidas en 150 días vendrá dado por la variable $Y=\sum X_i$, con $i=1...150$, que será una variable de Poisson de parámetro 1500. El problema nos pide $P(Y>1480)$, valor que podría calcularse con la distribución de Poisson, pero al ser el parámetro tan alto, utilizamos la aproximación a la distribución normal.

Las variables X_i tienen $E(X_i) = 10$ y $V(X_i) = 10$ por ser de Poisson con $\lambda = 10$.

Por el teorema central del límite tenemos:

$$\frac{\frac{1}{150}\sum_{i=1}^{150} X_i - \lambda}{\sqrt{\lambda}/\sqrt{n}} = \frac{\sum_{i=1}^{150} X_i - 150\lambda}{\sqrt{n\lambda}} \to N(0,1) \Rightarrow Y = \sum_{i=1}^{150} X_i \to N(150\lambda, \sqrt{n\lambda}) = N(1500;38,7)$$

© Alfaomega - RC Libros

El problema nos pide $P(Y>1480)=1-P(Y \leq 1480) = 1-F_{N(1500;38,7)}(1480) = 0,697$

```
> 1-pnorm(1480,1500,38.7)
[1] 0.6973507
```

Ejercicio 8-8. Hallar la mediana y el primer y tercer cuartil para una distribución beta de parámetros 3 y 2.

La mediana es el valor de la variable para el que la función de distribución vale 0,5; así pues, hemos de obtener k tal que $p(X<k) = 0,5$, cuyo valor se halla mediante el valor crítico de una beta(3,2) en el punto 0,5; es decir, $F^{-1}_{\beta(3,2)}(0,5)$ cuyo resultado es 0,614. Ya sabemos que el valor de la mediana es 0,614.

```
> qbeta(0.5,3,2)
[1] 0.6142724
```

El primer cuartil es el valor de la variable para el que la función de distribución vale 0,25; así pues, hemos de hallar k tal que $p(X<k) = 0,25$, cuyo valor se halla mediante el valor crítico de una beta(3,2) en el punto 0,25; es decir, $F^{-1}_{\beta(3,2)}(0,25)$ cuyo resultado es 0,456. Ya conocemos que el valor del primer cuartil es 0,456.

```
> qbeta(0.25,3,2)
[1] 0.4563217
```

El tercer cuartil es el valor de la variable para el que la función de distribución vale 0,75; así pues, hemos de hallar k tal que $p(X<k) = 0,75$, cuyo valor se halla mediante el valor crítico de una beta(3,2) en el punto 0,75; es decir, $F^{-1}_{\beta(3,2)}(0,75)$ cuyo resultado es 0,7569. Ya tenemos que el valor del tercer cuartil es 0,7569.

```
> qbeta(0.75,3,2)
[1] 0.7569779
```

Ejercicio 8-9. Sea una sucesión de variables aleatorias independientes $X_1, X_2, \ldots, X_n$ con la misma función de densidad: $f(x) = 8\,e^{-8x}$ si $x > 0$, y $f(x) = 0$ si $x \leq 0$. Si consideramos 100 de estas variables, calcúlese:

- La probabilidad de que su suma sea menor o igual a 10.
- La probabilidad de que la suma esté entre 11 y 13.

© Alfaomega - RC Libros

Por la forma de su función de densidad, estas variables son exponenciales de parámetro $p = 8$. Por otra parte, la suma de 100 variables exponenciales independientes de parámetro 8 será una gamma (100,8).

Sea $Y = X_1+X_2+ \ldots +X_{100}$. Y es una variable aleatoria gamma (100,8).

En la primera pregunta, el problema nos pide $P(Y \leq 10) = F(10) = 0,0171$, donde F es la función de distribución de una gamma (100,8).

```
> pgamma(10,100,8)
[1] 0.01710831
```

En la segunda pregunta, el problema nos pide $P(11< Y <13)$, siendo Y una gamma (100,8).

$P(11< Y <13) = P(Y <13) - P(Y \leq 11) = F(13) - F(11) = 0,554$, donde F es la función de distribución de una variable gamma (100,8).

```
> pgamma(13,100,8)-pgamma(11,100,8)
[1] 0.554066
```

Ejercicio 8-10. Calcular las siguientes probabilidades:

$P(\chi 2\,(10) > 5,031)$
$P(6,821 < \chi 2(15) < 15,13)$
$P(\chi 2\,(150) > 128)$
$P(50 < \chi 2\,(65) < 60)$

Nota: los números entre paréntesis son los grados de libertad de la chi-cuadrado.

1) $P(\chi^2(10)> 5,031) = 1 - P(\chi^2(10) \leq 5,031,10)$, cuyo valor es 0,889097.

```
> 1-pchisq(5.031,10)
[1] 0.8890976
```

2) $P(6,821< \chi^2(15) < 15,13) = P(\chi^2(15) < 15,13) - P(\chi^2(15) < 6,821)$, cuyo valor es 0,5203.

```
> pchisq(15.13,15)-pchisq(6.821,15)
[1] 0.5203335
```

© Alfaomega - RC Libros

3) $P(\chi^2(150) > 128) = P(\sqrt{2\chi^2(150)} > \sqrt{2*128}) = P(Y > 16)$ siendo Y una $N(\sqrt{2*150-1},1)$. En este caso hemos aproximado la chi-cuadrado por la normal, ya que los grados de libertad son muy elevados (>30). El valor es 0,90175.

```
> 1-pnorm(16,sqrt(2*150-1),1)
[1] 0.901755
```

4) $P(50 < \chi^2(65) < 60) = P(\chi^2(65) < 60) - P(\chi^2(65) < 50) = P(\sqrt{2\chi^2(65)} < \sqrt{2*60}) - P(\sqrt{2\chi^2(65)} < \sqrt{2*50}) = p(Y < \sqrt{120}) - p(Y < 10)$, siendo Y una $N(\sqrt{2*65-1},1)$. El valor es 0,256.

```
> pnorm(sqrt(120),sqrt(2*65-1),1)-pnorm(10,sqrt(2*65-1),1)
[1] 0.2560787
```

Ejercicio 8-11. Calcular las siguientes probabilidades:

1. $P(T(5) < 1{,}239)$
2. $P(1{,}83 < T(22) < 2{,}04)$
3. $P(T(14) < -1{,}5)$
4. $P(-1 < T(17) < 2)$

siendo T una variable aleatoria de Student con los grados de libertad indicados entre paréntesis.

1) $P(T(5) < 1{,}239) = F(1{,}239)$, cuyo valor es 0,86483.

```
> pt(1.239,5)
[1] 0.8648334
```

2) $P(1{,}83 < T(22) < 2{,}04) = P(T(22) < 2{,}04) - P(T(22) < 1{,}83)$, cuyo resultado es 0,01365.

```
> pt(2.04,22)-pt(1.83,22)
[1] 0.01364937
```

3) $P(T(14) < -1{,}5) = 1 - P(T(14) < 1{,}5)$, ya que la T es simétrica respecto del origen (lo mismo que la N(0,1)). El resultado es 0,0779.

```
> 1-pt(1.5,14)
[1] 0.07791266
```


© Alfaomega - RC Libros

4) $P(-1 < T(17) < 2) = P(T(17) < 2) - P(T(17) < -1) = P(T(17) < 2) - (1 - P(T(17) < 1) = P(T(17) < 2) + P(T(17) < 1) - 1$, cuyo resultado es 0,803464.

```
> pt(2,17)-pt(1,17-1)
[1] 0.1352282
```

Ejercicio 8-12. Se supone que la estatura y el peso de las personas no son variables aleatorias independientes. Para verificar si esta hipótesis es cierta, y suponiendo que la distribución conjunta es normal, se recurre a contrastar la hipótesis de que el coeficiente de correlación poblacional es significativamente igual a cero mediante una muestra aleatoria simple de tamaño 10. Se obtiene la región crítica |T|>1,8595, siendo T el estadístico:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad r = \text{coeficiente de correlación muestral}$$

Hallar el nivel α de significación de este test.
Hallar el valor crítico correspondiente a un nivel α=0,05 (es decir, hallar k tal que $P(|T|>k)=0{,}05$).

1) El nivel de significación de un test de hipótesis es la probabilidad de la región crítica bajo la hipótesis nula. En este caso, la hipótesis nula es que el coeficiente de correlación poblacional es igual a cero, frente a la alternativa de que es distinto de cero. Pero con esta hipótesis nula, el estadístico T se distribuye según una variable aleatoria T de Student con $n-2 = 10 - 2 = 8$ grados de libertad. Luego: $\alpha = P(|T(8)|>1{,}8595)$, cuyo valor es aproximadamente 0,1.

$P(|T(8)|>1{,}8595) = 2(1-F_{T(8)}(1{,}8595)) = 0{,}1000074$

```
> 2*(1-pt(1.8595,8))
[1] 0.1000074
```

2) Para hallar k tal que $P(|T|>k) = 0{,}05$, hemos de hallar el valor crítico correspondiente a la probabilidad 0,05, cuyo valor es 2,306.

$P(|T|>k) = 0{,}05 \Rightarrow F_{T(8)}(k) = 0{,}975 \Rightarrow k = F^{-1}_{T(8)}(0{,}975) = 2{,}306$

```
> qt(0.975,8)
[1] 2.306004
```

© Alfaomega - RC Libros

Ejercicio 8-13. Comprobar gráficamente que la distribución exponencial de parámetro p es una gamma (1,p) y que una chi-cuadrado con n grados de libertad equivale a una gamma (n/2,1/2).

Para probar la equivalencia entre la exponencial y la gamma, tomamos *p = 6* o cualquier otro valor de p y graficamos las funciones de densidad de la exponencial de parámetro *p* y de la gamma (1,p). La Figura 8-3 muestra gráficas similares.

```
> windows()
> par(mfrow=c(1,2))
> x=seq(0,100,by=0.1)
> curve(dexp(x,6), type="l")
> curve(dgamma(x,1,6), type="l")
```

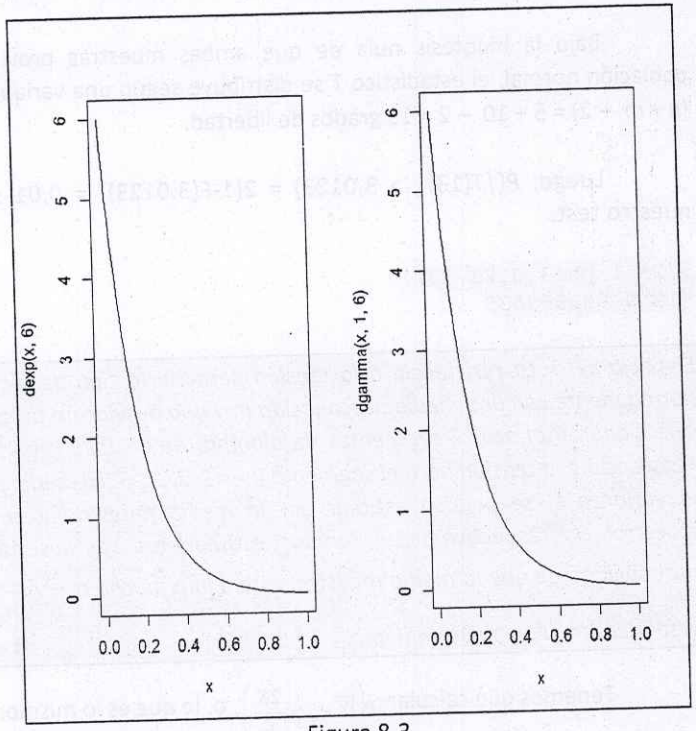


Figura 8-3

© Alfaomega - RC Libros

Para probar la equivalencia entre la chi-cuadrado con *n* grados de libertad y la gamma(n/2,1/2) tomamos *p = 15* o cualquier otro valor de *p* y graficamos las funciones de densidad de la chi-cuadrado con *n* grados de libertad y la gamma(n/2,1/2).

La Figura 8-4 muestra gráficas similares.

```
> windows()
> par(mfrow=c(1,2))
> x=seq(0,100,by=0.1)
> curve(dchisq(x,15), xlim=c(0,60),type="l")
> curve(dgamma(x,15/2,1/2), xlim=c(0,60),type="l")
```

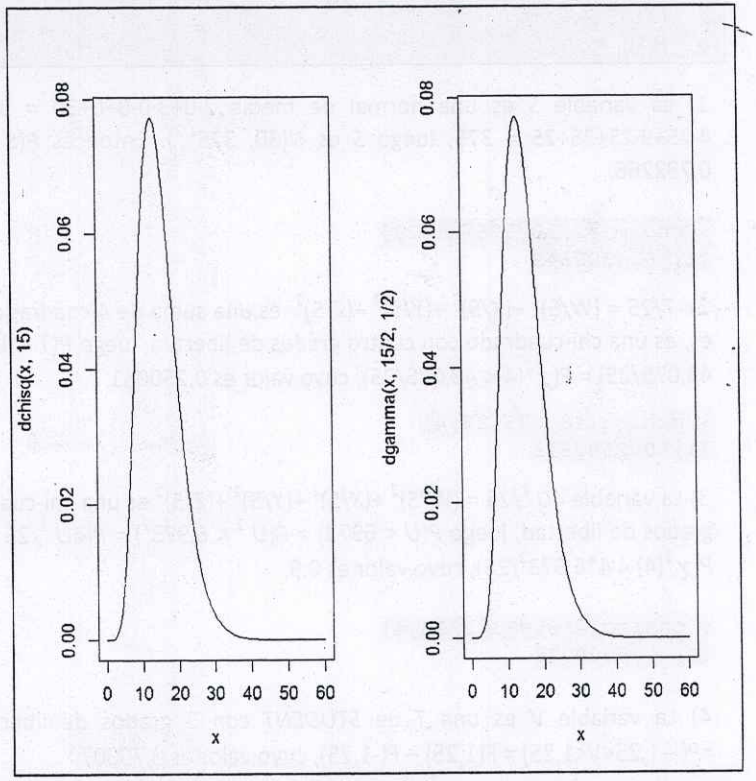


Figura 8-4

© Alfaomega - RC Libros

Ejercicio 8-14. Dadas cuatro variables aleatorias independientes N(0,5), W, X, Y, Z, se definen las siguientes variables aleatorias:

$$S = 2W + 3X - Y + Z + 30, \quad T = W^2 + X^2 + Y^2 + Z^2, \quad U = \sqrt{T/4}$$

$$V = \frac{W}{\sqrt{\frac{X^2 + Y^2 + X + Z^2}{3}}}$$

Calcular:

1. *P(S < 42)*
2. *P(T < 48,075)*
3. *P(U < 6,973)*
4. *P(|V| < 1,25)*

1) La variable *S* es una normal de media 2·0+3·0-0+0+30 = 30 y de varianza 4·25+9·25+25+25 = 375, luego *S* es $N(30, 375^{1/2})$. Entonces *P*(*S* < 42) = F(42) = 0,732266.

```
> pnorm(42,30,375^(1/2))
[1] 0.7322649
```

2) $T/25 = (W/5)^2 + (X/5)^2 + (Y/5)^2 + (Z/5)^2$ es una suma de 4 cuadrados de N(0,1), esto es, es una chi-cuadrado con cuatro grados de libertad, luego *P*(*T* < 48,075) = *P*(*T*/25 < 48,075/25) = $P(\chi^2(4) < 48{,}075/25)$, cuyo valor es 0,250081.

```
> pchisq(48.075/25,4)
[1] 0.2500813
```

3) La variable $4U^2/25 = (W/5)^2 + (X/5)^2 + (Y/5)^2 + (Z/5)^2$ es una chi-cuadrado con cuatro grados de libertad, luego $P(U < 6973) = P(U^2 < 6{,}973^2) = P(4U^2/25 < 4*6{,}973^2/25) = P(\chi^2(4) < 4*6{,}973^2/25)$, cuyo valor es 0,9.

```
> pchisq(4*6.973^2/25,4)
[1] 0.9000078
```

4) La variable *V* es una *T* de *STUDENT* con 3 grados de libertad. *P*(|*V*|<1,25) =*P*(–1,25<*V*<1,25) = F(1,25) – F(-1,25), cuyo valor es 0,70007.

```
> pt(1.25,3)-pt(-1.25,3)
[1] 0.7000705
```


© Alfaomega - RC Libros

Ejercicio 8-15. Se tienen dos muestras aleatorias simples de tamaños n=5 y m=10 obtenidas de dos variables X e Y, respectivamente. Se contrasta la hipótesis nula de que ambas muestras han sido extraídas de la misma población normal, y se obtiene la región crítica T > 3,0123, siendo T el estadístico:

$$T = \frac{\sqrt{\frac{nm}{n+m}}(\overline{Y} - \overline{X})}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{n+m-2}}}$$

Calcular el nivel de significación de este test.

El nivel de significación de un test de hipótesis es la probabilidad de la región crítica bajo la hipótesis nula.

Bajo la hipótesis nula de que ambas muestras provienen de la misma población normal, el estadístico *T* se distribuye según una variable *T* de Student con (*n* + *m* – 2) = 5 + 10 – 2 = 13 grados de libertad.

Luego, *P*(|*T*(13)| > 3,0123) = 2(1-F(3.0123)) = 0,01 será el nivel *a* de nuestro test.

```
> 2*(1-pt(3.0123,13))
[1] 0.009999535
```

Ejercicio 8-16. La resistencia a la tensión para cierto tipo de alambre se distribuye normalmente con una media desconocida m y una desviación típica desconocida s. Se seleccionaron al azar 6 segmentos de alambre de un rollo grande, y se midió Y_i la resistencia a la tensión para el segmento i, i = 1, 2, .., 6. La media de la población m y la varianza s^2 se pueden estimar por la media muestral y por la cuasivarianza muestral S^2, respectivamente (ambos estimadores son insesgados). Calcular la probabilidad de que la media muestral esté como mucho a $2s/\sqrt{n}$ de distancia de la verdadera media poblacional m.

Tenemos que calcular $P\left(|\overline{Y} - \mu| < \frac{2S}{\sqrt{n}}\right)$, o, lo que es lo mismo:

© Alfaomega - RC Libros

$$P\left(-\frac{2S}{\sqrt{n}}<\overline{Y}-\mu<\frac{2S}{\sqrt{n}}\right)=P\left(-2<\frac{\sqrt{n}(\overline{Y}-\mu)}{S}<2\right)=P(-2<T<2)$$

donde T es una variable T de Student con $n - 1 = 6 - 1 = 5$ grados de libertad. Tenemos que $P(-2 < T(5) < 2) = P(|T(5)| < 2) = F(2)-F(-2)$, cuyo valor es 0,89806.

```
> pt(2,5)-pt(-2,5)
[1] 0.8980605
```

Ejercicio 8-17. Supongamos que tenemos dos poblaciones normales de medias m_1 y m_2, y varianzas s_1^2 y s_2^2 desconocidas. Extraemos dos muestras aleatorias independientes de las dos poblaciones:

Muestra 1	Muestra 2	
0,9	1,6	0,8
0,1	2,8	0,0
0,2	0,8	1,2
1,1	2,3	1,6
0,8	2,3	1,1

Para contrastar la hipótesis nula $\sigma_2^2 / \sigma_1^2 = 1$, utilizamos el estadístico $T=\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$.

Siendo s_1^2 = cuasivarianza de la primera muestra y s_2^2 = cuasivarianza de la segunda muestra, obtenemos la región crítica $T > 2,8$. Se pide:

1. *Hallar el nivel de significación de este contraste.*
2. *Calcular la probabilidad de que la varianza de la segunda población sea mayor que la varianza de la primera.*
3. *Hallar el valor crítico correspondiente a un nivel de significación de 0,05.*

1) Bajo la hipótesis nula el estadístico T se distribuye según una variable aleatoria $F(n_1-1, n_2-1)$, donde n_1 y n_2 son los tamaños de las dos muestras (n_1= 5 y n_2=10). Por lo tanto, T es una $F(4,9)$ (estamos suponiendo que las dos poblaciones son normales). El nivel a de significación del test es $a = P(T > 2,8) = P(F(4,9)>2,8) = 1 - P(F(4,9) < 2,8)$, cuyo valor es 0,09193.

```
> 1-pf(2.8,4,9)
[1] 0.09193131
```

2) El problema nos pide:

© Alfaomega - RC Libros

$$P(\sigma_2^2>\sigma_1^2)=P\left(\frac{\sigma_2^2}{\sigma_1^2}>1\right)=P\left(\frac{\sigma_2^2\,S_1^2}{\sigma_1^2\,S_2^2}>\frac{S_1^2}{S_2^2}\right)=P\left(T>\frac{S_1^2}{S_2^2}\right)=P(T>\frac{\nu_1}{\nu_2})$$

siendo T una $F(4,9)$. Luego hemos de calcular $P(T>0,313/0,7161)$, cuyo valor es 0,7791.

```
> 1-pf(0.313/0.7161,4,9)
[1] 0.7791035
```

3) Tenemos que hallar un valor k tal que $P(T>k)=0,05$, o sea, $1 - P(T<k) = 0,05$, es decir, $P(T<k) = 0,95$. El valor de k es 3,63309.

```
> qf(0.95,4,9)
[1] 3.633089
```

Ejercicio 8-18. Un avión de apoyo táctico tiene una fiabilidad de 0,8 para determinada misión de apoyo a tierra. Calcular la probabilidad de efectuar más de 90 misiones sin fallo de un total de 100 salidas.

Se trata de calcular P(X > 90), siendo X una binomial (100, 0,8). Podemos resolver el problema directamente con la binomial, pero puede también utilizarse la aproximación de una binomial por la normal a través del teorema de Moivre de la siguiente forma:

$$P(X<90)=1-F(90)=1-F_{N(0,1)}\left(\frac{90-100\times 0,8}{\sqrt{100\times 0,8\times 0,2}}\right)=1-F_{N(0,1)}(2,5)=0,0062$$

```
> 1-pnorm(2.5,0,1)
[1] 0.006209665
```

Ejercicio 8-19. Utilizar la desigualdad de Tchebichev para hallar el número de veces que debe lanzarse una moneda para que la probabilidad de que la media muestral esté entre 0,4 y 0,6 sea al menos 0,90 suponiendo que la moneda está bien construida.

Observamos que puede escribirse lo siguiente:

$$P(0,4\le\overline{X}<0,6)\ge 0,90 \Leftrightarrow P(-0,1\le\overline{X}-0,5\le 0,1)\ge 0,90 \Leftrightarrow P(|\overline{X}-0,5|<0,1)\ge 0,90$$

© Alfaomega - RC Libros

Por otra parte, el teorema de Tchebichev aplicado a una media muestral de variables de Bernoulli (lanzamiento de una moneda bien construida) puede expresarse como sigue:

$$P(|\overline{X}-\mu|<k)\le 1-\frac{\sigma^2(\overline{X})}{k^2}=1-\frac{pq/n}{k^2}=1-\frac{pq}{nk^2}=1-\frac{\frac{1}{2}\frac{1}{2}}{nk^2}=1-\frac{1}{4nk^2}$$

Hemos aplicado que:

$$E(\overline{X})=\mu=p=1/2 \text{ y } \sigma^2(\overline{X})=\sigma^2/n=pq/n=1/(4n).$$

Si ahora comparamos las dos expresiones equivalentes anteriores:

$$P(|\overline{X}-0{,}5|<0{,}1)\ge 0{,}90$$

$$P(|\overline{X}-\mu|<k)\le 1-\frac{1}{4nk^2}$$

se observa que $1-\frac{1}{4nk^2}=0{,}90$ siendo k = 0,1. Luego hallaremos el tamaño muestral n (número de tiradas) resolviendo la siguiente ecuación:

$$1-\frac{1}{4n(0{,}1)^2}=0{,}90\Rightarrow 1-\frac{25}{n}=\frac{9}{10}\Rightarrow n=250$$

Ejercicio 8-20. ¿Cómo podría determinarse con más precisión el número necesario de tiradas en el problema anterior?

La forma de mejorar el resultado anterior es utilizar el teorema central del límite en vez del teorema de Tchebichev. Como el teorema central del límite es más preciso que el teorema de Tchebichev, es lógico que el tamaño muestral necesario (número de tiradas) sea menor cuando se aplica el teorema central del límite.

Aplicando el teorema central del límite para nuestros datos tenemos:

$$P(|\overline{X}-0{,}5|<0{,}1)\ge 0{,}90\Leftrightarrow P\left(\left|\frac{\overline{X}-0{,}5}{\sigma/\sqrt{n}}\right|<\frac{0{,}1}{\sigma/\sqrt{n}}\right)\ge 0{,}90\Leftrightarrow P\left(|Y_{N(0,1)}|<\frac{0{,}1\sqrt{n}}{\sigma}\right)\ge 0{,}90$$

© Alfaomega - RC Libros

Eliminando el módulo tenemos lo siguiente:

$$P\left(-\frac{0{,}1\sqrt{n}}{\sigma}<Y_{N(0,1)}<\frac{0{,}1\sqrt{n}}{\sigma}\right)\ge 0{,}90\Leftrightarrow F_{N(0,1)}\left(\frac{0{,}1\sqrt{n}}{\sigma}\right)-F_{N(0,1)}\left(-\frac{0{,}1\sqrt{n}}{\sigma}\right)=2F_{N(0,1)}\left(\frac{0{,}1\sqrt{n}}{\sigma}\right)-1\ge 0{,}9$$

Por tanto, puede calcularse n a partir de:

$$F_{N(0,1)}\left(\frac{0{,}1\sqrt{n}}{\sigma}\right)=0{,}95\Leftrightarrow F_{N(0,1)}\left(\frac{0{,}1\sqrt{n}}{\sqrt{1/4}}\right)=0{,}95\Leftrightarrow F_{N(0,1)}\left(0{,}2\sqrt{n}\right)=0{,}95\Leftrightarrow 0{,}2\sqrt{n}=1{,}645$$

```
> qnorm(0.95,0,1)
[1] 1.644854
```

Por tanto, n = 68. Efectivamente se observa que aplicando el teorema central del límite el tamaño muestral necesario (número de veces que ha de lanzarse la moneda para que la media muestral esté entre 0,4 y 0,6) ha disminuido sensiblemente.

Ejercicio 8-21. Si una población tiene $\sigma = 2$ y si $\overline{X}$ es la media de muestras de tamaño 100, hallar los límites entre los cuales estará situado $\overline{X}-\mu$ con probabilidad 0,90. Utilícese la desigualdad de Tchebichev y el teorema central del límite.

Aplicando el teorema central del límite tenemos:

$$P(-k<\overline{X}-u<k)=P\left(\frac{-k}{2/\sqrt{100}}<\frac{\overline{X}-u}{2/\sqrt{100}}<\frac{k}{2/\sqrt{100}}\right)=P(-5k<Y_{N(0,1)}<5k)=0{,}90$$

Tenemos:

$$F_{N(0,1)}(5k)-F_{N(0,1)}(-5k)=2F_{N(0,1)}(5k)-1=0{,}90\Leftrightarrow F_{N(0,1)}(5k)=0{,}95\Leftrightarrow 5k=1{,}645\Leftrightarrow k=0{,}329$$

Luego los límites pedidos para $\overline{X}-\mu$ son $\pm 0{,}329$

Aplicando el teorema de Tchebichev tenemos:

$$P(-k<\overline{X}-u<k)=P(|\overline{X}-u|<k)=1-\frac{\sigma^2/n}{k^2}=1-\frac{4/100}{k^2}=0{,}90\Rightarrow k=0{,}632$$

© Alfaomega - RC Libros

Luego los límites pedidos para $\overline{X}-\mu$ son $\pm 0{,}632$.

Se observa que los límites relativos a la desigualdad de Tchebichev son más amplios que los relativos al teorema central del límite.

Ejercicio 8-22. Supongamos que $\overline{X}_1$ y $\overline{X}_2$ son medias de dos muestras de tamaño n de una población con varianza σ^2. Hallar n de modo que la probabilidad de que las dos medias muestrales difieran en un valor superior a σ sea aproximadamente 0,01.

El problema se puede plantear en los siguientes términos:

$$P(|\overline{X}_1-\overline{X}_2|\geq\sigma)=0{,}01\Leftrightarrow P(|\overline{X}_1-\overline{X}_2|<\sigma)=0{,}99\Leftrightarrow P(-\sigma<|\overline{X}_1-\overline{X}_2|<\sigma)=0{,}99$$

Aplicando el teorema central del límite tenemos:

$$P\left(\frac{-\sigma}{\sqrt{2}\sigma/\sqrt{n}}<\frac{\overline{X}_1-\overline{X}_2-\overbrace{(\mu_1-\mu_2)}^{0}}{\sqrt{2}\sigma/\sqrt{n}}<\frac{\sigma}{\sqrt{2}\sigma/\sqrt{n}}\right)=P\left(-\sqrt{\frac{n}{2}}<Y_{N(0,1)}<\sqrt{\frac{n}{2}}\right)=0{,}99$$

No olvidemos que $\sigma^2(\overline{X}_1-\overline{X}_2)=\sigma^2(\overline{X}_1)+\sigma^2(\overline{X}_2)=\frac{\sigma^2}{n}+\frac{\sigma^2}{n}=2\frac{\sigma^2}{n}$

Realizando operaciones resulta:

$$F_{N(0,1)}\left(\sqrt{\frac{n}{2}}\right)-F_{N(0,1)}\left(-\sqrt{\frac{n}{2}}\right)=2F_{N(0,1)}\left(\sqrt{\frac{n}{2}}\right)-1=0{,}99\Leftrightarrow F_{N(0,1)}\left(\sqrt{\frac{n}{2}}\right)=0{,}995\Leftrightarrow\sqrt{\frac{n}{2}}=2{,}576$$

```
> qnorm(0.995,0,1)
[1] 2.575829
```

Tenemos entonces:

$$\frac{n}{2}=2{,}576^2\Leftrightarrow n=13$$

© Alfaomega - RC Libros

Ejercicio 8-23. Un investigador quiere estimar la media de una población usando una muestra suficientemente grande para que la probabilidad de que la media muestral no difiera de la media poblacional en más del 25% de la desviación típica sea 0,95. Hallar el tamaño de muestra necesario.

Aplicando el teorema central del límite tenemos:

$$P(|\overline{X}-u|<0{,}25\sigma)=0{,}95\Leftrightarrow P\left(\frac{-0{,}25\sigma}{\sigma/\sqrt{n}}<\frac{\overline{X}-u}{\sigma/\sqrt{n}}<\frac{0{,}25\sigma}{\sigma/\sqrt{n}}\right)=P\left(-0{,}25\sqrt{n}<Y_{N(0,1)}<0{,}25\sqrt{n}\right)=0{,}95$$

Realizando operaciones resulta:

$$2F_{N(0,1)}\left(0{,}25\sqrt{n}\right)-1=0{,}95\Leftrightarrow F_{N(0,1)}\left(0{,}25\sqrt{n}\right)=0{,}975\Leftrightarrow 0{,}25\sqrt{n}=1{,}96\Leftrightarrow n=62$$

```
> qnorm(0.975,0,1)
[1] 1.959964
```

Ejercicio 8-24. Se toma una muestra aleatoria de tamaño 1000 de una binomial de parámetros n=50 y p=0,5. Construir un histograma para la muestra obtenida.

La Figura 8-5 muestra el histograma pedido que se construye con la sintaxis siguiente:

```
> windows()
> muestra=rbinom(1000,50,0.5)
> hist(muestra)
```

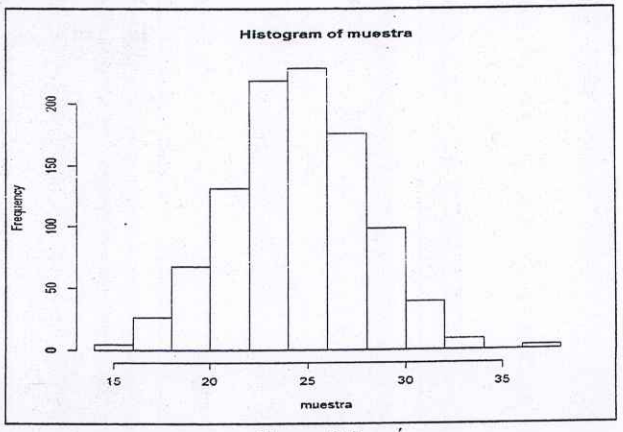


Figura 8-5

© Alfaomega - RC Libros

CAPÍTULO 9

INTERVALOS DE CONFIANZA Y TESTS DE HIPÓTESIS

ESTIMACIÓN POR INTERVALOS

El problema que presenta la estimación puntual de un parámetro reside en que no garantiza ni mide la precisión de la estimación. Solo la bondad de ajuste y el tamaño de la muestra pueden proporcionar una mayor o menor confianza en la estimación obtenida. Por esta razón es necesario dar, junto a la estimación, una medida del grado de confianza que nos merece. Esto se consigue mediante un intervalo de confianza que nos proporcione unos límites entre los cuales confiamos se encuentre el valor desconocido del parámetro. Esta confianza de inclusión se mide mediante un porcentaje, de tal manera que si este es el $\beta\%$, entonces, si obtenemos un gran número de intervalos como el anterior, tenemos la confianza de que el $\beta\%$ de ellos contendrán el valor exacto del parámetro desconocido.

© Alfaomega - RC Libros

Un intervalo de confianza para el parámetro θ de una población es de la forma $L_1 \leq \theta \leq L_2$, donde θ es el valor del parámetro desconocido, y L_1 y L_2 son funciones de los valores muestrales, que únicamente serán números cuando hayamos obtenido una muestra y sustituido sus valores en las funciones L_1 y L_2.

En el proceso de construcción de un intervalo de confianza se pueden distinguir dos etapas, que, para una mejor comprensión, podrían denominarse teórica y práctica. En la primera (etapa teórica), establecemos formalmente el intervalo aleatorio $[L_1, L_2]$, de tal manera que $P[L_1 \leq \theta \leq L_2] = \beta$. En esta etapa podemos hablar de la probabilidad de que el intervalo contenga el valor del parámetro θ. En la segunda etapa, tomamos una muestra de la población a la que pertenece el parámetro, y sustituida en L_1 y L_2, obtenemos dos números.

Una vez realizada tal sustitución, no es posible hablar de la probabilidad de que el valor desconocido del parámetro se encuentre comprendido entre ambos números, ya que, evidentemente, el parámetro lo estará o no. Sin embargo, como nosotros no sabemos si lo está, tenemos un *coeficiente de confianza* del β% de que lo esté. También se dice que el *nivel de confianza* es $\alpha = 1 - \beta$.

Expuesto el fundamento de la estimación por intervalos, para determinar un intervalo de confianza suele utilizarse el método general de Neyman, o el método de la cantidad pivotal, que pueden verse con detalle en los libros generales de estadística. Sin embargo, cuando la población es normal o sigue cualquier distribución aproximable por la normal, existen métodos puntuales de cálculo de intervalos de confianza para los parámetros poblacionales, de los cuales nos ocuparemos a continuación.

INTERVALOS DE CONFIANZA PARA PARÁMETROS DE DISTRIBUCIONES NORMALES

Si X es una variable aleatoria que sigue una distribución $N(\mu, \sigma)$, y $X_1, X_2, \ldots, X_n$ es una muestra aleatoria simple de tamaño n, podemos calcular intervalos de confianza para los parámetros μ y σ; dicho cálculo se basa en las distribuciones de determinados estadísticos, tal y como se indica en la siguiente tabla:

© Alfaomega - RC Libros

Parámetro	*Estadístico*	*Intervalo de confianza*
μ (σ conocida)	$\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$	$\bar{x} \pm Z_{\alpha/2}\,\sigma/\sqrt{n}$
μ (σ desconocida)	$\dfrac{\bar{x}-\mu}{S/\sqrt{n}} \to t_{n-1}$	$\bar{x} \pm t_{\alpha/2,n-1}\,S/\sqrt{n}$ $\bar{x} \pm Z_{\alpha/2}\,S/\sqrt{n}\,(sin > 30)$
σ^2	$\dfrac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}$	$\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$
σ		$\left(\sqrt{\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}, \sqrt{\dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}\right)$

Para cualquier a, los valores de Z_a, t_a, χ^2_a y F_a son tales que:

- Z_a es el valor de la abscisa de una $N(0,1)$ que deja a su derecha un área de valor a ($F(Z_a) = 1-a$), y F es la función de distribución de una $N(0,1)$).
- $T_{a,n}$ es el valor de la abscisa de una distribución T de Student con n grados de libertad que deja a su derecha un área de valor a ($F(t_{a,n}) = 1-a$, y F es la función de distribución de una T de Student con n grados de libertad).
- $\chi_{a,n}$ es el valor de la abscisa de una distribución chi-cuadrado con n grados de libertad que deja a su derecha un área de valor a ($F(\chi_{a,n}) = 1-a$, y F es la función de distribución de una chi-cuadrado con n grados de libertad).
- $F_{a,n1,n2}$ es el valor de la abscisa de una distribución F de Fisher Snedecor con n_1 y n_2 grados de libertad que deja a su derecha un área de valor a ($F(F_{a,n1,n2}) = 1-a$, y F es la función de distribución de una F de Fisher Snedecor con n_1 y n_2 grados de libertad).

INTERVALOS DE CONFIANZA PARA COMPARAR POBLACIONES NORMALES

Supongamos que tenemos dos poblaciones que siguen distribuciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, con σ_1 y σ_2 conocidas o desconocidas, y que a su vez pueden ser iguales o distintas. Siendo X_1 y X_2 dos muestras independientes de ambas, de tamaños n_1 y n_2, podemos resumir los intervalos de confianza para la diferencia de medias y cociente de varianzas según el siguiente cuadro:

© Alfaomega - RC Libros

Parámetro	*Estadístico*	*Intervalo de confianza*
$\mu_1-\mu_2$ (σ_1 y σ_2 dadas)	$\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \to N(0,1)$	$\bar{x}_1-\bar{x}_2 \pm Z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$
$\mu_1-\mu_2$ ($\sigma_1 \neq \sigma_2$ no dadas)	$\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}} \to t_f$ $f=\dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1+1}+\dfrac{(S_2^2/n_2)^2}{n_2+1}}-2$	$\bar{x}_1-\bar{x}_2 \pm t_{\alpha/2,f}\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}$ $\bar{x}_1-\bar{x}_2 \pm Z_{\alpha/2}\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}\,(n_1,n_2>30)$
$\mu_1-\mu_2$ ($\sigma_1=\sigma_2$ no dadas)	$\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \to t_{n_1+n_2-2}$ $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$\bar{x}_1-\bar{x}_2 \pm t_{\alpha/2,n_1+n_2-2}\,S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$
σ_1^2/σ_2^2	$\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \to F_{n_1-1,n_2-1}$	$\left(\dfrac{S_1^2/S_2^2}{F_{\alpha/2,n_1-1,n_2-1}},\dfrac{S_1^2/S_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\right)$
σ_1/σ_2		$\left(\sqrt{\dfrac{S_1^2/S_2^2}{F_{\alpha/2,n_1-1,n_2-1}}},\sqrt{\dfrac{S_1^2/S_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}}\right)$

RELACIÓN ENTRE INTERVALOS DE CONFIANZA Y CONTRASTES DE HIPÓTESIS

Los intervalos de confianza tienen la ventaja frente a los contrastes de hipótesis de que siempre nos dan una idea de la zona en la que se va a encontrar el verdadero valor del parámetro poblacional, mientras que en el caso de los tests, cuando se rechaza una hipótesis nula, no se conoce el valor del parámetro de la población en cuestión. Todo lo que se sabe es que es más verosímil que el valor del parámetro sea mayor o menor que un valor concreto.

© Alfaomega - RC Libros

En algunas ocasiones, en lugar de llevar a cabo un procedimiento de contraste de hipótesis sobre un parámetro, se puede construir un intervalo de confianza para el parámetro y extraer conclusiones del mismo. Supongamos que se desea contrastar H_0: $\mu = \mu_0$ frente a H_1: $\mu \neq \mu_0$. Si el nivel de significación es α, se pueden utilizar los datos de muestra para construir un intervalo de confianza para μ al $100(1-\alpha)\%$ de confianza. Si el valor de μ_0 no está incluido en el intervalo, hay que rechazar H_0; pero si el valor de μ_0 se encuentra en el intervalo calculado, no hay evidencia para rechazar H_0. Una ventaja de este procedimiento es que se pueden contrastar muchas hipótesis nulas de una vez, ya que se puede rechazar cualquier hipótesis nula de la forma H_0: $\mu = k$ para cualquier valor k que no pertenezca al intervalo.

Para un contraste de hipótesis unilateral el procedimiento es similar. Para contrastar H_0: $\mu = \mu_0$ frente a H_1: $\mu > \mu_0$, a un nivel de significación α, se construye un intervalo de confianza para μ al $100(1-2\alpha)\%$ de confianza. Si el valor de μ_0 es mayor que el extremo superior del intervalo, hay que rechazar H_0; si μ_0 es menor o igual que el extremo superior del intervalo, no hay evidencia para rechazar H_0. Por tanto, se puede rechazar cualquier hipótesis nula de la forma H_0: $\mu = k$, siendo k cualquier valor mayor que el extremo superior del intervalo de confianza.

Si se quiere contrastar H_0: $\mu = \mu_0$ frente a H_1: $\mu < \mu_0$, a un nivel de significación α, se puede razonar de una forma similar. Se construye un intervalo de confianza para μ al $100(1-2\alpha)\%$ de confianza. Si el valor de μ_0 es menor que el extremo inferior del intervalo, hay que rechazar H_0; si μ_0 es mayor o igual que el extremo inferior del intervalo, no hay evidencia para rechazar H_0.

Por tanto, se puede rechazar cualquier hipótesis nula de la forma H_0: $\mu = k$ siendo k cualquier valor menor que el extremo inferior del intervalo de confianza.

CONTRASTES DE HIPÓTESIS PARA POBLACIONES NORMALES

Pueden realizarse contrastes bilaterales o unilaterales sobre los valores de los parámetros de poblaciones normales basados en estadísticos de distribución conocida que permitirán calcular regiones críticas a un nivel α dado. Los dos cuadros siguientes resumen estos contrastes. El primero de ellos refleja contrastes para parámetros de dos poblaciones, generalmente utilizados para la comparación de ambas. El segundo presenta contrastes para el caso de parámetros de una única población normal.

© Alfaomega - RC Libros

Hipótesis	Estadístico	Regiones críticas
$H_o: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ (σ_1 y σ_2 dadas)	$Z_o = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\lvert Z_o \rvert > Z_{\alpha/2}$ $Z_o < -Z\alpha$ $Z_o > Z\alpha$
$H_o: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ ($\sigma_1 \neq \sigma_2$ no dadas)	$t_o = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ $f = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1+1} + \dfrac{(S_2^2/n_2)^2}{n_2+1}} - 2$	$\lvert t_o \rvert > t_{\alpha/2,f}$ $t_o < -t_{\alpha,f}$ $t_o > t_{\alpha,f}$
$H_o: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ ($\sigma_1 = \sigma_2$ no dadas)	$t_o = \dfrac{(\bar{x}_1 - \bar{x}_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ $g = n_1 + n_2 - 2$	$\lvert t_o \rvert > t_{\alpha/2,g}$ $t_o < -t_{\alpha,g}$ $t_o > t_{\alpha,g}$
$H_o: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$F_o = \dfrac{S_1^2}{S_2^2}$	$F_o > F_{\alpha/2,n_1-1,n_2-1}$ $F_o < F_{1-\alpha/2,n_1-1,n_2-1}$ $F_o < -F_{1-\alpha,n_1-1,n_2-1}$ $F_o > F_{\alpha,n_1-1,n_2-1}$

© Alfaomega - RC Libros

Hipótesis	Estadístico	Regiones críticas
$H_o: \mu = u_o$ $H_1: \mu \neq u_o$ $H_1: \mu < u_o$ $H_1: \mu > u_o$ (σ conocida)	$Z_o = \dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} \rightarrow N(0,1)$	$\lvert Z_o \rvert > Z_{\alpha/2}$ $Z_o < -Z_\alpha$ $Z_o > Z_\alpha$
$H_o: \mu = u_o$ $H_1: \mu \neq u_o$ $H_1: \mu < u_o$ $H_1: \mu > u_o$ (σ desconocida)	$t_o = \dfrac{\bar{x} - \mu_o}{S/\sqrt{n}} \rightarrow t_{n-1}$	$\lvert t_o \rvert > t_{\alpha/2,n-1}$ $t_o < -t_{\alpha/2,n-1}$ $t_o > t_{\alpha/2,n-1}$
$H_o: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$\dfrac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$	$\chi_o^2 > \chi^2_{\alpha/2,n-1}$ ó $\chi_o^2 < \chi^2_{1-\alpha/2,n-1}$ $\chi_o^2 < \chi^2_{1-\alpha,n-1}$ $\chi_o^2 > \chi^2_{\alpha,n-1}$

COMPARACIÓN DE DOS POBLACIONES NORMALES CON DATOS APAREADOS

Supongamos que tenemos dos poblaciones que siguen distribuciones normales, $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, siendo X e Y dos muestras de ambas poblaciones no necesariamente independientes, ambas de tamaño n. Se quiere contrastar la igualdad de medias de las dos poblaciones basándose en las muestras, para lo que se consideran los pares de valores (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) denominados datos apareados. Sea $d_i = x_i - y_i$ $i = 1,2,...,n$, de modo que los d_i puedan considerarse como una muestra aleatoria simple de la población normal de las diferencias.

Supongamos que estamos trabajando con *muestras grandes*, y que queremos realizar el *contraste bilateral* de igualdad de medias siguiente:

Hipótesis nula H_0: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$

Hipótesis alternativa H_1: $\mu_1 \neq \mu_2$

Bajo la hipótesis nula, el estadístico $Z = \dfrac{\bar{d}}{\hat{S}_d / \sqrt{n}}$ es una $N(0,1)$.

© Alfaomega - RC Libros

$$\bar{d} = \sum d_i/n \ \text{ y } \ \hat{S}_d^2 = \sum (d_i - \bar{d})^2 / (n-1)$$

La región de aceptación del contraste viene definida mediante el intervalo dado por $(-Z_{\alpha/2}, Z_{\alpha/2})$, aceptándose la hipótesis nula si se cumple:

$$\frac{|\bar{d}|}{\hat{S}_d / \sqrt{n}} \leq Z_{\alpha/2}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $\mu_1 \leq \mu_2$
Hipótesis alternativa H_1: $\mu_1 > \mu_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-\infty, Z_\alpha)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{\hat{S}_d / \sqrt{n}} \leq Z_\alpha$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $\mu_1 \geq \mu_2$
Hipótesis alternativa H_1: $\mu_1 < \mu_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-Z_\alpha, \infty)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{\hat{S}_d / \sqrt{n}} \geq -Z_\alpha$$

Para muestras grandes, un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ viene dado por $\bar{d} \pm Z_{\alpha/2} \hat{S}_d / \sqrt{n}$.

© Alfaomega - RC Libros

Supongamos que estamos trabajando con *muestras pequeñas*, y que queremos realizar el *contraste bilateral* siguiente:

Hipótesis nula H_0: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$
Hipótesis alternativa H_1: $\mu_1 \neq \mu_2$

Bajo la hipótesis nula, el estadístico $T = \dfrac{\bar{d}}{\hat{S}_d / \sqrt{n}}$

es una T de Student con $n-1$ grados de libertad.

La región de aceptación del contraste viene definida mediante el intervalo dado por $(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1})$, aceptándose la hipótesis nula si se cumple:

$$\frac{|\bar{d}|}{\hat{S}_d / \sqrt{n}} \leq t_{\alpha/2,n-1}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $\mu_1 \leq \mu_2$
Hipótesis alternativa H_1: $\mu_1 > \mu_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-\infty, t_{\alpha,n-1})$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{\hat{S}_d / \sqrt{n}} \leq t_{\alpha,n-1}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $\mu_1 \geq \mu_2$
Hipótesis alternativa H_1: $\mu_1 < \mu_2$

© Alfaomega - RC Libros

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-t_{\alpha,n-1}, \infty)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{\hat{S}_d / \sqrt{n}} \geq -t_{\alpha,n-1}$$

Para muestras pequeñas, un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ viene dado por $\bar{d} \pm t_{\alpha/2,n-1} \hat{S}_d / \sqrt{n}$.

EL LENGUAJE R Y LOS INTERVALOS DE CONFIANZA Y CONTRASTES DE HIPÓTESIS

El comando *t.test* de R permite trabajar con intervalos de confianza y test de hipótesis para la media y la diferencia de medias, incluido el caso de datos apareados. En todo caso las medias y varianzas poblacionales no podrán suponerse nunca conocidas, sino que se estimarán a partir de los datos de las variables que han de ser también especificadas.

La sintaxis general por defecto del comando *t.test* se presenta a continuación:

t.test(x,y=NULL, alternative="two.sided", mu=0, paired= "FALSE", var.equal = FALSE, conf.level=0.95)

En el argumento *x* introducimos la muestra sobre la que hacemos la inferencia. Si no se especifica la variable y se hace la inferencia solo para una muestra. Si se especifican las variables *x* e *y*, la inferencia se hace para comparación de muestras independientes o apareadas.

El argumento *alternative* permite las opciones *two.sided* (por defecto), *greater* y *less*. La primera de las opciones realiza el contraste bilateral y calcula el intervalo de confianza asociado (bien sea para la media o para la diferencia de medias). La segunda de las opciones considera la hipótesis nula *menor o igual* frente a la alternativa de *mayor*. La tercera opción considera la hipótesis nula *mayor o igual* contra la alternativa de *menor*.

© Alfaomega - RC Libros

El argumento *mu* indica el valor de la hipótesis nula a considerar para la media.

El argumento *paired* indica si consideramos datos apareados o no. Por defecto se usan datos independientes (*paired*=FALSE) y no es necesario especificar el argumento. Para el caso de datos apareados hay que especificar *paired*=TRUE.

El argumento *var.equal* permite especificar varianzas iguales (*var.equal*=TRUE) o varianzas desiguales (*var.equal*=FALSE, que es la opción por defecto y no es necesario especificarla).

El argumento *conf.level* permite especificar el coeficiente de confianza (0,95 por defecto, lo que es equivalente a un nivel de 0,05 para los tests).

Para realizar un contraste de igualdad de medias es necesario conocer previamente si las varianzas son iguales o no. Para contrastar la igualdad de varianzas se utiliza el comando *var.test*, que presenta la siguiente sintaxis general:

var.test(x, y, ratio, alternative="two.sided", conf.level=0.95)

Los argumentos *x* e *y* son las dos muestras a comparar.

En cuanto al argumento *ratio*, para contrastar la igualdad de varianzas se utiliza ratio=1.

La hipótesis *alternative* funciona como en el caso del comando *t.test*.

La hipótesis *conf.level* permite introducir el coeficiente de confianza.

En el caso de no normalidad para muestras grandes, si queremos contrastar un valor de la media de la muestra *x* o hallar el intervalo de confianza para la misma al 95%, se utiliza el comado:

t.test(x)

En el caso de los contrastes e intervalos de confianza para la diferencia de medias en muestras grandes de poblaciones no normales, se utiliza el comando *t.test* habitualmente, lo mismo que en el caso de poblaciones normales.

© Alfaomega - RC Libros

En el caso de poblaciones binomiales se utiliza el comando *prop.test*, cuya sintaxis general es la siguiente:

prop.test(x, n, ratio, alternative="two.sided", conf.level=0.95, correct=TRUE)

El argumento *x* es el vector de éxitos, el argumento *n* es el vector de número de pruebas realizadas, los argumentos *alternative* y *conf.level* tienen la finalidad ya estudiada anteriormente y el argumento *correct* se utiliza para indicar una corrección de Yates (su valor por defecto es TRUE y si no queremos utilizarla habrá que especificar *correct*=FALSE).

Como primer ejemplo consideramos el peso de varias piezas de motores de automóviles a una temperatura normal (variable X) y a otra temperatura elevada (variable Y) para contrastar si hay diferencias entre las medias de los pesos de las mismas piezas a diferentes temperaturas. Los datos son los siguientes:

X 169 168 165 177 179 168 169 167 181 163
Y 168 165 164 175 176 166 167 166 179 161

Como medimos los pesos de las mismas piezas a dos temperaturas diferentes, estamos ante datos apareados. El contraste de igualdad de medias se realizará mediante la siguente sintaxis:

```
> x=c(169,168,165,177,179,168,169,167,181,163)
> y=c(168,165,164,175,176,166,167,166,179,161)
> t.test(x,y,paired=TRUE)

        Paired t-test

data:  x and y
t = 8.1429, df = 9, p-value = 1.921e-05
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 1.372163 2.427837
sample estimates:
mean of the differences
                    1.9
```


© Alfaomega - RC Libros

Como el *p*-valor del contraste es muy pequeño, se rechaza la hipótesis nula de igualdad de medias. Además, el intervalo de confianza para la diferencia de medias es [1,372163 2,427837] que no contiene el valor cero, luego no puede aceptarse el valor cero para la diferencia de medias, es decir, no puede aceptarse que las medias sean iguales.

El contraste de igualdad de medias lo hemos realizado suponiendo varianzas distintas ya que si no se especifica el argumento *var.equal* en el comando *t.test*, se supone por defecto que las varianzas son desiguales (*var.equal*=FALSE).

Debemos entonces contastar la hipótesis de que las varianzas de *X* e *Y* son iguales. Para ello utilizamos la siguiente sintaxis.

```
> var.test(x,y,1)

        F test to compare two variances

data:  x and y
F = 1.0907, num df = 9, denom df = 9, p-value = 0.8992
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
 0.2709085 4.3910555
sample estimates:
ratio of variances
          1.090676
```

El *p*-valor muy alto de la salida indica que se acepta la hipótesis nula de igualdad de varianzas. Este hecho lo corrobora el intervalo de confianza para el cociente de varianzas [0,2709085 4,3910555] que contiene el valor 1.

Por lo tanto el contraste de la igualdad de medias debemos realizarlo bajo el supuesto de varianzas iguales mediante la siguiente sintaxis:

```
> t.test(x,y,paired=TRUE,var.equal=TRUE)

        Paired t-test

data:  x and y
t = 8.1429, df = 9, p-value = 1.921e-05
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
```

© Alfaomega - RC Libros

```
 1.372163 2.427837
sample estimates:
mean of the differences
                    1.9
```

Se obtiene un *p*-valor muy pequeño para el contraste que permite rechazar la hipótesis nula de igualdad de medias. Además, este resultado concuerda con el hecho de que el intervalo de confianza para la diferencia de medias no contiene el valor cero.

Como segundo ejemplo consideramos la variable Z que contiene los tiempos en minutos de ejecución de una tarea por 6 empledos de una empresa.

Z → 18 19 21 21 18 20

Se trata de contrastar si se puede aceptar la hipótesis de que el tiempo medio de ejecución de la tarea en la empresa es 20 minutos.

Utilizaremos la siguiente sintaxis:

```
> z=c(18,19,21,21,18,20)
> t.test(z,mu=20)

        One Sample t-test

data:  z
t = -0.8885, df = 5, p-value = 0.415
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 18.05345 20.94655
sample estimates:
mean of x
     19.5
```

Como el *p*-valor del contraste es alto, aceptamos la hipótesis nula de que el tiempo medio de ejecución de la tarea en la empresa es 20 minutos.

Como tercer ejemplo consideramos un medicamento que se administra en una población de 250 enfermos realizando efecto a 183 de ellos. El mismo medicamento se administra a otra población de 200 enfermos teniendo efecto en 90. Se trata de contrastar si la proporción de enfermos con efecto del medicamento es la misma en la dos poblaciones.

© Alfaomega - RC Libros

En este caso se trata de muestras independientes. Para realizar el contraste de la igualdad de proporciones utilizamos la siguiente sintaxis:

```
> x=c(183,90)
> n=c(250,200)
> prop.test(x,n,correct=F)

        2-sample test for equality of proportions without
continuity correction

data:  x out of n
X-squared = 37.0292, df = 1, p-value = 1.164e-09
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1938625 0.3701375
sample estimates:
prop 1 prop 2
 0.732  0.450
```

Como el *p*-valor del contraste es muy pequeño se rechaza la hipótesis nula de igualdad de proporciones. Además, el intervalo de confianza para la diferencia de proporciones es [0,1938625 0,3701375] y el cero no está en este intervalo, luego no se puede admitir diferencia de proporciones igual a cero (proporciones iguales).

CONTRASTES NO PARAMÉTRICOS

En los contrastes de hipótesis realizados hasta ahora hemos supuesto que la población de la cual tomamos la muestra pertenecía a una determinada familia de distribuciones y, por tanto, la estimación o contraste nos permitía especificar totalmente esa población. En los siguientes epígrafes vamos a analizar algunos procedimientos que no exigen ningún supuesto, o muy pocos, acerca de la familia de distribuciones a que pertenece la población, procedimientos que, además, soportan observaciones en donde las mediciones se realizan en forma cualitativa (clasificando caracteres cualitativos) o bien se refieren a alguna característica ordenable. A tales procedimientos se les denomina *contrastes no paramétricos*.

CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN: CONTRASTES DEL COEFICIENTE DE CORRELACIÓN Y MEDIDAS DE ASOCIACIÓN

En el capítulo 3 hemos utilizado coeficientes de correlación para estudiar el grado de asociación para variables numéricas. Además, cuando intentábamos

© Alfaomega - RC Libros

contrastar la independencia o el grado de asociación para variables carácter, nos encontrábamos con el problema de la no cuantificación de los atributos o caracteres y veíamos que cuando los caracteres estudiados pueden ordenarse de acuerdo con una cierta escala, se puede llegar a unos coeficientes de correlación que midan el grado de asociación entre ellos de manera parecida a como lo hicimos para medir la asociación entre variables numéricas. Estos coeficientes están basados en los rangos u órdenes de las observaciones y se denominan coeficientes de correlación por rangos. Considerábamos especialmente ***el coeficiente de correlación por rangos r de Sperman***. Cuando r=1 la concordancia entre los rangos es perfecta y el grado de asociación de las variables cualitativas es óptimo. Cuando r = -1 la discordancia entre rangos es muy alta y el grado de asociación entre las variables cualitativas es ínfimo. Si no se rechaza la hipótesis nula de independencia entre los factores, se dará por finalizado el análisis, pero si se rechaza tal hipótesis y, por consiguiente, se acepta la existencia de asociación significativa entre ellos, aún se pueden plantear cuestiones como: ¿cuál es la intensidad de la asociación entre las variables? o ¿cuál es la dirección de la asociación detectada?

La respuesta a tales preguntas ya constituyó objeto de estudio anteriormente en el capítulo 3 mediante el uso de los distintos coeficientes o medidas de asociación (cuadrado medio de la contingencia, coeficiente de contingencia C de K. Pearson, coeficiente T de Tschuprow, coeficiente V de Cramer, coeficientes λ de Goodman y Kruskall, etc.).

Contraste del coeficiente de correlación

Bajo la hipótesis nula de que el coeficiente de correlación poblacional sea nulo ($\rho = 0$), el estadístico:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

se distribuye según una t de Student con n-2 grados de libertad, siendo r el coeficiente de coeficiente de correlación muestral y n el tamaño de la muestra.

Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de correlación poblacional es cero*. Para ello, se halla el valor k tal que $P(T_{n-2} \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico T para los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de incorrelación al nivel fijado α. En caso contrario se acepta la incorrelación.

© Alfaomega - RC Libros

Bajo la hipótesis nula de que el coeficiente de correlación poblacional sea $\rho_0 \neq 0$, se tiene:

$$Z = \frac{1}{2} Ln \frac{1+r}{1-r} \rightarrow N\left(\frac{1}{2} Ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)}, \sqrt{\frac{1}{n-3}}\right)$$

Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de correlación poblacional es un valor dado* ($\rho = \rho_0$). Para ello, se halla el valor k tal que $P(Z \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico Z para los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de que el coeficiente de correlación poblacional ρ vale ρ_0. En caso contrario se acepta $\rho = \rho_0$.

Contraste del coeficiente de regresión

Bajo la hipótesis nula de que el coeficiente de regresión poblacional de Y sobre X sea β_{12}, el estadístico:

$$T = \frac{S_2\sqrt{n-2}}{S_1\sqrt{1-r^2}}(b_{12} - \beta_{12})$$

se distribuye según una t de Student con n-2 grados de libertad, siendo b_{12} el coeficiente de regresión muestral de Y sobre X, r el coeficiente de correlación muestral de X e Y, S_1 y S_2 las cuasivarianzas muestrales de X e Y, y n el tamaño de la muestra. Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de regresión poblacional de Y sobre X es* β_{12}.

Para ello, se halla el valor k tal que $P(T_{n-2} \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico T para los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de que el coeficiente de regresión poblacional de Y sobre X sea β_{12} al nivel fijado α. En caso contrario se acepta β_{12} como coeficiente de regresión de Y sobre X.

CONTRASTE CHI-CUADRADO DE INDEPENDENCIA

Se pretende contrastar, con un nivel de significación A, la hipótesis nula de independencia entre dos variables que pueden ser dos atributos A y B, que presentan h y k niveles exhaustivos y mutuamente excluyentes, respectivamente. Para ello se toma una muestra aleatoria simple de tamaño N de una población bidimensional, y los N valores muestrales se clasifican en una tabla de doble entrada, denominada Tabla de

© Alfaomega - RC Libros

Contingencia, en función de los niveles de *A* y de *B* que posean. Así, la información muestral obtenida se clasifica como indica la tabla siguiente:

$A, B \rightarrow$ ↓	B_1	B_2	$\cdots$	B_j	$\cdots$	B_k	$n_{i\cdot}$
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_h	n_{h1}	n_{h2}	$\cdots$	n_{hj}	$\cdots$	n_{hk}	$n_{h.}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	$n_{..}$

Tenemos que la observación simultánea de dos atributos da lugar a una tabla de doble entrada, en donde n_{ij} indica el número de objetos o individuos que poseen conjuntamente las modalidades indicadas en la fila *i*-ésima y en la columna *j*-ésima de la tabla de contingencia. Sabemos que dos atributos *A* y *B* son independientes cuando entre ellos no existe ningún tipo de influencia mutua. Si dos atributos, *A y B*, son independientes estadísticamente, la frecuencia relativa conjunta será igual al producto de las frecuencias marginales respectivas. Para que *A* y *B* sean independientes habrá de cumplirse que $n_{ij} = (n_{i\cdot}n_{\cdot j})/N$ para todo *i,j*. En la práctica basta con que la relación se verifique para $(h-1)(k-1)$ valores de n_{ij}, ya que entonces se verificará para todos los restantes.

Si designamos por n_{ij} la frecuencia conjunta correspondiente a las modalidades A_i del atributo *A* y B_j de *B*, y por e_{ij} la frecuencia teórica que correspondería en el caso de que ambos atributos fuesen independientes, esto es, $e_{ij} = (n_{i\cdot}n_{\cdot j})/N$, $i=1,\ldots,h$, $j=1,\ldots,k$, siendo *N* el total de elementos que se estudian, definimos el *coeficiente de contingencia* χ^2 como sigue:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

Este coeficiente también se denomina en la literatura estadística *cuadrado de la contingencia*, y puede expresarse de forma más sencilla para el cálculo como sigue:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{e_{ij}} - N$$

El coeficiente de contingencia χ^2 se utiliza para realizar un contraste formal para la hipótesis nula de independencia de los atributos *A* y *B* cuya información muestral se recoge en la tabla de contingencia dada. La hipótesis alternativa es la existencia de asociación entre los atributos *A* y *B*. El contraste se basa en que, bajo la hipótesis nula de independencia de los atributos *A* y *B*, el estadístico χ^2 se distribuye según una chi-cuadrado con $(h-1)(k-1)$ grados de libertad.

Para realizar el contraste se halla el valor *k* tal que $P(\chi^2_{(h-1)(k-1)} \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico χ^2 para los datos dados de la tabla de contingencia es mayor que *k* se rechaza la hipótesis nula de independencia de los atributos *A* y *B* al nivel fijado α. En caso contrario se acepta la independencia.

Cuando el tamaño muestral es pequeño (*N* menor que 150) se utiliza el test exacto de Fisher para contrastar la independencia de atributos. En este caso suele introducirse una corrección por continuidad en el estadístico de la chi-cuadrado, tomando en su lugar para el contraste de independencia el estadístico corregido de Yates, cuya expresión es la siguiente:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(|n_{ij}-e_{ij}|-\frac{1}{2})^2}{e_{ij}}$$

Como concepto contrario al de independencia tenemos el de *asociación.* Se dice que *A y B* están asociados cuando aparecen juntos en mayor número de casos que el que cabría esperar si fuesen independientes. Según que esa tendencia a coincidir o no coincidir esté más o menos marcada, tendremos distintos grados de asociación. Para medirlos se han ideado diversos procedimientos, denominados *coeficientes de asociación*, entre los que destacan los ya estudiados en el capítulo 3.

CONTRASTES MÚLTIPLES DE HOMOGENEIDAD: TEST DE LA $\chi 2$

Sean $\Omega_1, \Omega_2, \ldots, \Omega_r$ *r* poblaciones, y sean $B_1, B_2, \ldots, B_s$ *s* sucesos excluyentes, de modo que $\Omega_i = B_1+B_2+\ldots+B_s$ para i = 1, 2, ..., *r*.

Queremos contrastar si las probabilidades de los sucesos B_j son homogéneas a lo largo de las *r* poblaciones, es decir, establecemos la hipótesis nula H_0 siguiente:

$$Ho: P_1(B_j) = P_2(B_j) = \ldots = P_r(B_j) \quad j = 1, 2, \ldots, s$$

Obtenidas muestras independientes de tamaños $n_1, n_2, \ldots, n_r$ de las r poblaciones, si n_{ij} es la frecuencia de B_j en Ω_i, el contraste de hipótesis nula H_o se resuelve aplicando el test de la chi-cuadrado a la tabla de contingencia $[n_{ij}]$. El estadístico a utilizar es:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

que bajo la hipótesis nula sigue una distribución chi-cuadrado con $(r-1)(s-1)$ grados de libertad.

La diferencia de este contraste con el contraste de independencia de la chi-cuadrado que veremos luego, es que aquí los números $n_1, n_2, \ldots, n_r$ no son aleatorios, sino tamaños muestrales fijos, mientras que en el contraste de independencia las frecuencias $f_{1.}, f_{2.}, \ldots, f_{r.}$ son frecuencias con distribución conjunta multinomial.

CONTRASTES DE LA BONDAD DE AJUSTE: TEST DE LA CHI-CUADRADO

Este tipo de tests tratan de contrastar si de los datos obtenidos en una muestra se puede deducir o no que proceden de una población con una distribución determinada. Entre ellos destaca el contraste chi-cuadrado de la bondad de ajuste que se describe a continuación.

Sea una muestra aleatoria simple de tamaño N de una cierta variable. El contraste de la chi-cuadrado de la bondad de ajuste se lleva a cabo para saber si hay razones suficientes para aceptar que esta variable tiene una distribución de probabilidad dada. Por tanto, la hipótesis nula H_o hará referencia a que las observaciones muestrales constituyen un conjunto de N valores procedentes de una variable aleatoria con una distribución de probabilidad dada por $P(S)$, y la alternativa de que no sucede esto. Pasemos a estudiar este contraste distinguiendo dos casos:

© Alfaomega - RC Libros

a) Parámetros poblacionales conocidos

La distribución de probabilidad $P(S)$ está completamente especificada, de modo que en su expresión no aparecen parámetros desconocidos. Si H_o es cierta, la distribución de la muestra, que es la distribución uniforme discreta que se obtiene al asignar una probabilidad $P_i = 1/N$ a cada uno de los N valores observados, puede considerarse como una imagen estadística de la distribución poblacional $P(S)$.

Sin embargo, debido a fluctuaciones aleatorias del muestreo, lo razonable será que las distribuciones poblacional y muestral no coincidan, pero para muestras grandes cabe esperar que la distribución de la muestra constituya una aproximación a la distribución poblacional. En este sentido, parece natural introducir alguna medida de la desviación entre ambas distribuciones y basar nuestro contraste en las propiedades de la distribución en el muestreo de esa medida.

Supongamos que el espacio de la variable aleatoria se divide en r intervalos $S_1, S_2, \ldots, S_r$ disjuntos, y sea la probabilidad de que un valor pertenezca al intervalo i-ésimo $p_i = P(S_i)$, donde tales intervalos pueden ser los r grupos o clases en que se han dispuesto los valores muestrales a efectos de tabulación. Sean $n_1, n_2, \ldots, n_r$ las correspondientes frecuencias absolutas en los r grupos muestrales, de tal manera que n_i valores muestrales pertenecen a la clase S_i, siendo $\sum n_i = N$ con $i = 1, \ldots, r$. Ahora se elabora una medida de la desviación entre la distribución de la muestra y la distribución supuesta para la población bajo la hipótesis nula, cuya expresión es:

$$\chi^2 = \sum \frac{(n_i - Np_i)^2}{Np_i}$$

donde n_i son las frecuencias observadas en la muestra, y Np_i la estimación de las frecuencias esperadas si la distribución poblacional fuese la que indica H_o. El estadístico χ^2 tiene, en el límite ($N \to \infty$), una distribución chi-cuadrado con $r - 1$ grados de libertad.

Determinada la distribución en el muestreo para la medida de desviación χ^2, construiremos el contraste para H_0 de la siguiente forma: fijado el nivel de significación A, buscamos en las tablas de la chi-cuadrado un valor C tal que $P(\chi^2_{(r-1,\,A)} > C) = A$, y si el valor del estadístico χ^2 es mayor que C, entonces rechazamos la hipótesis H_0 de que la muestra proviene de la población indicada. Si el valor del estadístico χ^2 es menor que C, entonces aceptamos H_0 y aseguramos que la muestra proviene de la población indicada.

© Alfaomega - RC Libros

b) Parámetros poblacionales desconocidos

Supongamos que según la hipótesis nula H_0 la muestra procede de una población en donde hay que estimar previamente una serie de k parámetros por ser desconocidos. Fisher demostró que si previamente se estiman los k parámetros mediante la información muestral, el límite de la distribución del estadístico χ^2 es una chi-cuadrado con $r - k - 1$ grados de libertad, y el contraste se realiza como en el caso anterior. La única diferencia son los grados de libertad del estadístico χ^2.

Una vez expuesto el contraste χ^2 de bondad del ajuste, tanto para el caso en que no se estiman parámetros como para cuando sí se hace, hemos de señalar que, para que no se produzcan perturbaciones, los intervalos S_i deberían ser tales que los valores de $P(S_i)$, para todo i, fueran aproximadamente iguales. Además, cuando el número de intervalos en que se divide el espacio de la variable aleatoria sea 2 el test χ^2 no debe utilizarse si existe alguna frecuencia esperada inferior a 5.

Si el número de intervalos es mayor que 2 tampoco deberá usarse el test χ^2 si más del 20% de las frecuencias esperadas son menores que 5 o alguna es inferior a 1. No obstante, las frecuencias esperadas a veces pueden incrementarse realizando uniones de intervalos adyacentes, uniones que, por supuesto, tengan sentido en la realidad.

CONTRASTE DE KOLMOGOROV-SMIRNOV DE LA BONDAD DE AJUSTE

El contraste de Kolmogorov-Smirnov es un test para bondad de ajuste alternativo al de la chi-cuadrado. Al igual que en el contraste de la chi-cuadrado de bondad del ajuste, consideramos que la masa total de probabilidad discreta está repartida uniformemenie entre los N valores muestrales de forma que, ordenados los valores muestrales de menor a mayor, la función de distribución empírica de la muestra es $F_n(x) = N_i/N$.

El contraste de Kolmogorov-Smirnov se aplica solo a variables continuas y trata de medir el ajuste entre la función de distribución empírica de una muestra y la función de distribución teórica. Se trata por tanto de un contraste de ajuste de la distribución de una muestra dada a una distribución continua determinada.

La función de distribución empírica de una muestra $x_1\ x_2 \ldots x_n$, se define como:

© Alfaomega - RC Libros

$$F_n(x) = \frac{\text{nº de valores del conjunto } \{x_1, x_2, .., x_n\} \text{ que son} \leq x}{n}$$

Para contrastar la hipótesis de que la muestra se ajusta a una distribución teórica F(x), se calcula el estadístico:

$$D_n = Máx\ |F_n(x) - F(x)|$$

cuya distribución es conocida y está tabulada. Si la distancia calculada D_n es mayor que la encontrada en las tablas, para un nivel A, rechazamos la distribución F(x) para la muestra.

Para n y A dados, hallamos $D(A,n)$ tal que $P(D_n > D(A,n)) = A$. La región crítica del test será $D_n > D(A,n)$.

Este contraste tiene la ventaja de que no requiere agrupar los datos y el inconveniente de que si calculamos $F(x)$ estimando parámetros de la población, mediante la muestra, la distribución de D_n es solo aproximada.

Normalmente se usa el estadístico $\sqrt{n}\ D_n$ en vez del D_n, lo que nos permite contrastar tablas para tamaños muestrales muy grandes.

Tendremos en cuenta las siguientes consideraciones:

- El test de Kolmogorov-Smirnov es de más fácil aplicación que el test de la χ^2.
- El test de Kolmogorov-Smirnov no se ve afectado por reagrupaciones de las observaciones, mientras que en el test de la χ^2, al disminuir los grupos, se pierde información, así como grados de libertad.
- El test de Kolmogorov-Smirnov es aplicable a pequeñas muestras, mientras que el test de la χ^2 está diseñado para grandes muestras.
- La potencia del test de Kolmogorov-Smirnov es mayor que la del de la χ^2, si bien tienden a igualarse cuando el tamaño de la muestra crece.
- El test de la χ^2 puede fácilmente ser modificado cuando hay parámetros desconocidos mientras que el test de Kolmogorov-Smirnov no tiene tal flexibilidad.

© Alfaomega - RC Libros

- El test de la χ^2 es aplicable cuando la población es discreta o continua y el test de Kolmogorov-Smirnov requiere la continuidad de $F(x)$.

Se puede utilizar para estos contrastes (como para todos) el criterio del *p*-valor, rechazando la hipótesis nula al nivel α cuando el *p*-valor es menor que α, y aceptándola en caso contrario.

CONTRASTES DE ALEATORIEDAD: TEST DE RACHAS Y TEST DE DANIEL

Test de rachas de Wald-Wolfowitz y test del cuadrado medio de diferencias sucesivas

Se trata de tests para contrastar la aleatoriedad de una muestra basándose en el número de rachas observadas en la misma, siendo una racha un conjunto de números iguales o de la misma naturaleza.

Hipótesis	*Estadístico*	*Región crítica*
H_0 : *Muestra aleatoria*	$R = Total\ de\ rachas$	$R \le k'_{\alpha/2}\ con\ P(R \le k'_{\alpha/2} / H_0) \le \alpha/2$
H_1 : *Muestra no aleatoria*		$R \ge k'_{\alpha/2}\ con\ P(R \ge k'_{\alpha/2} / H_0) \le \alpha/2$

Para tamaños de muestra grandes (n_1,n_2>10), bajo la hipótesis nula se verifica:

$$R \to N\left(\frac{2n_1n_2}{n}+1, \sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}}\right) \Rightarrow Z = \frac{R-\frac{2n_1n_2}{n}}{\sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}}} \to N(0,1)$$

Las región crítica para el contraste bilateral es:

$$|Z| \ge Z_{\alpha/2}$$

También se puede utilizar el criterio del *p*-valor, rechazando la hipótesis nula al nivel α cuando el *p*-valor es menor que α, y aceptándola en caso contrario.

© Alfaomega - RC Libros

Cuando existen variaciones de naturaleza cíclica en las observaciones, el test de aleatoriedad de Wald-Wolfowitz no es eficiente y será necesario utilizar el test del cuadrado medio de diferencias sucesivas para detectar la aleatoriedad.

Hipótesis	*Estadístico*	*Región crítica*
H_0 : *Muestra aleatoria* H_1 : *Muestra no aleatoria*	$R = \dfrac{\sum_{i=1}^{n-1}(X_{i+1}-X_i)^2}{2\sum_{i=1}^{n}(X_i-\overline{X})^2}$	$R \le k'_\alpha\ con\ P(R \le k'_\alpha / H_0) \le \alpha$

Si la muestra proviene de una población normal y R es grande (R>20) se tiene:

$$R \to N\left(1, \sqrt{\frac{1}{n+1}\left(1-\frac{1}{n-1}\right)}\right) \Rightarrow Z = \frac{R-1}{\sqrt{\frac{1}{n+1}\left(1-\frac{1}{n-1}\right)}} \to N(0,1)$$

Las región crítica para el contraste es:

$$Z \le Z_\alpha$$

También se puede utilizar el criterio del *p*-valor, rechazando la hipótesis nula al nivel α cuando el *p*-valor es menor que α, y aceptándola en caso contrario.

Test de Daniel

Se trata de un test para contrastar la hipótesis nula de aleatoriedad de una muestra basándose en el coeficiente de correlación por rangos de Spearman.

Dados los T valores de la serie Y_t, calculamos $d_t = Rango(y_t)-t$ (t = 1,2,...,T), bajo la hipótesis nula de aleatoriedad, el coeficiente de correlación de Spearman cumple:

$$\tau_s = 1-\frac{6\sum_{t=1}^{T} d_t^2}{T(T^2-1)} \to N\left(0, \frac{1}{T-1}\right) \Rightarrow Z = \sqrt{T-1}\,\tau_s \to N(0,1)$$

La región crítica para el contraste bilateral a un nivel de significación α es:

$$|Z| \ge Z_{\alpha/2}$$

© Alfaomega - RC Libros

También se puede utilizar el criterio del *p*-valor, rechazando la hipótesis nula al nivel α cuando el *p*-valor es menor que α, y aceptándola en caso contrario.

EL LENGUAJE R Y LOS CONTRASTES NO PARAMÉTRICOS

Los comandos esenciales de R para el trabajo con distribuciones continuas se presentan a continuación:

t.test(x,y=NULL,alternative="two.sided", mu=0, paired= "FALSE", var.equal = FALSE, conf.level=0.95)

cor.test (x,y)	*Contraste del coeficiente de correlación de Pearson para las variables x e y*
cor.test (x,y,method="spearman")	*Contraste del coeficiente de correlación de Spearman para las variables x e y*
chisq.test(x,correct=TRUE,p)	*Realiza el **test de la bondad de ajuste a una distribución**. El argumento x es el vector de frecuencias absolutas, correct=TRUE permite utilizar la corrección de Yates y p es el vector de probabilidades teóricas que comparamos con las observadas*
chisq.test(x,correct=TRUE)	*Realiza el **test de homogeneidad de varias muestras**. El argumento x es la tabla de doble entrada para los datos de las muestras y correct=TRUE permite utilizar la corrección de Yates*
chisq.test(x,correct=TRUE)	*Realiza el **test de independencia de caracteres**. El argumento x es la tabla de contingencia de los dos caracteres y correct=TRUE permite utilizar la corrección de Yates*
ks.test(x,y,..., alternative="two.sided")	*Realiza el **test de la bondad de ajuste a una distribución de Kolmogorov-Smirnov**. El argumento x es el vector de observaciones, el argumento y es el nombre de la función de distribución que queremos ajustar a las observaciones, los puntos suspensivos son los parámetros de la distribución de ajuste y el argumento alternative es el mismo de todos los contrastes*

© Alfaomega - RC Libros

ks.test(x,y, alternative="two.sided")	*Realiza el **test de Kolmogorov-Smirnov para la homogeneidad de dos muestras**. Los argumentos x e y son las dos muestras y el argumento alternative es el mismo de todos los contrastes*
wilcox.test(x,y, alternative="two.sided", mu=0, exact=TRUE, correct=TRUE)	*Realiza el **test de Wilcoxon-Mann-Whitney para la igualdad de medianas de las dos poblaciones (homegeneidad de poblaciones)** cuyas muestras son x e y. El argumento mu permite introducir el valor de la diferencia de medianas a contrastar, exact=TRUE permite utilizar la distribución exacta del estadístico U y no su distribución límite y correct=TRUE permite utilizar la corrección por continuidad. El argumento alternative tiene la función de siempre*
wilcox.test(x, alternative="two.sided", mu=0, exact=TRUE, correct=TRUE)	*Realiza el **test de Wilcoxon-Mann-Whitney para la media de una población** cuya muestra es x. El argumento mu permite introducir el valor de la media a contrastar, exact=TRUE permite utilizar la distribución exacta del estadístico U y no su distribución límite y correct=TRUE permite utilizar la corrección por continuidad. El argumento alternative tiene la función de siempre*
kruskal.test (modelo)	*Realiza el **test múltiple de Kruskal-Wallis para la igualdad de medias de varias poblaciones.** El argumento modelo incluye la tabla de datos de las poblaciones como una lista de vectores columna*
friedman.test (modelo)	*Realiza el test múltiple de Kruskal-Wallis para la igualdad de medias de varias poblaciones. El argumento modelo incluye la tabla de datos de las poblaciones como una matriz*

Como primer ejemplo consideramos las variables *x* e *y* que se especifican a continuación y realizamos un contraste del coeficiente de correlación.

© Alfaomega - RC Libros

X	34675873548558885
Y	5587791047410579 1057

La sintaxis adecuada sería la siguiente:

```
> x=c(3, 4, 6, 7, 5, 8, 7, 3, 5, 4, 8, 5, 5, 8, 8, 8, 5)
> y=c(5, 5, 8, 7, 7, 9, 10, 4, 7, 4, 10, 5, 7, 9, 10, 5, 7)
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 4.5151, df = 15, p-value = 0.0004108
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4382534 0.9082981
sample estimates:
      cor
0.7590159
```

Se observa que el *p*-valor del test es muy pequeño con lo que se rechaza la hipótesis nula de variables incorreladas. Esto concuerda con el valor alto del propio coeficiente de correlación (0,759).

Como segundo ejemplo consideramos una encuesta realizada realizada a 100 hogares sobre el número de coches X de los que se dispone en el hogar y que arroja los siguientes resultados:

X_i	0	1	2	3
n_i	22	42	28	8

¿Pueden ajustarse estos datos a una distribución binomial (3,1/2)?

Definimos el vector de frecuencias absolutas *x* y el vector de probabilidades teóricas y a continuación efectuamos el test de ajuste mediante la siguiente sintaxis:

```
> x=c(22,42,28,8)
>
p1=c(dbinom(0,3,1/2),dbinom(1,3,1/2),dbinom(2,3,1/2),dbinom(3,3
,1/2))
> chisq.test(x,p=p1)

        Chi-squared test for given probabilities

data:  x
X-squared = 11.7867, df = 3, p-value = 0.008151
```


© Alfaomega - RC Libros

El *p*-valor muy pequeño nos lleva a rechazar la hipótesis nula de que los datos se ajustan bien a una distribución binomial.

Como segundo ejemplo se considera una muestra de 100 individuos extraídos de 3 barrios A, B y C y se les pregunta si son deportistas o no. Los datos acumulados se presentan en la tabla siguiente:

Barrio	Deportistas	No deportistas	Total
A	13	87	100
B	17	83	100
C	18	82	100
	48	252	300

Se trata de contrastar si los barrios son homogéneos en cuanto a personas deportistas.

Realizaremos el contraste de homogeneidad de la chi-cuadrado mediante la siguiente sintaxis:

```
> deportista=matrix(c(13,17,18,87,83,82),ncol=2)
> colnames(deportista)=c("Deportistas","No deportistas")
> rownames(deportista)=c("A","B","C")
> chisq.test(deportista)

        Pearson's Chi-squared test

data:  deportista
X-squared = 1.0417, df = 2, p-value = 0.594
```

Se obtiene un *p*-valor alto del contraste de homogeneidad, lo que permite aceptar la hipótesis nula de que los tres barrios son homogéneos a cuanto a nivel de deportistas.

Como tercer ejemplo se consideran dos variables categóricas *X* e *Y* cuya distribución conjunta viene dada por la tabla de contingencia siguiente:

Y \ X	x1	x2
y1	38	9
y2	31	32

© Alfaomega - RC Libros

Se trata de contrastar si ambas variables categóricas son independientes.

Realizaremos el contraste de independencia de caracteres de la chi-cuadrado mediante la siguiente sintaxis:

```
> tabla=matrix(c(38,31,9,22),ncol=2)
> colnames(tabla)=c("x1","x2")
> rownames(tabla)=c("y1","y2")
> chisq.test(tabla)

        Pearson's Chi-squared test with Yates' continuity
correction

data:  tabla
X-squared = 4.8243, df = 1, p-value = 0.02806
```

Se obtiene un *p*-valor pequeño en el contraste, lo que sugiere rechazar la hipótesis nula de independencia de las dos variables categóricas *X* e *Y*.

Como cuarto ejemplo consideramos la variable *X* que toma los siguientes valores:

X	2.3	2.9	1.6	2	4.2	3.1	2.3	2	1.6	3.3	4.1	4	3	2.8

¿Podríamos asegurar que el valor medio de esta variable puede ser 2,5?

Para contrastar esta hipótesis utilizaremos el test de los signos rangos de Wilcoxon. La sintaxis adecuada será la siguiente:

```
> x=c(2.3, 2.9, 1.6, 2, 4.2, 3.1, 2.3, 2, 1.6, 3.3, 4.1, 4, 3,
2.8)

> wilcox.test(x,mu=2.5,exact=F,correct=F)

        Wilcoxon signed rank test

data:  x
V = 69, p-value = 0.2996
alternative hypothesis: true location is not equal to 2.5
```

Hemos realizado el contraste sin corrección de continuidad y con el *p*-valor aproximado para muestras grandes (no se calcula el *p*-valor exacto de Wilcoxon). Hemos

obtenido un *p*-valor alto que indica que aceptamos la hipótesis nula de que el valor medio de la variable puede ser 2,5.

Como quinto ejemplo consideramos consideramos la variable *X* del ejemplo anterior y tratamos de contrastar si es normal. Para ello utilizaremos el test de Kolmogorov-Smirnov para la normalidad mediante la sintaxis siguiente:

```
> ks.test(x,"pnorm",3,1)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1866, p-value = 0.7142
alternative hypothesis: two-sided
```

El *p*-valor alto del contraste permite aceptar la hipótesis nula de normalidad de la variable. Además, el gráfico QQ de normalidad no se aleja mucho del diagonal del primer cuadrante (Figura 9-1), como se observa con la siguiente sintaxis:

```
> qqnorm(x)
```

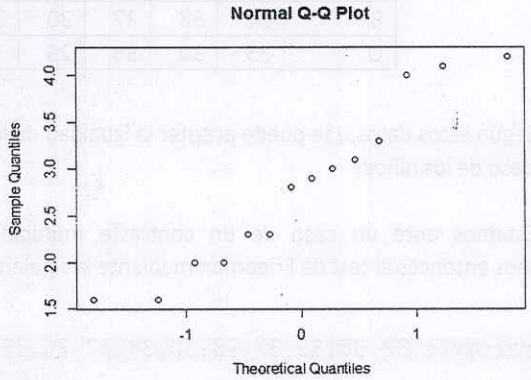


Figura 9-1

Como sexto ejemplo consideramos las variables *X* e *Y* siguientes:

X	3 4 6 7 5 8 7 3 5 4 8 5 5 8 8 8 5
Y	5 5 8 7 7 9 10 4 7 4 10 5 7 9 10 5 7

¿Podemos aceptar que ambas poblaciones son homogéneas en mediana?

Para ello utilizamos el contraste de Wilcoxon-Mann-Whitney para la igualdad de medianas de muestras independientes. Utilizaremos la siguiente sintaxis:

```
> x=c(3, 4, 6, 7, 5, 8, 7, 3, 5, 4, 8, 5, 5, 8, 8, 8, 5)
> y=c(5, 5, 8, 7, 7, 9, 10, 4, 7, 4, 10, 5, 7, 9, 10, 5, 7)
> wilcox.test(x,y,exact=F,correct=F)

        Wilcoxon rank sum test

data:  x and y
W = 102.5, p-value = 0.1409
alternative hypothesis: true location shift is not equal to 0
```

El *p*-valor alto del test permite aceptar la hipótesis nula de homogeneidad de poblaciones en mediana.

Como último ejemplo consideramos los pesos de 5 niños según tres dietas alimenticias distintas A, B y C que se reflejan en la tabla siguiente:

Dieta	P1	P2	P3	P4	P5
A	32	37	34	33	30
B	36	38	37	30	34
C	35	30	36	29	31

Según estos datos, ¿se puede aceptar la igualdad de los efectos de las tres dietas sobre el peso de los niños?

Estamos ante un caso de un contraste múltiple de igualdad de medias. Utilizaremos entonces el test de Friedman mediante la siguiente sintaxis:

```
>
datos=matrix(c(32,36,35,37,38,30,34,37,36,33,30,29,30,34,31),nc
ol=5)
> datos
     [,1] [,2] [,3] [,4] [,5]
[1,]   32   37   34   33   30
[2,]   36   38   37   30   34
[3,]   35   30   36   29   31
> colnames(datos)=c("P1","P2","P3","P4","P5")
> rownames(datos)=c("A","B","C")
> friedman.test(datos)
```


© Alfaomega - RC Libros


```
        Friedman rank sum test

data:  datos
Friedman chi-squared = 6.6667, df = 4, p-value = 0.1546
```

El *p*-valor alto del test de Friedman permite aceptar la hipótesis nula de igualdad de efectos de las tres dietas sobre el peso de los niños.

También podíamos haber utilizado el test de Kruskal-Wallis mediante la siguiente sintaxis:

```
>
datos1=list(c(32,36,35),c(37,38,30),c(34,37,36),c(33,30,29),c(3
0,34,31))
> kruskal.test(datos1)

        Kruskal-Wallis rank sum test

data:  datos1
Kruskal-Wallis chi-squared = 6.481, df = 4, p-value = 0.166
```

Observamos que para este test, los 5 vectores columna de pesos que componen la tabla se declaran como una lista. Al igual que en el test de Friedman, el *p*-valor alto del test de Kruskal-Wallis permite aceptar la hipótesis nula de igualdad de efectos de las tres dietas sobre el peso de los niños.

Ejercicio 9-1. Los valores de las longitudes en micras de 50 filamentos de la producción de una máquina (que se supone normal) son los siguientes:

102	98	93	100	98	105	115	110	99	120
115	130	100	86	95	103	105	92	99	134
116	118	89	102	128	99	119	128	110	130
112	114	106	114	100	116	108	113	106	105
120	106	110	100	106	117	109	108	105	106

a) Hallar un intervalo de confianza para la media de la producción basado en la muestra de los 50 filamentos al 95% de coeficiente de confianza. Hallarlo al 90%.
b) Contrastar la hipótesis nula de que la longitud media de los filamentos de la producción es de 100 a los niveles $\alpha = 5\%$ y $\alpha = 10\%$, basándose en la muestra dada.
c) Comprobar si la producción puede considerarse efectivamente normal.
d) ¿Cuál es la longitud de filamento más frecuente?
e) Hallar el valor de la longitud de los filamentos tal que la mitad de los restantes valores sea inferior a él.

© Alfaomega - RC Libros

Comenzamos introduciendo las longitudes de los filamentos en una variable *z*. A continuación realizamos el contraste de la media que también ofrece adicionalmente el intervalo de confianza al 95%.

```
>
z=c(102,98,93,100,98,105,115,110,99,120,115,130,100,86,95,103,1
05,92,99,134,116,118,89,102,128,99,119,128,110,130,112,114,106,
114,100,116,108,113,106,105,120,106,110,100,106,117,109,108,105
,106)
> t.test(z,mu=100,conf.level=0.95)

        One Sample t-test

data:  z
t = 5.4816, df = 49, p-value = 1.456e-06
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 105.3079 111.4521
sample estimates:
mean of x
   108.38
```

La media real es 108,38 y el intervalo de confianza al 95% será [105,308; 111,452]. Por otro lado, como el *p*-valor del contraste de la T es muy pequeño podemos rechazar la hipótesis de que la media es 100 al 95% de confianza.

Para calcular el intervalo de confianza para la media al 90% y realizar el contraste de la media, utilizamos la siguiente sintaxis:

```
> t.test(z,mu=100,conf.level=0.90)

        One Sample t-test

data:  z
t = 5.4816, df = 49, p-value = 1.456e-06
alternative hypothesis: true mean is not equal to 100
90 percent confidence interval:
 105.817 110.943
sample estimates:
mean of x
   108.38
```


© Alfaomega - RC Libros

El nuevo intervalo de confianza será [105,817; 110,943]. Por otro lado, como el *p*-valor del contraste de la T es muy pequeño podemos rechazar la hipótesis de que la media es 100 al 90% de confianza.

Además, como el valor 100 no pertenece a ninguno de los dos intervalos de confianza, se rechaza la hipótesis de que la longitud media de los filamentos de la producción es de 100 a los niveles α= 5% y α= 10%.

Para realizar el contraste formal de normalidad al 95% de la variable utilizamos el test de Kolmogorov-Smirnov. Sabemos que la media es 108,38. A continuación calculamos la desviación típica.

```
> ((length(z)-1)/length(z))*sd(z)
[1] 10.59364
```

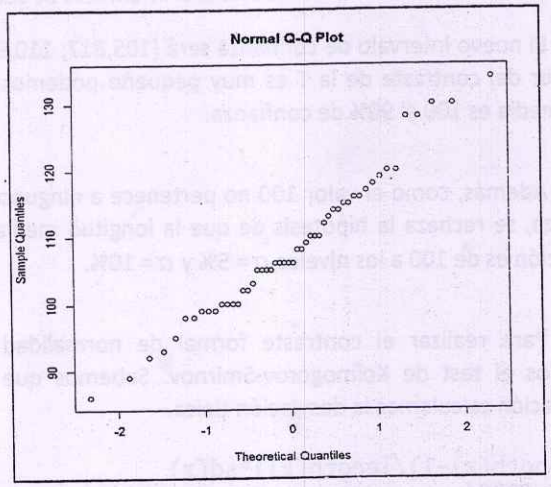
El contraste de normalidad de Kolmogorov-Smirnov se realiza mediante la siguiente sintasis:

```
> ks.test(x,"pnorm",108.38,10.59)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

El *p*-valor pequeño indica falta de normalidad formal al 95% de confianza. Pero, esa exigencia al 95 % es muy fuerte. Para ver si hay aproximación a la normal utilizamos el gráfico Q-Q de normalidad.

```
> qqnorm(z)
```

© Alfaomega - RC Libros



Se observa que la nube de puntos se ajusta bien a la diagonal del primer cuadrante, lo que indica que se puede aceptar la normalidad débil de la variable.

Los dos últimos apartados del ejercicio nos piden la moda y la mediana de la variable.

Para calcular la moda observamos el valor de mayor frecuencia absoluta.

```
> t=table(z)
> t

z
 86  89  92  93  95  98  99 100 102 103 105 106 108 109 110 112
113 114 115 116
  1   1   1   1   1   2   3   4   2   1   4   5   2   1   3   1
1   2   2   2
117 118 119 120 128 130 134
  1   1   1   2   2   2   1
```

La moda resulta ser el valor 106.

A continuación se calcula la mediana.

```
> median(z)
[1] 106
```

La mediana también resulta ser 106.

© Alfaomega - RC Libros

Ejercicio 9-2. Los valores del pH de una solución en 10 determinaciones diferentes son los siguientes: 6,80; 6,78; 6,77; 6,80; 6,78; 6,80; 6,82; 6,81; 6,80 y 6,79. Suponiendo normal la distribución de la población de todas las determinaciones del pH de esa solución:

a) Hallar un intervalo de confianza al 95% para la media poblacional.
b) Hallar un intervalo de confianza al 65% para la media poblacional.
c) Contrastar la hipótesis nula de que la media poblacional es 6,8 frente a la alternativa de que sea distinta, a un nivel de significación $\alpha = 0,05$, y a un nivel $\alpha' = 0,35$.

Comenzaremos introduciendo los datos de la muestra como una variable y a continuación realizamos los contrastes y el cálculo de los intervalos de confianza.

```
> z=c(6.80, 6.78, 6.77, 6.80, 6.78, 6.80, 6.82, 6.81, 6.80,
6.79)
> t.test(z,mu=6.8,conf.level=0.95)

        One Sample t-test

data:  z
t = -1.0476, df = 9, p-value = 0.3221
alternative hypothesis: true mean is not equal to 6.8
95 percent confidence interval:
 6.784204 6.805796
sample estimates:
mean of x
    6.795

> t.test(z,mu=6.8,conf.level=0.65)

        One Sample t-test

data:  z
t = -1.0476, df = 9, p-value = 0.3221
alternative hypothesis: true mean is not equal to 6.8
65 percent confidence interval:
 6.790295 6.799705
sample estimates:
mean of x
    6.795
```

El intervalo de confianza al 95% para la media es [6,7842; 6,8058].

© Alfaomega - RC Libros

El intervalo de confianza al 65% para la media es [6,7903; 6,7997].

Por otra parte, se acepta la hipótesis de que la media es 6,8 con $\alpha = 0{,}05$, ya que el intervalo de confianza hallado para la media al 95% contiene el valor 6,8. Por otra parte, se rechaza la hipótesis de que la media es 6,8 con $\alpha = 0{,}35$, ya que el intervalo de confianza hallado para la media al 65% no contiene el valor 6,8.

Ejercicio 9.3 Las presiones críticas de dos grupos independientes de recipientes de distintos vidrios dan los siguientes valores:

Grupo 1º: 100; 102; 96; 106; 110; 110; 120; 112; 112; 90
Grupo 2º: 104; 88; 100; 98; 102; 92; 96; 100; 96; 96

a) Suponiendo que las dos poblaciones son normales y de varianzas iguales y desconocidas, calcular un intervalo de confianza del 95% para la diferencia de presiones medias, y contrastar la hipótesis de igualdad de presiones medias (α=0,05). Contrastar también la hipótesis de igualdad de varianzas al mismo nivel.

b) ¿Qué ocurre en el caso de varianzas desiguales?

Comenzaremos introduciendo los datos de las dos muestras como dos variables llamadas *x* e *y*. El contraste de igualdad de medias de variables independientes normales con varianzas iguales y el intervalo de confianza para la diferencia de medias se calculan con la siguiente sintaxis de R:

```
> x=c(100, 102, 96, 106, 110, 110, 120, 112, 112, 90);
> y=c(104, 88, 100, 98, 102, 92, 96, 100, 96, 96);

> t.test(x,y,alternative="two.sided", mu=0, paired= FALSE,
var.equal = TRUE, conf.level=0.95)

        Two Sample t-test

data:  x and y
t = 2.7058, df = 18, p-value = 0.01447
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.922431 15.277569
sample estimates:
mean of x mean of y
    105.8      97.2
```


© Alfaomega - RC Libros

El contraste de igualdad de varianzas de variables independientes normales y el intervalo de confianza para el cociente de varianzas se calculan con la siguiente sintaxis de R:

```
> var.test(x, y, 1, conf.level=0.95)

        F test to compare two variances

data:  x and y
F = 3.5099, num df = 9, denom df = 9, p-value = 0.07538
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
  0.8718146 14.1309200
sample estimates:
ratio of variances
          3.509921
```

Se observa que el intervalo de confianza para la diferencia de medias es [1,92242; 15,2776], que no contiene el valor cero, lo que nos lleva a rechazar la hipótesis de igualdad de medias al 95%. El *p*-valor = 0,01447 que es menor que 0,05 corrobora este resultado. También se observa que el intervalo de confianza para el cociente de varianzas es [0,8718; 14,1309] que sí contiene el valor uno, lo que nos lleva a aceptar la hipótesis de igualdad de varianzas al 95%. El *p*-valor = 0,07538 que es mayor que 0,05 corrobora este resultado.

Si se quiere el intervalo de confianza para la diferencia de medias sin asumir varianzas iguales, se utilizará la sintaxis R siguiente:

```
> x=c(100, 102, 96, 106, 110, 110, 120, 112, 112, 90);
> y=c(104, 88, 100, 98, 102, 92, 96, 100, 96, 96);
> t.test(x,y,alternative="two.sided", mu=0, paired= FALSE,
var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  x and y
t = 2.7058, df = 13.743, p-value = 0.01729
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.771046 15.428954
sample estimates:
mean of x mean of y
    105.8      97.2
```

© Alfaomega - RC Libros

El intervalo de confianza obtenido para la diferencia de medias es ahora [1,771; 15,428]. Este intervalo no contiene el valor cero, por lo que se rechazará la hipótesis de que las medias son iguales (diferencia de medias nula) al 95%. El *p*-valor = 0,01729 que es menor que 0,05 corrobora el resultado.

El intervalo de confianza para el cociente de varianzas se realiza igual que en el caso anterior, ya que no depende del hecho de que las varianzas desconocidas sean o no iguales.

Ejercicio 9.4 Las presiones críticas de dos grupos independientes de recipientes de distintos vidrios dan los siguientes valores:

Grupo 1º: 100; 102; 96; 106; 110; 110; 120; 112; 112; 90
Grupo 2º: 104; 88; 100; 98; 102; 92; 96; 100; 96; 96

a) Suponiendo que las dos poblaciones son normales y de varianzas iguales y desconocidas, contrastar la hipótesis de igualdad de medias al nivel 0,05.
b) Realizar el contraste para varianzas desiguales.
c) Realizar el contraste para muestras pareadas.
d) Contrastar también la igualdad de varianzas.

Comenzaremos introduciendo los datos de las dos muestras en dos variables llamadas *X* e *Y*.

Para contrastar la hipótesis de la igualdad de medias con varianzas iguales (2 colas) utilizamos la siguiente sintaxis de R:

```
> x=c(100, 102, 96, 106, 110, 110, 120, 112, 112, 90);
> y=c(104, 88, 100, 98, 102, 92, 96, 100, 96, 96);
> t.test(x,y,alternative="two.sided", mu=0, paired= FALSE,
var.equal = TRUE, conf.level=0.95)

        Two Sample t-test

data:  x and y
t = 2.7058, df = 18, p-value = 0.01447
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.922431 15.277569
sample estimates:
mean of x mean of y
    105.8      97.2
```


© Alfaomega - RC Libros

Se observa que el intervalo de confianza para la diferencia de medias no contiene el valor cero, por lo que se rechazará la hipótesis de que las medias son iguales (diferencia de medias nula) al 95%. El *p*-valor = 0,01447 que es menor que 0,05 corrobora el resultado.

Para contrastar la hipótesis de la igualdad de medias con varianzas desiguales (2 colas) utilizamos la siguiente sintaxis de R:

```
> t.test(x,y,alternative="two.sided", mu=0, paired= FALSE,
var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  x and y
t = 2.7058, df = 13.743, p-value = 0.01729
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.771046 15.428954
sample estimates:
mean of x mean of y
    105.8      97.2
```

Se observa que el intervalo de confianza para la diferencia de medias no contiene el valor cero, por lo que se rechazará la hipótesis de que las medias son iguales (diferencia de medias nula) al 95%. El *p*-valor = 0,01729 que es menor que 0,05 corrobora el resultado.

Para contrastar la hipótesis de la igualdad de medias con varianzas iguales para muestras pareadas (2 colas) utilizamos la siguiente sintaxis de R:

```
> t.test(x,y,alternative="two.sided", mu=0, paired= TRUE,
var.equal = TRUE, conf.level=0.95)

        Paired t-test

data:  x and y
t = 2.6437, df = 9, p-value = 0.02675
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.241129 15.958871
sample estimates:
mean of the differences
                    8.6
```

© Alfaomega - RC Libros

Se observa que el intervalo de confianza para la diferencia de medias no contiene el valor cero, por lo que se rechazará la hipótesis de que las medias son iguales (diferencia de medias nula) al 95%. El *p*-valor = 0,02675 que es menor que 0,05 corrobora el resultado.

Para contrastar la hipótesis de la igualdad de medias con varianzas desiguales para muestras pareadas (2 colas) utilizamos la siguiente sintaxis de R:

```
> t.test(x,y,alternative="two.sided", mu=0, paired= TRUE,
var.equal = FALSE, conf.level=0.95)

        Paired t-test

data:  x and y
t = 2.6437, df = 9, p-value = 0.02675
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  1.241129 15.958871
sample estimates:
mean of the differences
                    8.6
```

Se observa que el intervalo de confianza para la diferencia de medias no contiene el valor cero, por lo que se rechazará la hipótesis de que las medias son iguales (diferencia de medias nula) al 95%. El *p*-valor = 0,02675 que es menor que 0,05 corrobora el resultado.

Para contrastar la hipótesis de igualdad de varianzas utilizamos la siguiente sintaxis de R:

```
> var.test(x, y, 1, conf.level=0.95)

        F test to compare two variances

data:  x and y
F = 3.5099, num df = 9, denom df = 9, p-value = 0.07538

alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
  0.8718146 14.1309200
sample estimates:
```


© Alfaomega - RC Libros


```
ratio of variances
          3.509921
```

Se observa que el intervalo de confianza para el cociente de varianzas contiene el valor 1, lo que indica que se acepta la hipótesis de igualdad de varianzas al 95% de confianza. Asimismo, el *p*-valor del contraste p= 0,07538 es mayor que 0,05, lo que indica la aceptación de la hipótesis de igualdad de varianzas.

Ejercicio 9.5. Una central de productos lácteos recibe diariamente la leche de dos granjas, X e Y. Con el fin de estudiar la calidad de los productos recibidos, se extraen dos muestras, una de cada granja, y se analiza el contenido de materia grasa, obteniendo los siguientes resultados:

X→ 0,32; 0,29; 0,30; 0,28; 0,33; 0,31; 0,30; 0,29; 0,33; 0,32; 0,30; 0,29.
Y→ 0,28; 0,30; 0,32; 0,29; 0,31; 0,29; 0,33; 0,32; 0,29; 0,32; 0,31; 0,29; 0,32; 0,31; 0,32; 0,33.

Realizar el contraste de hipótesis de igualdad de varianzas.

Comenzaremos introduciendo los datos de las dos muestras como dos variables llamadas *X* e *Y*. Para contrastar la hipótesis de la igualdad de varianzas, utilizamos la siguiente sintaxis de R.

```
> x=c(0.32, 0.29, 0.30, 0.28, 0.33, 0.31, 0.30, 0.29, 0.33,
0.32, 0.30, 0.29)
> y=c(0.28, 0.30, 0.32, 0.29, 0.31, 0.29, 0.33, 0.32, 0.29,
0.32, 0.31, 0.29, 0.32, 0.31, 0.32, 0.33)
> var.test(x, y, 1, conf.level=0.95)

        F test to compare two variances

data:  x and y
F = 1.0998, num df = 11, denom df = 15, p-value = 0.8449
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
 0.3656387 3.6621900
sample estimates:
ratio of variances
          1.099778
```

© Alfaomega - RC Libros

Se observa que el *p*-valor es mayor que 0,05, lo que implica que el cociente de varianzas es significativamente igual a la unidad, o sea, la aceptación de la igualdad de varianzas. Por otra parte, el intervalo de confianza para el cociente de varianzas contiene el valor 1, lo que corrobora la aceptación de la hipótesis de igualdad de varianzas.

Ejercicio 9.6. Se desean comparar dos nuevas clases de trigo. Para ello, se toman 10 fincas al azar, plantando en cada una de ellas, en dos partes distintas de las mismas, ambas clases. La producción en las 10 fincas fue la siguiente:

Clase A → 57 49 60 55 57 48 50 61 52 56
Clase B → 55 48 58 56 54 48 52 56 50 58

¿Podemos aceptar la homogeneidad de producciones para ambas clases de trigo a un nivel $\alpha = 0{,}05$ suponiendo distribución normal bivariante?

Comenzaremos introduciendo los datos de las dos muestras como dos variables *X* e *Y*. El contraste de homogeneidad de producciones lo realizamos a través del test de Kolmogorov-Smirnov de homogeneidad mediante la siguiente sintaxis de R:

```
> ks.test(x,y)
        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.2, p-value = 0.9883
alternative hypothesis: two-sided
```

Según los resultados, se acepta la homogeneidad de muestras ya que el *p*-valor es mayor que 0,05. Esto indica que las producciones de las dos clases de trigos pueden considerarse idénticas al 95% de confianza.

Ejercicio 9.7. Se supone que la estatura E y el peso P de las personas no son variables independientes. Verificar si esta hipótesis es cierta al 90%, basándose en la siguiente muestra de estaturas y pesos:

E	P
178,8	92,9
168,1	87,2
167,9	85,1
166,9	76,1

© Alfaomega - RC Libros

184,5	99,5
168,1	86,5
173,5	93,2
167	77,1
168,7	92,6
165,7	65

Dado que la distribución conjunta es normal, para contrastar la independencia bastará con contrastar la hipótesis de que el coeficiente de correlación poblacional es significativamente igual a cero con un nivel de significación del 10%.

Para contrastar la hipótesis de que el coeficiente de correlación poblacional es igual a cero, frente a la alternativa de que es distinto de cero, utilizaremos el estadístico:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

que se distribuye como una *T* de Student con $n - 2$ grados de libertad.

La sintaxis en R sería la siguiente:

```
> E=c(178.8,168.1,167.9,166.9,184.5,168.1,173.5,167,168.7)
> P=c(92.9, 87.2, 85.1, 76.1, 99.5, 86.5, 93.2, 77.1, 92.6, 65)
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 4.9569, df = 8, p-value = 0.001111
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5272635 0.9685261
sample estimates:
      cor
0.8685496
```

Se observa que el *p*-valor del test es muy pequeño con lo que se rechaza la hipótesis nula de variables incorreladas. Esto concuerda con el valor alto del propio coeficiente de correlación (0,759). Por lo tanto, se rechaza la independencia.

© Alfaomega - RC Libros

Se concluye que, como las variables estatura y peso tienen una distribución conjunta normal y están correlacionadas, no son independientes.

Ejercicio 9.8. En un examen final de estadística, los estudiantes recibieron las siguientes calificaciones:

80, 70, 90, 75, 55, 80, 80, 65, 100, 75, 60, 60,75, 95, 80, 80, 90, 85, 70, 95, 75, 70, 85, 80, 80, 65, 65, 50, 75, 75, 85, 85, 90, 70.

Compruébese si las calificaciones fueron o no distribuidas según una ley normal a un nivel 0,05.

Comenzaremos introduciendo los datos de las calificaciones como una variable de nombre *x*.

```
> x=c(80, 70, 90, 75, 55, 80, 80, 65, 100, 75, 60, 60,75, 95,
80, 80, 90, 85, 70, 95, 75, 70, 85, 80, 80, 65, 65, 50, 75, 75,
85, 85, 90, 70)
> ((length(x)-1)/length(x))*sd(z)
[1] 0.01464842
> mean(x)
[1] 76.76471
> ks.test(x,"pnorm",76.76471,0.01464842)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.5, p-value = 8.28e-08
alternative hypothesis: two-sided
```

El *p*-valor del contraste es menor que 0,05, lo que nos lleva a rechazar la hipótesis nula de ajuste de los datos a una normal.

Ejercicio 9.9. En el transcurso de 2 horas, el número de llamadas por minuto solicitadas a una centralita telefónica presenta la siguiente distribución:

Nº llamadas/minuto	0	1	2	3	4	5	6	7
Frecuencias	6	18	32	35	17	10	2	0

Compruébese si la variable número de llamadas/minuto se distribuye o no según una ley de Poisson a un nivel 0,01.

© Alfaomega - RC Libros

Definimos el vector de valores de la variable *x* y el vector de frecuencias absolutas *n*. A continuación calculamos el parámetro *l* de la distribución de Poisson que es su media.

```
> x=c(0,1,2,3,4,5,6,7)
> n=c(6,18,32,35,17,10,2,0)
> l=(1/sum(n))*sum(x*n)
> l
[1] 2.641667
```

A continuación calculamos el vector de probabilidades teóricas *p1* y efectuamos el test de ajuste mediante la siguiente sintaxis:

```
>
p1=c(dpois(0,l),dpois(1,l),dpois(2,l),dpois(3,l),dpois(4,l),dpo
is(5,l),dpois(6,l),dpois(7,l))
> p1
[1] 0.07124243 0.18819876 0.24857920 0.21888779 0.14455715
0.07637436 0.03362593
[8] 0.01268979

> chisq.test(n,p1)

        Pearson's Chi-squared test

data:  n and p1
X-squared = 56, df = 49, p-value = 0.2289
```

Observamos que el *p*-valor del contraste es 0,2289, que es mayor que 0,05, lo que nos lleva a aceptar la hipótesis nula de ajuste de los datos a una ley de Poisson.

Ejercicio 9.10. Se considera una muestra de 100 hogares unifamiliares con 5 hijos cada uno y se analiza la distribución de la variable X = número de hijos varones obteniéndose los siguientes resultados:

X	*0*	*1*	*2*	*3*	*4*	*5*
Nº de familias	31	168	319	308	150	24

Ajustar estos datos a una distribución binomial y medir la calidad del ajuste.

© Alfaomega - RC Libros

Estamos ante un caso de ajuste de un conjunto de datos a una distribución binomial $X \rightarrow B(5,p)$.

El parámetro p se determinará a partir de la propiedad de la distribución binomial que asegura que su media es precisamente np. Por tanto haremos:

$$\bar{x} = np \Rightarrow p = \frac{\bar{x}}{n} = \frac{2{,}455}{5} = 0{,}49$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{6} n_i x_i = \frac{1}{1000}(31*0+168*1+319*2+308*3+150*4+24*5) = 2{,}455$$

Los cálculos con R se realizan de la siguiente forma:

```
> x=c(0,1,2,3,4,5)
> n=c(31,168,319,308,150,24)
> p=(1/5)*(1/sum(n))*sum(x*n)
> p
[1] 0.49
```

Las probabilidades teóricas o esperadas p_i se calcularán mediante la ley de probabilidad de la distribución binomial, con lo que tenemos la siguiente tabla:

x_i	n_i	$p_i = P(X = x_i)$	np_i
0	31	$\binom{5}{0} 0{,}49^0 0{,}51^5 = 0{,}0345$	34,5
1	168	$\binom{5}{1} 0{,}49^1 0{,}51^4 = 0{,}1657$	165,7
2	319	$\binom{5}{2} 0{,}49^2 0{,}51^3 = 0{,}3185$	318,5
3	308	$\binom{5}{3} 0{,}49^3 0{,}51^2 = 0{,}3060$	306,0
4	150	$\binom{5}{4} 0{,}49^1 0{,}51^4 = 0{,}1470$	147,0
5	24	$\binom{5}{5} 0{,}49^5 0{,}51^0 = 0{,}0283$	28,3

© Alfaomega - RC Libros

Estos cálculos se realizan en R de la siguiente forma:

```
>
p1=c(dbinom(0,5,p),dbinom(1,5,p),dbinom(2,5,p),dbinom(3,5,p),db
inom(4,5,p),dbinom(5,5,p))
> p1
[1] 0.03450253 0.16574742 0.31849505 0.30600505 0.14700243
0.02824752
```

Ahora ya podemos realizar el contraste de ajuste mediante la siguiente sintaxis:

```
> chisq.test(n,p1)

        Pearson's Chi-squared test

data:  n and p1
X-squared = 30, df = 25, p-value = 0.2243
```

Como el p-valor del contraste es mayor que 0,05 se acepta la hipótesis nula de que el ajuste de los datos a una binomial es correcto con un nivel de significación $\alpha = 0{,}5$.

Ejercicio 9.11. Se realiza un análisis de la duración de 40 pilas alcalinas obteniéndose los siguientes resultados:

Duraciones x_i	*Frecuencias absolutas* n_i
1,55 – 1,95	2
1,95 – 2,45	1
2,45 – 2,95	4
2,95 – 3,45	15
3,45 – 3,95	10
3,95 – 4,45	5
4,45 – 4,95	3

Ajustar las duraciones de las pilas alcalinas a una distribución normal con $\mu = 3{,}5$ y $\sigma = 0{,}7$.

© Alfaomega - RC Libros

Las probabilidades teóricas o esperadas p_i se calcularán mediante la distribución normal mediante:

$$P(1{,}45 \le x \le 1{,}95) = P\left(\frac{1{,}45-3{,}5}{0{,}7} \le \frac{x-3{,}5}{0{,}7} \le \frac{1{,}95-3{,}5}{0{,}7}\right) = P(-293 \le z \le -2{,}21) = 0{,}012$$

De igual forma se calculan las del resto de las clases, con lo que tenemos la siguiente tabla:

Clases	n_i	p_i	np_i
1,45 – 1,95	2	0,012	0,5
1,95 – 2,45	1	0,0525	2,1
2,45 – 2,95	4	0,1475	5,9
2,95 – 3,45	15	0,2573	10,3
3,45 – 3,95	10	0,2674	10,7
3,95 – 4,45	5	0,175	7,0
4,45 – 4,95	3	0,0874	3,5

No obstante, suele ser habitual agrupar los intervalos consecutivos cuyas frecuencias teóricas son menores que 5, en cuyo caso las clases se reducen a 4 y la tabla es la siguiente:

Clases	n_i	p_i	np_i
1,55 – 2,95	7	0,212	8,5
2,95 – 3,45	15	0,2573	10,3
3,45 – 3,95	10	0,2674	10,7
3,95 – 4,95	8	0,2624	10,5

Con estos datos ya podemos realizar el contraste de la chi-cuadrado de ajuste a una normal. Se utilizará la siguiente sintaxis de R:

```
> p1=c(pnorm(2.95,3.5,0.7)-pnorm(1.5,3.5,0.7),
pnorm(3.45,3.5,0.7)-pnorm(2.95,3.5,0.7),pnorm(3.95,3.5,0.7)-
pnorm(3.45,3.5,0.7),pnorm(4.95,3.5,0.7)-pnorm(3.95,3.5,0.7))
> p1
[1] 0.2138801 0.2555109 0.2683133 0.2409990
> n=c(7,15,10,8)
> n
```


© Alfaomega - RC Libros


```
[1] 7 15 10  8
> chisq.test(n,p1)

        Pearson's Chi-squared test

data:  n and p1
X-squared = 12, df = 9, p-value = 0.2133
```

Como el *p*-valor del contraste es mayor que 0,05 se acepta la hipótesis nula de que el ajuste de los datos a una binomial es correcto con un nivel de significación $\alpha = 0{,}5$.

Ejercicio 9.12. Se trata de realizar un estudio para averiguar si con el 95% de confianza se puede concluir que el desarrollo del cáncer de pulmón está relacionado con el hábito de fumar. Para ello se encuesta a un grupo de personas obteniendo los siguientes resultados:

Enfermedad / calidad de fumador	*No fumador*	*Moderado*	*Empedernido*
SÍ	350	1200	1450
NO	525	900	575

¿Se puede concluir realmente que con una probabilidad del 95% el desarrollo del cáncer de pulmón está relacionado con el hábito de fumar?

Realizaremos un contraste de independencia entre el desarrollo del cáncer de pulmón y el hábito de fumar a través de la chi-cuadrado. Para ello utilizamos la siguiente sintaxis R:

```
> tabla=matrix(c(350,525,1200,900,1450,575),ncol=3)
> colnames(tabla)=c("No fumador", "Moderado", "Empedernido")
> rownames(tabla)=c("SI","NO")
> chisq.test(tabla)

        Pearson's Chi-squared test

data:  tabla
X-squared = 266.61, df = 2, p-value < 2.2e-16
```

© Alfaomega - RC Libros

Como el *p*-valor es muy pequeño se rechaza la hipótesis nula de independencia entre el desarrollo del cáncer de pulmón y el hábito de fumar con un nivel de significación $\alpha = 0{,}5$. Es decir, el hábito de fumar influye en el cáncer de pulmón con una probabilidad del 95%.

Ejercicio 9.13. En un análisis de satisfacción de los ciudadanos con los cuidados hospitalarios se realiza una encuesta a 300 pacientes de los tres hospitales A, B y C de una ciudad. Se obtuvieron los siguientes resultados:

	Satisfechos	*Insatisfechos*	$n_{i.}$
A	47	53	100
B	40	60	100
C	27	73	100
$n_{.j}$	114	186	300

Contrastar la hipótesis de que hay homogeneidad en la satisfacción de los pacientes con los cuidados recibidos en los tres hospitales de la ciudad.

Se trata de realizar un contraste de homogeneidad a través del contraste de la chi-cuadrado para los datos de la tabla dada. La sintaxis apropiada de R sería la siguiente:

```
> tabla=matrix(c(47,40,27,53,60,73),ncol=2)
> colnames(tabla)=c("Satisfechos", "Insatisfechos")
> rownames(tabla)=c("A","B","C")
> chisq.test(tabla)

        Pearson's Chi-squared test

data:  tabla
X-squared = 8.7436, df = 2, p-value = 0.01263
```

Como el *p*-valor es menor que 0,05, se rechaza la hipótesis nula de homogeneidad al 95% entre la satisfacción de los ciudadanos con la atención recibida en los tres hospitales de la ciudad.

© Alfaomega - RC Libros

Ejercicio 9.14. Se extrae una muestra de tamaño 10 de una variable aleatoria X cuyos resultados son los siguientes:

0,62 0,36 0,23 0,76 0,65 0,09 0,55 0,26 0,38 0,24

Se quiere contrastar la hipótesis de que la muestra proviene de una variable uniforme [0,1] utilizando el contraste de Kolmogorov-Smirnov con $\alpha = 0{,}05$.

Para utilizar el contraste de Kolmogorov-Smirnov de ajuste a una distribución uniforme unitaria uasmos la siguiente sintaxis de R.

```
> x=c(0.62,0.36, 0.23,0.76, 0.65, 0.09, 0.55, 0.26, 0.38, 0.24)
> x
 [1] 0.62 0.36 0.23 0.76 0.65 0.09 0.55 0.26 0.38 0.24

> ks.test(x,"punif",0,1)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.25, p-value = 0.4841
alternative hypothesis: two-sided
```

Como el *p*-valor del contraste es mayor que 0,05, se acepta la hipótesis nula de que los datos provienen de una distribución uniforme [0,1] con un nivel de significación $\alpha = 0{,}05$.

© Alfaomega - RC Libros

CAPÍTULO 10

GRÁFICOS 2-D Y 3-D EN R

GRAFICOS BÁSICOS EN R. EL COMANDO PLOT PARA GRÁFICOS DE SECUENCIA Y DISPERSIÓN

El paquete gráfico del módulo básico de R permite realizar una importante variedad de gráficos. Es posible representar funciones y realizar gráficos estadísticos de todo tipo. El comando más básico para realizar gráficos es *plot*. Permite realizar gráficos de secuencia y gráficos de dispersión. Su sintaxis es la siguiente:

plot(x)	*Representa la variable x contra un índice. Es decir, representa los puntos (t,x_t)*
plot(x, type="tipo")	*Representa la variable x contra un índice y une los puntos (t,x_t) con el carácter indicado en el argumento tipo. Los caracteres posibles son:* • *"p" para puntos* • *"l" para líneas* • *"b" para puntos y líneas* • *"c" para puntos y líneas al lado* • *"o" para puntos y líneas superpuestos* • *"h" para histograma con líneas verticales* • *"s" para escalones* • *"S" para otros escalones* • *"n" para no pintar el gráfico*
plot(x, main="título")	*Representa los puntos (t,x) y sitúa un título principal en el gráfico*

plot(x, sub="subtítulo")	*Representa los puntos (t,x_t) y sitúa un subtítulo en el gráfico*
plot(x, xlab="título del eje x")	*Representa los puntos (t,x_t) y sitúa un título principal en el gráfico*
plot(x, ylab="título del eje y")	*Representa los puntos (t,x_t) y sitúa un subtítulo en el gráfico*
plot(x, asp="ratio")	*Representa los puntos (t,x_t) y sitúa un ratio de aspecto y/x en el gráfico*
plot(f, a,b)	*Representa la función f en el intervalo [a,b]*
plot(x,y)	*Representa el gráfico de dispersión (x_t,y_t). Admite los mismos argumentos que plot(x)*

Consideramos el conjunto de datos *coches.sav* en formato SPSS que se encuentra situado en el subdirectorio *H:\R\data* del disco duro y que contiene las variables realtivas al consumo (*consumo*), motorización (*motor*), potencia (*cv*), peso (*peso*), aceleración (*acel*), año de fabricación (*año*), origen geográfico (*origen*) y cilindrada (*cilindr*).

En primer lugar leemos el fichero, habilitamos sus variables y graficamos el consumo contra el número de caso (Figura 10-1).

```
> library(foreign)
coches=read.spss("H:/R/data/coches.sav", use.value.labels =
TRUE)
> attach(coches)
> plot(consumo)
```

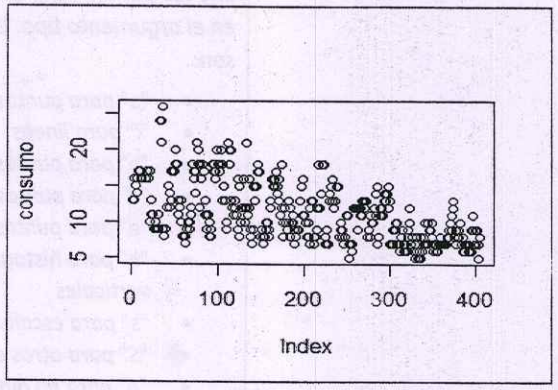


Figura 10-1

© Alfaomega - RC Libros

A continuación realizamos el gráfico anterior uniendo sus puntos por líneas y situando un título principal y un subtítulo (Figura 10-2).

```
> plot(consumo, type="l", main="Consumo de los automóviles",
sub="Litros a los 100 kilómetros")
```

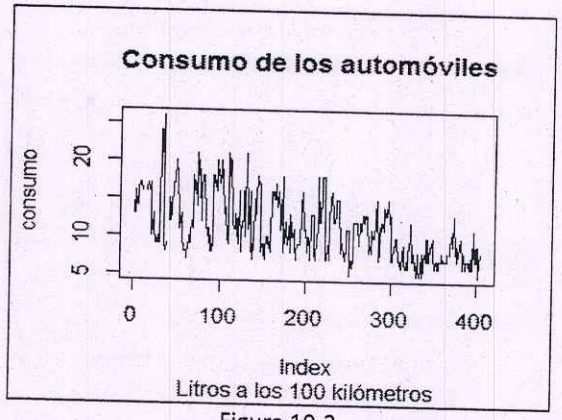


Figura 10-2

A continuación situamos etiquetas adecuadas a los ejes *X* e *Y* (Figura 10-3).

```
> plot(consumo, type="l", main="Consumo de los automóviles",
ylab="Consumo", xlab="Número de caso")
```

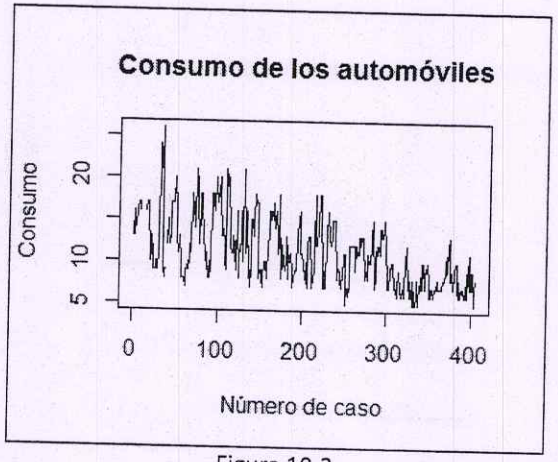


Figura 10-3

© Alfaomega - RC Libros

A continuación se realiza un gráfico de dispersión del consumo contra la potencia.

```
> plot(consumo,cv)
```

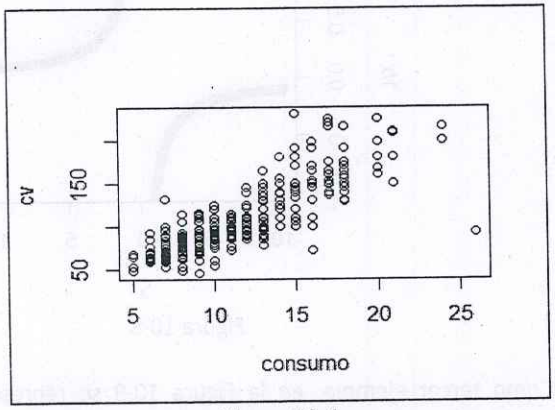


Figura 10-4

El comando *plot* también permite representar funciones. A continuación graficamos la parábola $y=x^2$ para x variando entre 0 y 100 (Figura 10-5).

```
> x=(1:100)
> plot(x,x^2,type="l",main="Gráfico de la parábola y=x^2")
```

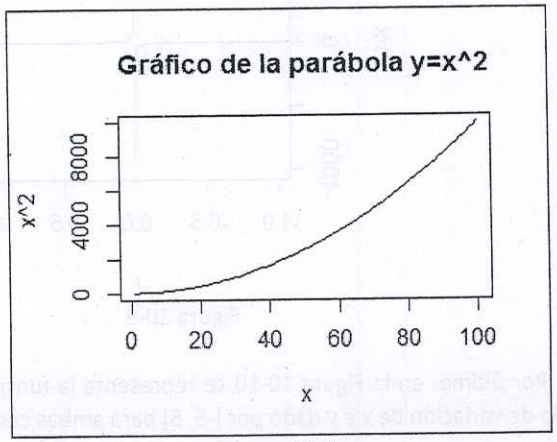


Figura 10-5

© Alfaomega - RC Libros

CURVAS PLANAS EN COORDENADAS PARAMÉTRICAS

El comando *plot* permite representar curvas planas en coordenadas paramétricas. Como ejemplo representamos la curva en coordenadas paramétricas:

$$x = x(t) = 4cos(t) - cos(4t)$$
$$y = y(t) = 4sen(t)\text{-}sen(4t)$$

cuando el parámetro t varía entre $-\pi$ y π.

La sintaxis de R adecuada sería la siguiente:

```
> t=seq(-pi,pi, by=pi/100)
> plot(4*cos(t)-cos(4*t),4*sin(t)-sin(4*t),type="l",
main="Gráfico en coordenadas paramétricas")
```

Se obtiene el gráfico de la Figura 10-6.

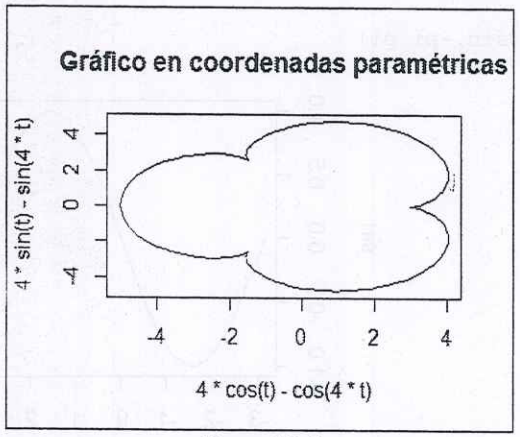


Figura 10-6

CURVAS PLANAS EN COORDENADAS EXPLÍCITAS

La función o comando *plot* de R permite representar gráficamente curvas planas en coordenadas explícitas. Las sintaxis más habituales de *plot* para representar curvas en coordenadas explícitas son las siguientes:

© Alfaomega - RC Libros

plot(f, a,b)	*Representa la función f en el intervalo [a,b]*
plot(x,f(x))	*Representa la función f(x) en el campo de variación de x previamente definido*
plot(x,f(x), xlim = c(a,b))	*Representa la función f(x) para x variando entre a y b*
plot(x,f(x), ylim = c(d,e))	*Representa la función f(x) en el intervalo de variación de x previamente definido y para y variando entre d y e*
plot(x,f(x), xlim=c(a,b), ylim=c(d,e))	*Representa la función f(x) para x variando entre a y b e y variando entre d y e*
plot(x,f(x), xlim=range(c(a,b)), ylim=range(c(d,e)))	*Representa la función f(x) para x variando entre a y b e y variando entre d y e*

Como primer ejemplo, en la Figura 10-7 se representa la función *f(x)=sen(x)* en el intervalo [-π, π].

```
> plot(sin,-pi,pi)
```

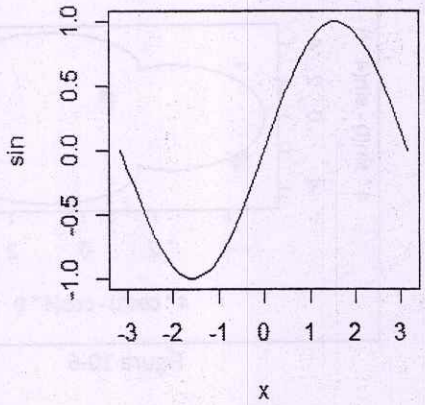


Figura 10-7

Como segundo ejemplo, en la Figura 10-8 se representa la función *f(x)=1/x* en el intervalo de variación de *x* dado por [-10, 10] definido de milésima en milésima y para el intervalo de variación y entre -1 y 1.

© Alfaomega - RC Libros


```
> x=seq(-10,10, by=1/1000)
> plot(x,1/x, ylim=c(-1,1))
```

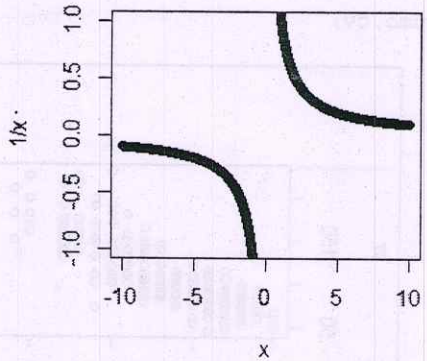


Figura 10-8

Como tercer ejemplo, en la Figura 10-9 se representa la función anterior *f(x)=1/x* en el intervalo de variación de *x* dado por [-1, 1].

```
> plot(x,1/x, xlim=c(-1,1), type="l")
```

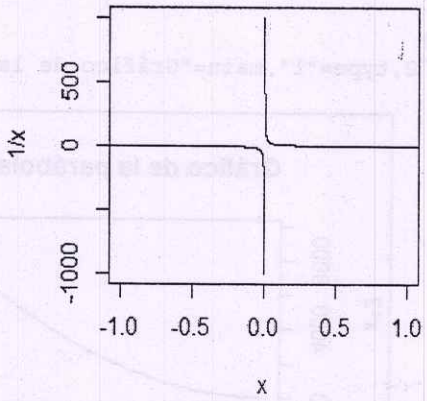


Figura 10-9

Por último, en la Figura 10-10 se representa la función anterior *f(x)=1/x* en el intervalo de variación de *x* e *y* dado por [-5, 5] para ambas coordenadas.

© Alfaomega - RC Libros

```
> plot(x,1/x, xlim=c(-5,5), ylim=c(-5,5))
```

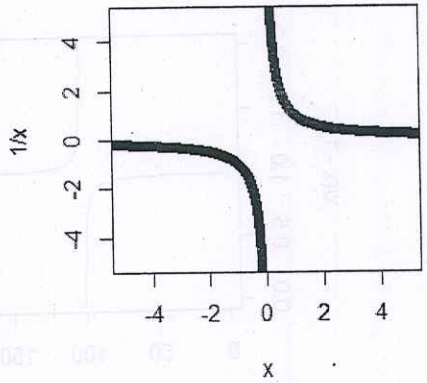


Figura 10-10

La función o comando *curve* de R permite representar gráficamente curvas planas en coordenadas explícitas. Las sintaxis más habituales de *curve* para representar curvas en coordenadas explícitas son las siguientes:

curve(f(x), a,b)	*Representa la función f(x) para x en el intervalo [a,b]*
curve(f(x), xlim = c(a,b))	*Representa la función f(x) para x variando entre a y b*
curve(f(x), ylim = c(d,e))	*Representa la función f(x) en el intervalo de variación de x previamente definido y para y variando entre d y e*
curve(f(x), xlim=c(a,b), ylim=c(d,e))	*Representa la función f(x) para x variando entre a y b e y variando entre d y e*
curve(f(x), xlim=range(c(a,b)), ylim=range(c(d,e)))	*Representa la función f(x) para x variando entre a y b e y variando entre d y e*

Como primer ejemplo, en la Figura 10-11 se representa la función $f(x)=x^3-3x$ en el intervalo de variación de *x* definido por [-π, π].

© Alfaomega - RC Libros


```
> curve(x^3-3*x,-pi,pi)
```

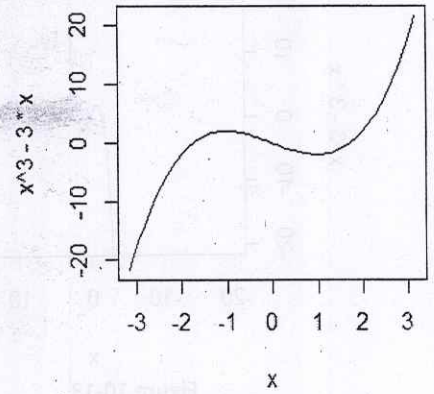


Figura 10-11

Como segundo ejemplo, en la Figura 10-12 se representa la función $f(x)=x^3-3x$ en el intervalo de variación de *x* definido por [-20, 20].

```
> curve(x^3-3*x,xlim=c(-20,20))
```

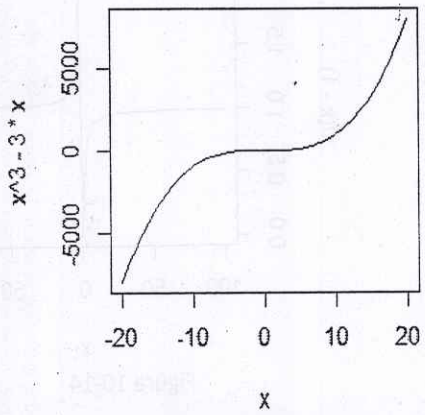


Figura 10-12

Como tercer ejemplo, en la Figura 10-13 se representa la función $f(x)=x^3-3x$ en el intervalo de variación de *x* y de *y* definido por [-20, 20] para ambas variables.

© Alfaomega - RC Libros

```
> curve(x^3-3*x,xlim=c(-20,20),ylim=c(-20,20))
```

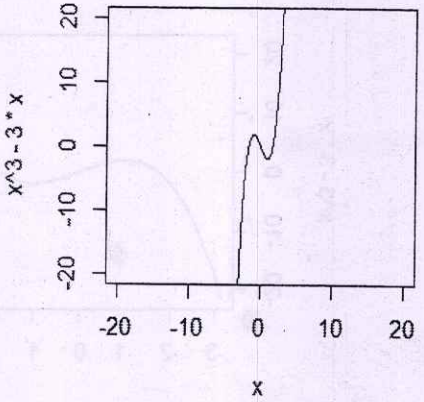


Figura 10-13

Como cuarto ejemplo, en la Figura 10-14 se representa la función *f(x)=x/(1-x)* en el intervalo de variación de *x* definido por [-100, 100].

```
> curve(x/(x-1), -100, 100)
```

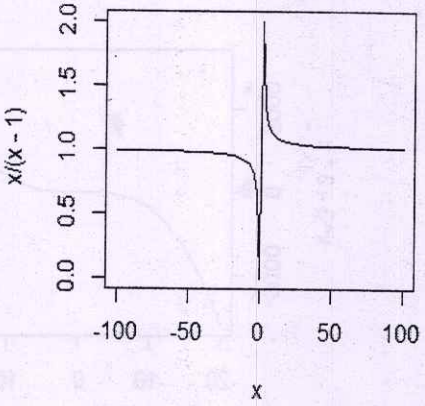


Figura 10-14

La misma gráfica (Figura 10-15) también puede obtenerse mediante la siguiente sintaxis:

© Alfaomega - RC Libros


```
> x=(-100:100)
> plot(x/(x-1), type="l")
```

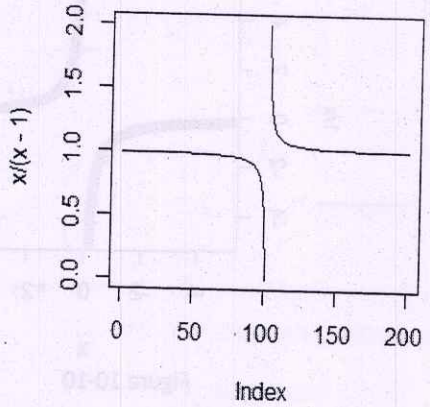


Figura 10-15

CURVAS PLANAS EN COORDENADAS POLARES

En el plano de ejes *xy* con centro de coordenadas en el punto O se puede definir un sistema de coordenadas polares de un punto **M** del plano, definidas por la distancia **r** al centro de coordenadas, y el ángulo φ del vector de posición sobre el eje x. Según la Figura 10-16, la curva en polares de ecuación **r=r(φ)** puede parametrizarse mediante:

x=r cos(φ)
y=r sen(φ)

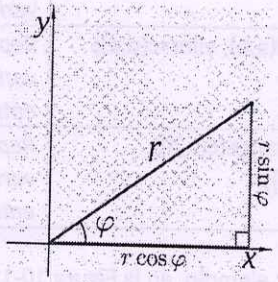


Figura 10-16

© Alfaomega - RC Libros

El comando *plot* permite representar curvas planas en coordenadas polares. Sencillamente basta con parametrizar previamente la ecuación de la curva. Como ejemplo representamos la curva en coordenadas polares *r=sen(2t)cos(2t)* t∈ [0,2π]:

```
> t=seq(0,2*pi, by=0.1)
> r=sin(2*t)*cos(2*t)
> x=r*cos(t)
> y=r*sin(t)
> plot(x,y,type="l")
```

La Figura 10-17 muestra la representación de la curva.

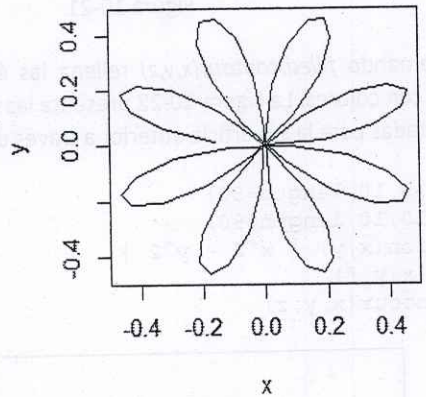


Figura 10-17

GRÁFICOS DE SUPERFICIES

En R es posible representar superficies definidas por su ecuación: *z=f(x,y)*. Se utiliza el comando *persp* con la siguiente sintaxis genérica:

x=campo de definición de la variable x
y=campo de definición de la variable y
z=outer(x,y,f)
persp(x,y,z)

Como ejemplo vamos a representar la superficie $z(x,y)=x^2-y^2$

La sintaxis de R adecuada sería la siguiente:

© Alfaomega - RC Libros


```
> x = seq(-10,10,length=50)
> y = seq(-10,10,length=50)
> f = function(x,y) { x^2 - y^2 }
> z = outer(x,y,f)
> persp(x,y,z)
```

Se obtiene la representación de la Figura 10-18.

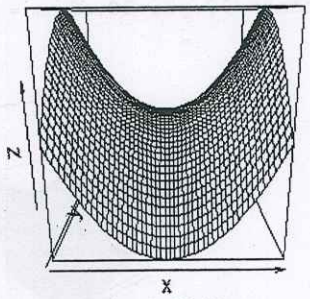


Figura 10-18

El comando *persp* admite como parámetros adicionales las coordenadas del punto de vista desde el cual se quiere observar la gráfica. En la Figura 10-19 se muestra la superficie observada desde el punto de vista (30º,30º).

```
> x = seq(-10,10,length=50)
> y = seq(-10,10,length=50)
> f = function(x,y) { x^2 - y^2 }
> z = outer(x,y,f)
> persp(x,y,z,theta=30,phi=30)
```

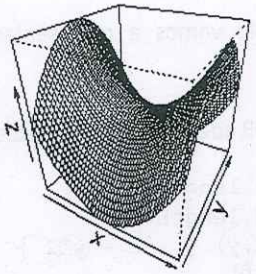


Figura 10-19

© Alfaomega - RC Libros

El comando *persp* también admite determinados parámetros adicionales relativos al sombreado de la superficie (*shade*), al establecimiento de los límites de variación de las variables (*xlim, ylim* y *zlim*), al establecimiento de marcas en los ejes (*axes*=TRUE) y una caja para el gráfico (*box*=TRUE), así como otros parámetros típicos de todos los gráficos.

En el ejemplo siguiente se representa la superficie anterior sombreada y con los límites en el interior de la sintaxis del comando *persp* (Figura 10-20).

```
> persp(x,y,z,theta=30,phi=30,xlim = c(-10,10), ylim=c(-10,10),
shade=0.9)
```

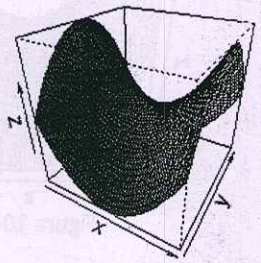


Figura 10-20

En R es posible representar contornos (curvas de nivel) de superficies definidas por su ecuación: *z=f(x,y)*. Para esta finalidad se utiliza el comando *contour* con la siguiente sintaxis genérica:

x=campo de definición de la variable x
y=campo de definición de la variable y
z=outer(x,y,f)
contour(x,y,z)

Como ejemplo vamos a representar las curvas de nivel de la superficie $z(x,y)=x^2-y^2$

La sintaxis de R adecuada sería la siguiente:

```
> x = seq(-10,10,length=50)
> y = seq(-10,10,length=50)
> f = function(x,y) { x^2 - y^2 }
> z = outer(x,y,f)
> contour(x,y,z)
```


© Alfaomega - RC Libros

Se obtiene la representación de la Figura 10-21.

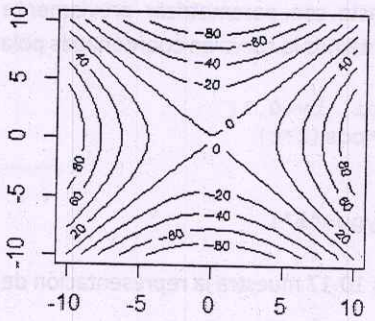


Figura 10-21

El comando *filled.contour(x,y,z)* rellena las áreas de los contornos de las curvas de nivel con colores. La Figura 10-22 presenta las curvas de nivel con densidades de color etiquetadas para la superficie anterior a través de la siguiente sintaxis:

```
> x = seq(-10,10,length=50)
> y = seq(-10,10,length=50)
> f = function(x,y) { x^2 - y^2 }
> z = outer(x,y,f)
> filled.contour(x,y,z)
```

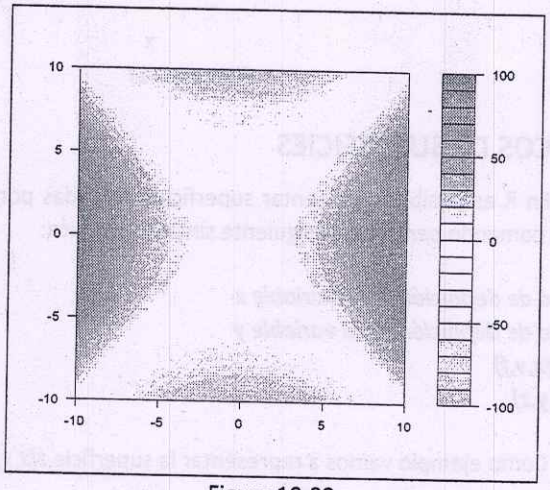


Figura 10-22

© Alfaomega - RC Libros

El comando *image(x,y,z)* rellena las áreas de los contornos de las curvas de nivel con colores, pero no los etiqueta. La Figura 10-23 presenta las curvas de nivel con colores para la superficie anterior a través de la siguiente sintaxis:

```
> x = seq(-10,10,length=50)
> y = seq(-10,10,length=50)
> f = function(x,y) { x^2 - y^2 }
> z = outer(x,y,f)
> image(x,y,z)
```

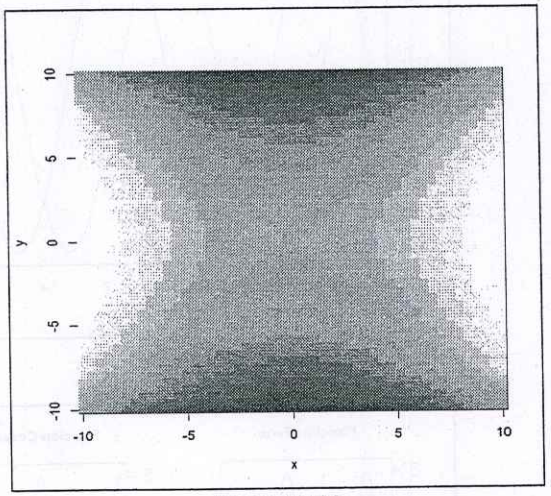


Figura 10-23

En la tabla siguiente se resumen los comandos esenciales relativos a representación de superficies en el lenguaje R.

persp(x,y,z)	*Representa la superficie definida por su ecuación: z=f(x,y)*
contour(x,y,z)	*Representa las curvas de nivel de la superficie definida por su ecuación: z=f(x,y)*
filled.contour(x,y,z)	*Rellena las áreas de los contornos de las curvas de nivel con densidades de colores etiquetados para la superficie z=f(x,y)*
image(x,y,z)	*Rellena las áreas de los contornos de las curvas de nivel con colores sin etiquetar para la superficie z=f(x,y)*

© Alfaomega - RC Libros

VENTANAS GRÁFICAS EN R

En R se puede trabajar con ventanas gráficas que permiten obtener las representaciones con mayor resolución. Los comandos más importantes relativos a ventanas gráficas son los siguientes:

windows( )	*Abre una ventana gráfica de alta resolución al trabajar con el sistema operativo Windows*
x11( )	*Abre una ventana gráfica de alta resolución al trabajar con los sistemas operativos Linux o UNIX*
quartz( )	*Abre una ventana gráfica de alta resolución al trabajar con el sistema operativo Mac OSX*
dev.off( )	*Desactiva la ventana gráfica de alta resolución actual*

Como ejemplo representamos la función $x^2/(x^2-1)$ en una ventana de alta resolución sobre Windows (Figura 10-24). La sintaxis sería la siguiente:

```
> windows()
> curve(x^2/(x^2-1),xlim=c(-5,5),ylim=c(-5,5))
```

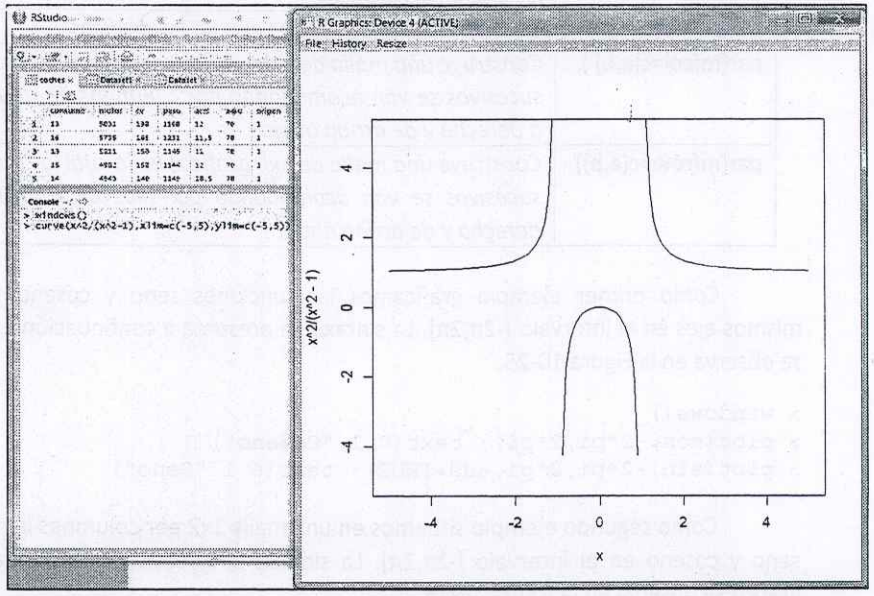


Figura 10-24

© Alfaomega - RC Libros

Los siguientes comandos permiten gobernar la situación de los diferentes gráficos sobre la ventana gráfica de salida.

windows( )	*Abre una ventana gráfica de alta resolución al trabajar con el sistema operativo Windows*
x11( )	*Abre una ventana gráfica de alta resolución al trabajar con los sistemas operativos Linux o UNIX*
quartz( )	*Abre una ventana gráfica de alta resolución al trabajar con el sistema operativo Mac OSX*
dev.off( )	*Desactiva la ventana gráfica de alta resolución actual*

Los siguientes comandos permiten situar los gráficos sobre la ventana gráfica a medida.

add=TRUE	*Opción de la mayoría de los comandos gráficos que permite la superposición de gráficos consecutivos en los mismos ejes*
add=FALSE	*No permite la superposición de gráficos consecutivos en los mismos ejes*
axes=TRUE	*Opción de la mayoría de los comandos gráficos que dibuja los ejes en los gráficos*
axes=FALSE	*No permite el dibujo de los ejes en los gráficos*
par(mfcol=c(a,b))	*Construye una malla de axb gráficos en la cual los gráficos sucesivos se van acomodando por columnas de izquierda a derecha y de arriba abajo*
par(mfrow=c(a,b))	*Construye una malla de axb gráficos en la cual los gráficos sucesivos se van acomodando por filas de izquierda a derecha y de arriba abajo*

Como primer ejemplo graficamos las funciones seno y coseno sobre los mismos ejes en el intervalo [-2π,2π]. La sintaxis se presenta a continuación y el gráfico se observa en la Figura 10-25.

```
> windows()
> plot(cos,-2*pi,2*pi); text(0,1,"Coseno")
> plot(sin,-2*pi,2*pi,add=TRUE); text(6,1,"Seno")
```

Como segundo ejemplo situamos en una malla 1x2 por columnas las funciones seno y coseno en el intervalo [-2π,2π]. La sintaxis se presenta a continuación y el gráfico se observa en la Figura 10-26.

© Alfaomega - RC Libros


```
> windows()
> par(mfrow=c(1,2))
> plot(sin,-2*pi,2*pi,main="Función Seno")
> plot(cos,-2*pi,2*pi,main="Función Coseno")
```

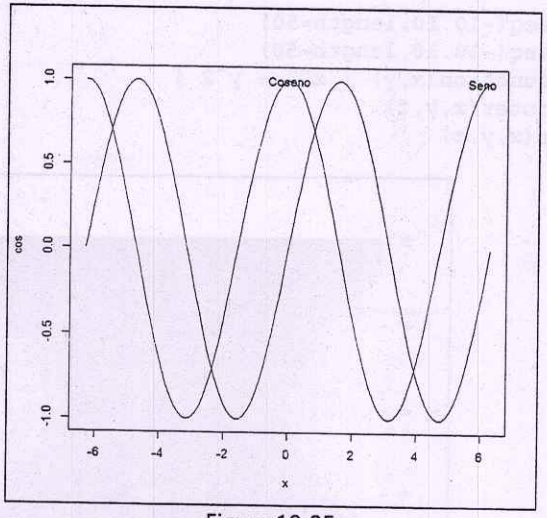


Figura 10-25

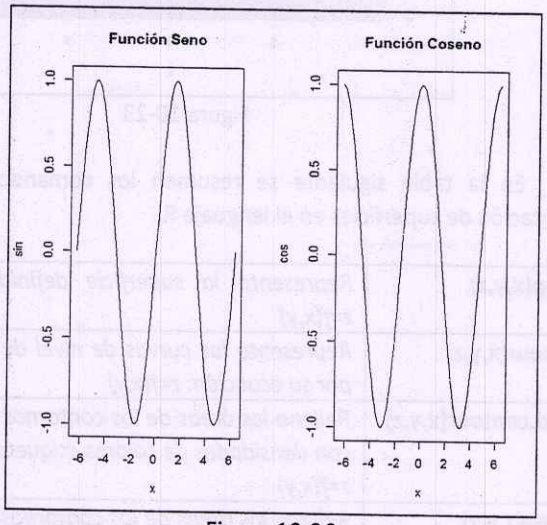


Figura 10-26

© Alfaomega - RC Libros

Ejercicio 10-1. Represente las curvas en coordenadas explícitas de las ecuaciones:

$$x\ sen(1/x) \quad x\varepsilon[0,\pi/6]$$
$$x^2/(x+1) \quad x,y\varepsilon[-12,12]x[-8,8]$$

Para representar la primera curva (Figura 10-27) utilizamos la sintaxis siguiente:

```
> x=seq(0.001,pi/6,by=1/1000)
> plot(x,x*sin(1/x), type="l")
```

Alternativamente podríamos haber utilizado la sintaxis siguiente:

```
> x=seq(0.001,pi/6,by=1/1000)
> curve(x*sin(1/x),0.0001,pi/6)
```

Otra sintaxis alternativa sería la siguiente:

```
> x=seq(0.001,pi/6,by=1/1000)
> plot(x*sin(1/x), type="l")
```

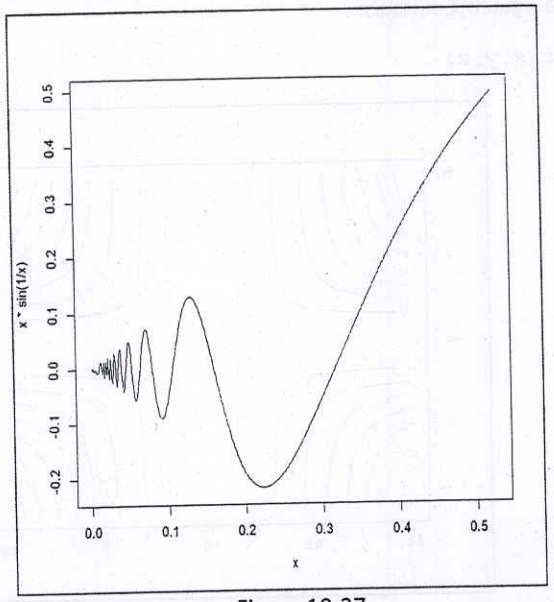


Figura 10-27

© Alfaomega - RC Libros

La segunda curva (Figura 10-28) se representa mediante la sintaxis siguiente:

```
> curve(x^2/(x+1), xlim=c(-12,12),ylim=c(-8,8))
```

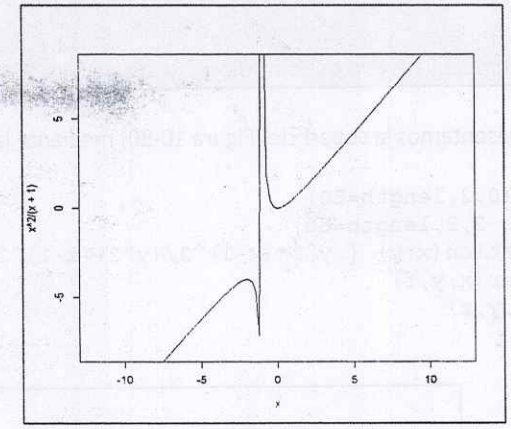


Figura 10-28

Alternativamente podríamos haber representado la curva (Figura 10-29) utilizando la sintaxis siguiente:

```
> x=seq(-12,12,by=1/1000)
> plot(x^2/(x+1),ylim=c(-8,8))
```

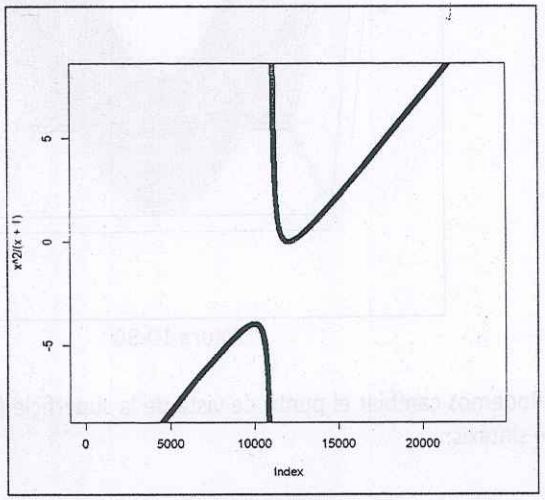


Figura 10-29

© Alfaomega - RC Libros

Ejercicio 10-2. Represente la superficie definida por la ecuación:

$$f(x,y)=\frac{(x-1)^2 y^2}{(x-1)^2+y^2}$$

Represente también sus curvas de nivel.

Representamos la superficie (Figura 10-30) mediante la siguiente sintaxis:

```
> x = seq(0,2,length=50)
> y = seq(-2,2,length=50)
> f = function(x,y) { y^2*(x-1)^2/(y^2+(x-1)^2})
> z = outer(x,y,f)
> persp(x,y,z)
```

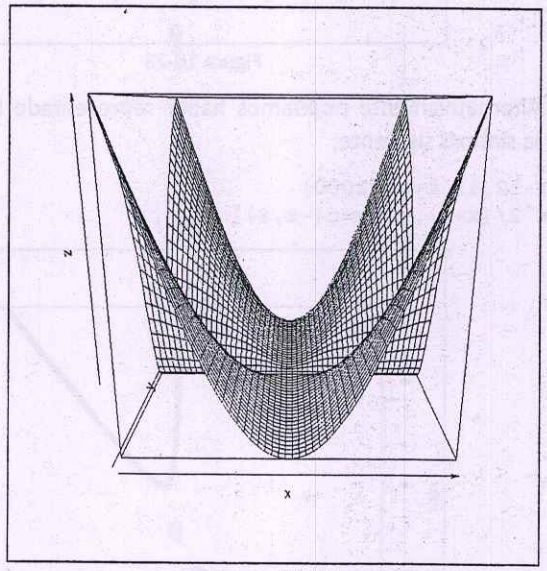


Figura 10-30

Podemos cambiar el punto de vista de la superficie (Figura 10-31) mediante la siguiente sintaxis:

© Alfaomega - RC Libros


```
> persp(x,y,z,theta=30,phi=30)
```

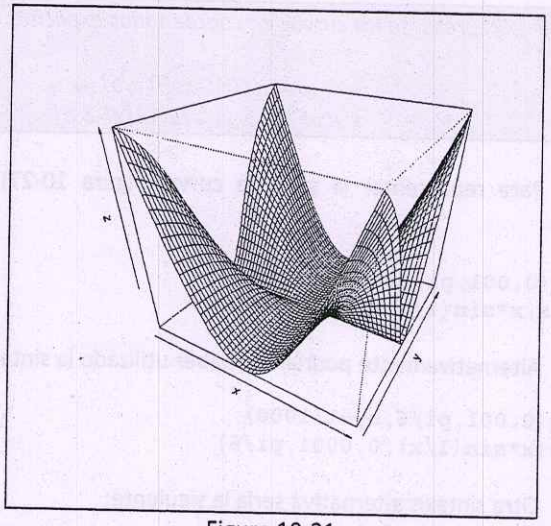


Figura 10-31

Podemos representar las curvas de nivel de la superficie (Figura 10-32) mediante la siguiente sintaxis:

```
> contour(x,y,z)
```

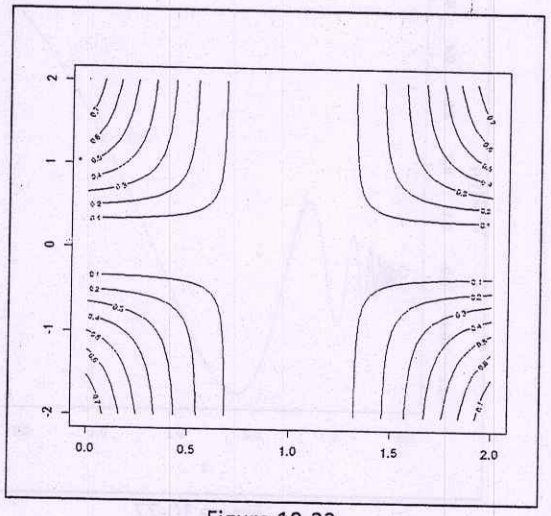


Figura 10-32

© Alfaomega - RC Libros

Ejercicio 10-3. Sea $f:R^2 \rightarrow R$ una función definida por:

$$f(x,y)=\frac{[1-\cos(x)]Sen(y)}{x^3+y^3}$$

Represéntela gráficamente en un entorno de (0,0).

La sintaxis adecuada para obtener la Figura 10-33 podría ser la siguiente:

```
> x = seq(-1/100,1/100,by=0.0009)
> y = seq(-1/100,1/100,by=0.0009)
> f = function(x,y) {(1-cos(x))*sin(y)/(x^3+y^3)}
> z = outer(x,y,f)
> persp(x,y,z,theta=-15,phi=50)
```

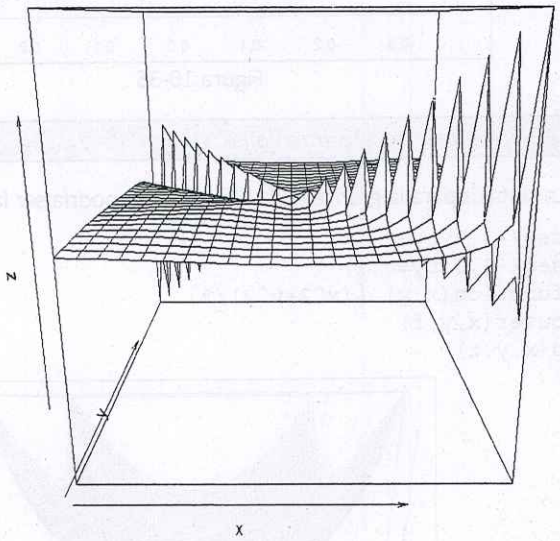


Figura 10-33

Ejercicio 10-4. Represente sobre el mismo gráfico y de forma adyacente las curvas en polares:

$$r=\sqrt{Cos(2a)} \quad y \quad r=Sen(2a)$$

© Alfaomega - RC Libros

La sintaxis para representar las dos curvas de forma adyacente (Figura 10-34) podría ser la siguiente:

```
> windows()
> par(mfrow=c(1,2))
> t=seq(0,2*pi, by=0.001)
> r=(cos(2*t))^0.5
> x=r*cos(t)
> y=r*sin(t)
> plot(x,y,type="l",main="r=sqrt(Cos(2t))")
> t=seq(0,2*pi, by=0.001)
> r=sin(2*t)
> x=r*cos(t)
> y=r*sin(t)
> plot(x,y,type="l",main="r=sen(2t)")
```

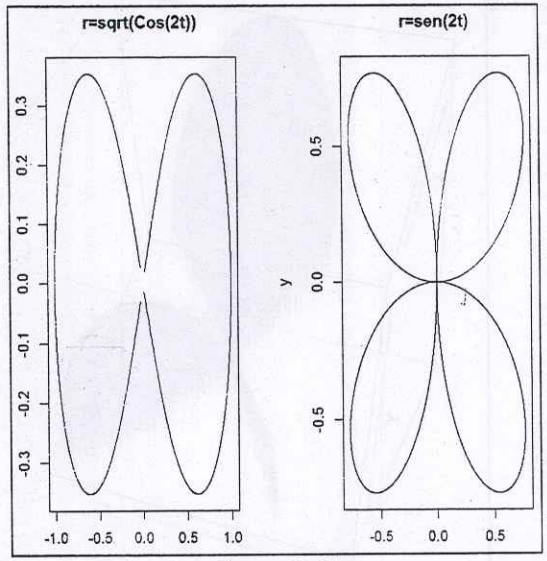


Figura 10-34

Ejercicio 10-5. Represente el volumen engendrado al girar la cúbica $12x - 9x^2 + 2x^3$ alrededor del eje OX, comprendida entre x=-1/3 y x=1/3

La superficie de revolución tendrá de ecuación $y^2+z^2=(12x-9x^2+2x^3)^2$, y para graficarla (Figura 10-35) puede utilizarse la siguiente sintaxis:

© Alfaomega - RC Libros

```
> x= seq(-1/3,1/3,by=0.01)
> y = seq(-2.5,2.5,by=0.01)
> f = function(x,y) {(y^2-(12*x-9*x^2+2*x^3)^2)^(1/2)}
> z = outer(x,y,f)
> persp(x,y,z,theta=-15,phi=50)
```

Para representar las curvas de nivel (Figura 10-36) se utiliza la siguiente sintaxis:

```
> x = seq(-1/3,1/3,by=0.01)
> y = seq(-2.5,2.5,by=0.01)
> f = function(x,y) {(y^2-(12*x-9*x^2+2*x^3)^2)^(1/2)}
> z = outer(x,y,f)
> contour(x,y,z)
```

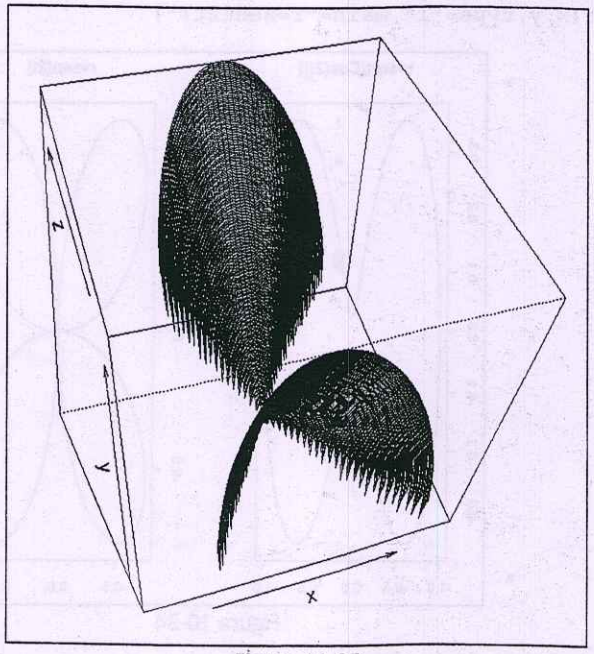


Figura 10-35

© Alfaomega - RC Libros

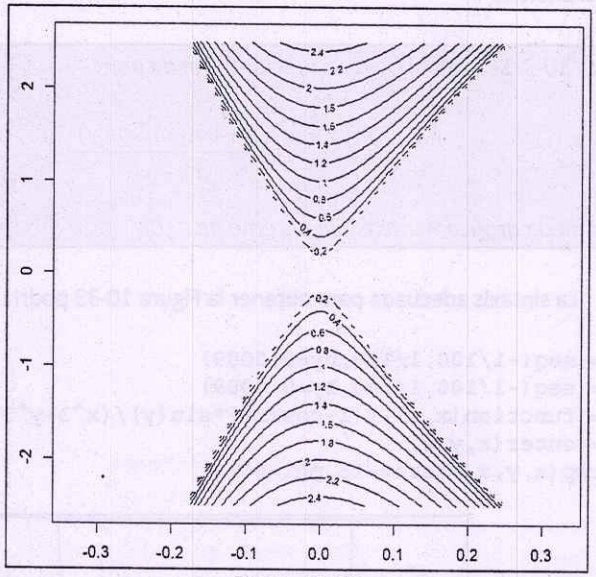


Figura 10-36

Ejercicio 10-6. Represente el paraboloide $x^2 + y^2 = 2z$.

La sintaxis para la gráfica de la Figura 10-37 podría ser la siguiente:

```
> x = seq(-3,3,by=0.1)
> y = seq(-3,3,by=0.1)
> f = function(x,y) {(x^2+y^2)/2}
> z = outer(x,y,f)
> persp(x,y,z)
```

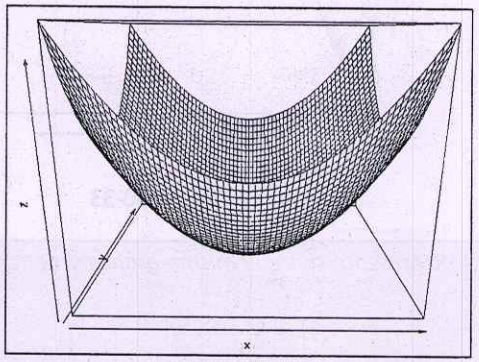


Figura 10-37

© Alfaomega - RC Libros

Ejercicio 10-7. Represente sobre los mismos ejes las parábolas $y^2 = x$ y $x^2 = y$. Represente también sobre los mismos ejes la parábola $y^2=4x$ y la recta $x+y=3$.

La primera gráfica (Figura 10-38) se obtiene con la siguiente sintaxis:

```
> curve(x^2,0,3/2)
> curve(x^(1/2),0,3/2,add=TRUE)
```

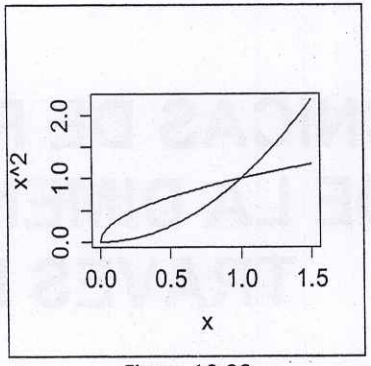


Figura 10-38

La segunda gráfica (Figura 10-39) se obtiene con la siguiente sintaxis:

```
> curve(3-x,xlim=c(0,4),ylim=c(0,4))
> curve((4*x)^(1/2),xlim=c(0,4),ylim=c(0,4),add=TRUE)
```

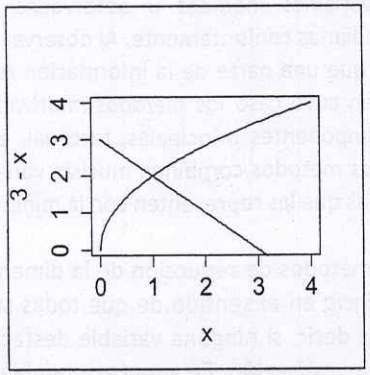


Figura 10-39

© Alfaomega - RC Libros

Ejercicio 10-8. Represente la curva en coordenadas paramétricas:

$$x(t) = tSen(t)$$
$$y(t) = tCos(t)$$

La sintaxis para la gráfica de la Figura 10-40 podría ser la siguiente:

```
> t = seq(-3*pi/2,3*pi/2,by=0.01)
> x=t*sin(t)
> y=t*cos(t)
> plot(x,y,type='l')
```

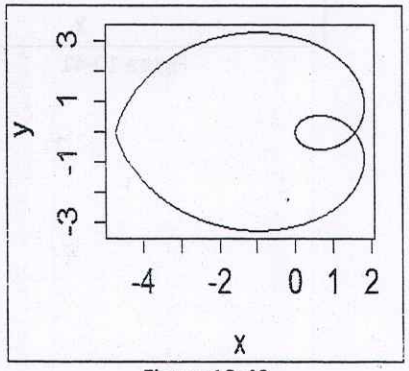


Figura 10-40

Ejercicio 10-9. Represente la curva en coordenadas polares:

$$\rho=1-cos(\theta)$$

La sintaxis para la gráfica de la Figura 10-41 podría ser la siguiente:

```
> t=seq(0,2*pi,by=0.01)
> r=1-cos(t)
> x=r*cos(t)
> y=r*sin(t)
> plot(x,y,type="l",main="r=1-cos(t)")
```

© Alfaomega - RC Libros

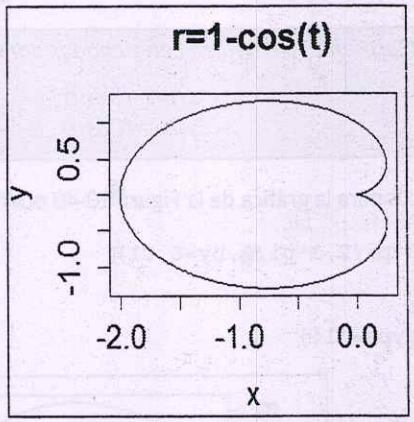


Figura 10-41

© Alfaomega - RC Libros

CAPÍTULO 11

TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN A TRAVÉS DE R

TRANSFORMACIÓN DE DATOS MEDIANTE TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN

En el mundo de la información de hoy en día es habitual disponer de gran cantidad de variables medidas u observadas en una colección de individuos y pretender estudiarlas conjuntamente. Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los *métodos multivariantes de reducción de la dimensión* (análisis en componentes principales, factorial, escalamiento óptimo, etc.) tratan de eliminarla. Estos métodos combinan muchas variables observadas para obtener pocas variables ficticias que las representen con la mínima pérdida de información.

Estos métodos de reducción de la dimensión son *métodos multivariantes de la interdepedencia* en el sentido de que todas sus variables tienen una importancia equivalente, es decir, si ninguna variable destaca como dependiente principal en el objetivo de la investigación. En este caso también deberá tener en cuenta el tipo de variables que se maneja. Si son variables cuantitativas, las técnicas de reducción de la

© Alfaomega - RC Libros

son variables cualitativas, puede acudirse al *Análisis de Correspondencias* y al *Escalamiento Óptimo*.

Los métodos de la interdependencia se contraponen a los denominados *métodos multivariantes de la dependencia,* en los cuales no es aceptable una importancia equivalente en las variables, porque alguna se destaca como dependiente principal. En este caso habrá de utilizar técnicas multivariantes analíticas o inferenciales considerando la variable dependiente como explicada por las demás variables independientes explicativas, y tratando de relacionar todas las variables por medio de una posible ecuación o modelo que las ligue. El método elegido podría ser entonces la Regresión Lineal, generalmente con todas las variables cuantitativas. Una vez configurado el modelo matemático se podrá llegar a predecir el valor de la variable dependiente conocido el perfil de todas las demás. Si la variable dependiente fuera cualitativa dicotómica (1,0; sí o no) podrá usarse como clasificadora, estudiando su relación con el resto de variables clasificativas a través de la Regresión Logística. Si la variable dependiente cualitativa observada constatara la asignación de cada individuo a grupos previamente definidos (dos, o más de dos), puede ser utilizada para clasificar nuevos casos en que se desconozca el grupo a que probablemente pertenecen, en cuyo caso estamos ante el Análisis Discriminante, que resuelve el problema de asignación en función de un perfil cuantitativo de variables clasificativas. Si la variable dependiente es cuantitativa y las explicativas son cualitativas estamos ante los modelos del análisis de la varianza, que puede extenderse a los modelos loglineales para el análisis de tablas de contingencia de dimensión elevada. Si la variable dependiente puede ser cualitativa o cuantitativa y las independientes cualitativas, estamos ante la Segmentación.

Por otra parte, las técnicas de reducción de la dimensión juegan un papel muy importante dentro de las *técnicas emergentes de análisis multivariante* de datos. La disponibilidad de grandes volúmenes de datos y el uso generalizado de herramientas informáticas han transformado el análisis multivariante orientándolo hacia determinadas técnicas especializadas englobadas en la *minería de datos*.

COMPONENTES PRINCIPALES

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de interdependencia. Se trata de un método multivariante de simplificación o reducción de la dimensión y que se aplica

© Alfaomega - RC Libros

cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionadas entre sí persiguiendo obtener un menor número de variables, combinación lineal de las primitivas e incorrelacionadas, que se denominan componentes principales o factores, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información y cuya posterior interpretación permitirá un análisis más simple del problema estudiado.

Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, *clusters*, etc.). El elevado número de variables iniciales $x_1, x_2,..., x_p$ se resumen en unas pocas variables $C_1, C_2,..., C_k$ (*componentes principales*) *perfectamente calculables* y que sintetizan la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$
$$\vdots$$
$$C_p = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{pp}x_p$$

Pero solo se retienen las k componentes principales ($k \leq p$) que explican un porcentaje alto de la variabilidad de las variables iniciales ($C_1, C_2,..., C_k$).

Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquella que tenga mayor varianza, mientras que, por el contrario, la última es la de menor varianza.

En general, la extracción de componentes principales se efectúa sobre variables *tipificadas* para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en *desviaciones* respecto a la media.

Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. Si las variables originales estuvieran completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

© Alfaomega - RC Libros

Cálculo de las componentes principales

En el análisis en componentes principales se dispone de una muestra de tamaño n acerca de p variables $X_1, X_2,..., X_p$ (tipificadas o expresadas en desviaciones respecto de su media) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k \leq p$ de variables incorrelacionadas $C_1, C_2,..., C_k$ que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad.

La primera componente principal, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$C_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \cdots + u_{1p}X_{pi} \quad i = 1,.., n$$

Para el conjunto de las n observaciones muestrales y para todas las componentes tenemos:

$$\begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} \cdots & X_{p1} \\ X_{12} & X_{22} \cdots & X_{p2} \\ & \vdots & \\ X_{1n} & X_{2n} \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tendremos: $C_1 = X\,u_1$ y:

$$V(C_1) = \frac{\sum_{i=1}^{n} C_{1i}^2}{n} = \frac{1}{n} C_1' C_1 = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X'X \right] u_1 = u_1' V u_1$$

La primera componente C_1 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones $(u_{11}, u_{12},..., u_{1p})'$ se toma normalizada. Se trata entonces de hallar C_1 maximizando $V(C_1) = u_1'Vu_1$, sujeta a la restricción:

$$\sum_{j=1}^{p} u_{1i}^2 = u_1' u_1 = 1$$

© Alfaomega - RC Libros

Se demuestra que, para maximizar $V(C_1)$ se toma el mayor valor propio λ de la matriz V. Sea λ_1 el citado mayor valor propio de V y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera componente principal, componente que vendrá definida como:

$$C_1 = u_1 X = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p$$

Para maximizar $V(C_2)$ hemos de tomar el segundo mayor valor propio λ de la matriz V (el mayor ya lo había tomado al obtener la primera componente principal).

Tomando λ_2 como el segundo mayor valor propio de V y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2$ = 1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la segunda componente principal, componente que vendrá definida como:

$$C_2 = u_2 X = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p$$

De forma similar, la componente principal h-ésima se define como $C_h = Xu_h$ donde u_h es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h eje factorial h-ésimo.

Se demuestra que la proporción de la variabilidad total recogida por la componente principal h-ésima (*porcentaje de inercia explicada por la componente principal h-ésima*) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p} \lambda_h} = \frac{\lambda_h}{traza(V)}$$

Si las variables están tipificadas, $traza(V) = p$, con lo que la proporción de la componente h-ésima en la variabilidad total será λ_h/p. También se define el *porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales)* como:

$$\frac{\sum_{h=1}^{k} \lambda_h}{\sum_{h=1}^{p} \lambda_h} = \frac{\sum_{h=1}^{k} \lambda_h}{traza(V)}$$

© Alfaomega - RC Libros

Puntuaciones o medición de las componentes

El análisis en componentes principales es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por las componentes obtenidas. Por ejemplo, en el caso de estimación de modelos afectados de multicolinealidad o correlación serial (autocorrelación). Por ello, es necesario conocer los valores que toman las componentes en cada observación. Una vez calculados los coeficientes u_{hj} (componentes del vector propio normalizado asociado al valor propio h-ésimo de la matriz $V = X'X/n$ relativo a la componente principal Z_h), se pueden obtener las puntuaciones Z_{hj}, es decir, los valores de las componentes correspondientes a cada observación, a partir de la siguiente relación:

$$Z_{hi} = u_{h1}X_{1i} + u_{h2}X_{2i} + \cdots + u_{hp}X_{pi} \quad h = 1,\ldots,p \quad i = 1,\ldots,n$$

Número de componentes principales a retener

En general, el objetivo de la aplicación de las componentes principales es reducir las dimensiones de las variables originales, pasando de p variables originales a $k < p$ componentes principales.

El problema que se plantea es cómo fijar k, o, dicho de otra forma, ¿qué número de componentes se deben retener? Aunque para la extracción de las componentes principales no hace falta plantear un modelo estadístico previo, algunos de los criterios para determinar cuál debe ser el número óptimo de componentes a retener requieren la formulación previa de hipótesis estadísticas.

- ***Criterio de la media aritmética***

Según este criterio se seleccionan aquellas componentes cuya raíz característica λ_j excede de la media de las raíces características. Recordemos que la raíz característica asociada a una componente es precisamente su varianza. Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_h > \overline{\lambda} = \frac{\sum_{j=1}^{p} \lambda_h}{p}$$

© Alfaomega - RC Libros

Si se utilizan variables tipificadas, entonces, como ya se ha visto, se verifica que $\sum_{j=1}^{p} \lambda_h = p$, con lo que para variables tipificadas se retienen aquellas componentes tales que $\lambda_h > 1$.

- ***Criterio del gráfico de sedimentación***

El *gráfico de sedimentación* se obtiene al representar en ordenadas las raíces características y en abscisas los números de las componentes principales correspondientes a cada raíz característica en orden decreciente. Uniendo todos los puntos se obtiene una figura que, en general, se parece al perfil de una montaña con una pendiente fuerte hasta llegar a la base, formada por una meseta con una ligera inclinación. Continuando con el símil de la montaña, en esa meseta es donde se acumulan los guijarros caídos desde la cumbre, es decir, donde se sedimentan.

Por esta razón, a este gráfico se le conoce con el nombre de gráfico de sedimentación. Su denominación en inglés es *scree plot*. De acuerdo con el criterio gráfico se retienen todas aquellas componentes previas a la zona de sedimentación.

Matriz de cargas factoriales, comunalidad y círculos de correlación

La dificultad en la interpretación de los componentes estriba en la necesidad de que tengan sentido y midan algo útil en el contexto del fenómeno estudiado. Por tanto, es indispensable considerar el peso que cada variable original tiene dentro del componente elegido, así como las correlaciones existentes entre variables y factores.

Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionado con algunas de ellas, y menos con otras. Ya hemos visto que el coeficiente de correlación entre una componente y una variable se calcula multiplicando el peso de la variable en esa componente por la raíz cuadrada de su valor propio:

$$r_{jh} = u_{hj}\sqrt{\lambda_h}$$

Se demuestra también que estos coeficientes r representan la parte de varianza de cada variable que explica cada factor. De este modo, cada variable puede ser representada como una función lineal de los k componentes retenidos, donde los

© Alfaomega - RC Libros

pesos o cargas de cada componente o factor (*cargas factoriales*) en la variable coinciden con los coeficientes de correlación.

El cálculo matricial permite obtener de forma inmediata la tabla de coeficientes de correlación variables-componentes (*pxk*), que se denomina *matriz de cargas factoriales*. Las ecuaciones de las variables en función de las componentes (factores), traspuestas las inicialmente planteadas, son de mayor utilidad en la interpretación de los componentes, y se expresan como sigue:

$$\begin{aligned} Z_1 &= r_{11}X_1 + \cdots + r_{1p}X_p \\ Z_2 &= r_{21}X_1 + \cdots + r_{2p}X_p \\ &\vdots \\ Z_k &= r_{k1}X_1 + \cdots + r_{kp}X_p \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= r_{11}Z_1 + \cdots + r_{k1}Z_k \\ X_2 &= r_{12}Z_1 + \cdots + r_{k2}Z_k \\ &\vdots \\ X_p &= r_{1p}Z_1 + \cdots + r_{kp}Z_k \end{aligned}$$

Para la primera variable, la comunalidad será $r^2_{11} + \ldots + r^2_{k1} = V(X_1) = h_1^2$. Por consiguiente, la suma de las comunalidades de todas las variables representa la parte de inercia global de la nube original explicada por los *k* factores retenidos, y coincide con la suma de los valores propios de estas componentes. La comunalidad proporciona un criterio de calidad de la representación de cada variable, de modo que, variables totalmente representadas tienen de comunalidad la unidad.

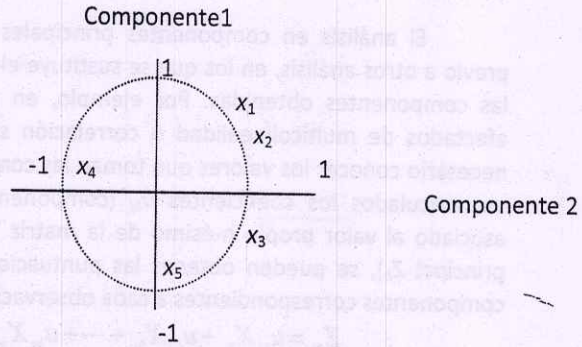
También se demuestra que la suma en vertical de los cuadrados de las cargas factoriales de todas las variables en un componente es su valor propio. Por ejemplo, el valor propio del primer componente será $r^2_{11} + \ldots + r^2_{1p} = \lambda_1$.

Es evidente que, al ser las cargas factoriales los coeficientes de correlación entre variables y componentes, su empleo hace comparables los pesos de cada variable en la componente y facilita su interpretación.

En este mismo sentido, su representación gráfica puede orientar al investigador en una primera aproximación a la interpretación de los componentes. Como es lógico, esta representación sobre un plano solo puede contener los factores de dos en dos, por lo que se pueden realizar tantos gráficos como parejas de factores retenidos. Estos gráficos se denominan *círculos de correlación*, y están formados por puntos que representan cada variable por medio de dos coordenadas que miden los coeficientes de correlación de dicha variable con los dos factores o componentes

© Alfaomega - RC Libros

consideradas. Todas las variables estarán contenidas dentro del círculo de radio unidad siguiente:



Rotación de las componentes

Es frecuente no encontrar interpretaciones verosímiles a los factores (componentes) obtenidos. Sería deseable, para una más fácil interpretación, que cada componente estuviera relacionada muy bien con pocas variables (coeficientes de correlación *r* próximos a 1 o -1) y mal con las demás (*r* próximos a 0).

Esta optimización se obtiene por una adecuada *rotación de los ejes* que definen los componentes principales. Rotar un conjunto de componentes no cambia la proporción de inercia total explicada, como tampoco cambia las comunalidades de cada variable, que no son sino la proporción de varianza explicada por todos ellos. Las rotaciones más utilizadas son la rotación VARIMAX y la QUARTIMAX (ortogonales) y PROMAX (oblicua).

Sin embargo, los coeficientes, que dependen directamente de la posición de los componentes respecto a las variables originales (cargas factoriales y valores propios), se ven alterados por la rotación.

ANÁLISIS FACTORIAL

El ***análisis factorial*** tiene como objeto simplificar las múltiples y complejas relaciones que puedan existir entre un conjunto de variables observadas $X_1, X_2, \ldots, X_p$. Para ello trata de encontrar dimensiones comunes o *factores* que ligan a las aparentemente no relacionadas variables. Concretamente, se trata de encontrar un conjunto de $k < p$ *factores no directamente observables* $F_1, F_2, \ldots, F_k$ que expliquen

© Alfaomega - RC Libros

suficientemente a las variables observadas perdiendo el mínimo de información, de modo que sean fácilmente interpretables (*principio de interpretabilidad*) y que sean los menos posibles, es decir, *k* pequeño (*principio de parsimonia*). Además, los factores han de extraerse de forma que resulten independientes entre sí, es decir, que sean ortogonales. En consecuencia, el análisis factorial es una técnica de reducción de datos que examina la interdependencia de variables y proporciona conocimiento de la estructura subyacente de los datos.

El aspecto más característico del análisis factorial lo constituye su capacidad de reducción de datos. Las relaciones entre las variables observadas $X_1, X_2, \ldots, X_p$ vienen dadas por su matriz de correlaciones, cuyo determinante ha de ser pequeño (hay relación entre ellas).

El análisis de componentes principales y el análisis factorial tienen en común que son técnicas de reducción de la dimensión para examinar la interdependencia de variables, pero difieren en su objetivo, sus características y su grado de formalización. La diferencia entre análisis en componentes principales y análisis factorial radica en que en el análisis factorial trata de encontrar variables sintéticas latentes, inobservables y aún no medidas cuya existencia se sospecha en las variables originales y que permanecen a la espera de ser halladas, mientras que en el análisis en componentes principales se obtienen variables sintéticas combinación de las originales y cuyo cálculo es posible basándose en aspectos matemáticos independientes de su interpretabilidad práctica.

En el análisis en componentes principales la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (sus componentes). Pero esto no ocurre en el análisis factorial.

En el análisis factorial solo una parte de la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (*factores comunes* $F_1, F_2, \ldots, F_p$). Esta parte de la variabilidad de cada variable original explicada por los factores comunes se denomina *comunalidad*, mientras que la parte de varianza no explicada por los factores comunes se denomina *unicidad* (*comunalidad + unicidad* = 1) y representa la parte de variabilidad propia f_i de cada variable x_i.

$$x_1 = r_{11}F_1 + r_{12}x_2 + \cdots + r_{1p}F_p + f_1$$
$$\vdots$$
$$x_n = r_{n1}F_1 + r_{n2}F_2 + \cdots + r_{np}F_p + f_n$$

© Alfaomega - RC Libros

Cuando la comunalidad es unitaria (*unicidad nula*) el análisis en componentes principales coincide con el factorial. Es decir, el análisis en componentes principales es un caso particular del análisis factorial en el que los factores comunes explican el 100% de la varianza total.

Entre los métodos para obtener los factores destacan los siguientes:

- *Método de las componentes principales.*
- *Métodos MINRES (minimización residual), ULS (mínimos cuadrados no ponderados) y GLS (mínimos cuadrados generalizados).*
- *Método de máxima verosimilitud.*
- *Método de componentes principales iteradas o ejes principales.*
- *Método del factor principal.*
- *Método alfa.*
- *Método de factorización imagen.*
- *Método del centroide.*
- *Método de Turstone.*

A continuación se presentan las características de los métodos más importantes de extracción de los factores.

- *Método de las componentes principales.* Método de extracción de factores utilizado para formar combinaciones lineales no correlacionadas de las variables observadas. La primera componente tiene la varianza máxima. Las componentes sucesivas explican progresivamente proporciones menores de la varianza y no están correlacionadas las unas con las otras. El análisis de componentes principales se utiliza para obtener la solución factorial inicial. Puede utilizarse cuando una matriz de correlaciones es singular.
- *Método de mínimos cuadrados no ponderados.* Método de extracción factorial que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlaciones observada y reproducida, ignorando las diagonales.
- *Método de mínimos cuadrados generalizados.* Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida. Las correlaciones se ponderan por el inverso de su unicidad, de manera que las variables que tengan un valor alto de unicidad reciban un peso menor que las que tengan un valor bajo de unicidad.

© Alfaomega - RC Libros

- *Método de máxima verosimilitud*. Método de extracción factorial que proporciona las estimaciones de los parámetros que con mayor probabilidad han producido la matriz de correlaciones observada, si la muestra procede de una distribución normal multivariada. Las correlaciones se ponderan por el inverso de la unicidad de las variables y se emplea un algoritmo iterativo.

- *Factorización de ejes principales*. Método de extracción de factores que parte de la matriz de correlaciones original con los cuadrados de los coeficientes de correlación múltiple insertados en la diagonal principal como estimaciones iniciales de las comunalidades. Las saturaciones factoriales resultantes se utilizan para estimar de nuevo las comunalidades y reemplazan a las estimaciones previas en la diagonal de la matriz. Las iteraciones continúan hasta que los cambios en las comunalidades, de una iteración a la siguiente, satisfagan el criterio de convergencia para la extracción.

- *Alfa*. Método de extracción factorial que considera a las variables incluidas en el análisis como una muestra del universo de las variables posibles. Este método maximiza el Alfa de Cronbach para los factores.

- *Factorización imagen*. Método de extracción de factores, desarrollado por Guttman y basado en la teoría de las imágenes. La parte común de una variable, llamada la imagen parcial, se define como su regresión lineal sobre las restantes variables, en lugar de ser una función de los factores hipotéticos.

Contrastes en el modelo factorial

En el modelo factorial pueden realizarse varios tipos de contrastes. Estos contrastes suelen agruparse en dos bloques, según se apliquen previamente a la extracción de los factores o que se apliquen después. Con los contrastes aplicados previamente a la extracción de los factores trata de analizarse la pertinencia de la aplicación del análisis factorial a un conjunto de variables observables. Con los contrastes aplicados después de la obtención de los factores se pretende evaluar el modelo factorial una vez estimado.

Dentro del grupo de *contrastes que se aplican previamente a la extracción de los factores* tenemos el contraste de esfericidad de Barlett y la medida de adecuación muestral de Kaiser, Meyer y Olkin.

© Alfaomega - RC Libros

Evidentemente, antes de realizar un análisis factorial nos plantearemos si las p variables originales están correlacionadas entre sí o no lo están. Si no lo estuvieran no existirían factores comunes y, por lo tanto, no tendría sentido aplicar el análisis factorial. Esta cuestión suele probarse utilizando el contraste de esfericidad de Barlett que se basa en que la matriz de correlación poblacional R_p recoge la relación entre cada par de variables mediante sus elementos ρ_{ij} situados fuera de la diagonal principal.

Los elementos de la diagonal principal son unos, ya que toda variable está totalmente relacionada consigo misma. En caso de que no existiese ninguna relación entre las p variables en estudio, la matriz R_p sería la identidad, cuyo determinante es la unidad.

Por lo tanto, para decidir la ausencia o no de relación entre las p variables puede plantearse el siguiente contraste:

$$H_0 : |R_p| = 1$$
$$H_1 : |R_p| \neq 1$$

Barlett introdujo un estadístico para este contraste basado en la matriz de correlación muestral R, que bajo la hipótesis H_0 tiene una distribución *Chi-cuadrado* con $p(p-1)/2$ grados de libertad. La expresión de este estadístico es la siguiente:

$$-[n-11-(2p+5)/6]\, Ln|R|$$

Por otro lado, Kaiser-Meyer y Olkin definen la medida *KMO* de adecuación muestral global al modelo factorial basada en los coeficientes de correlación observados de cada par de variables y en sus coeficientes de correlación parcial mediante la expresión siguiente:

$$KMO = \frac{\sum_j \sum_{h \neq j} r_{jh}^2}{\sum_j \sum_{h \neq j} r_{jh}^2 + \sum_j \sum_{h \neq j} a_{jh}^2}$$

r_{jh} son los coeficientes de correlación observados entre las variables X_j y X_h
a_{jh} son los coeficientes de correlación parcial entre las variables X_j y X_h

© Alfaomega - RC Libros

- En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador, que recoge los coeficientes a_{jh}, será pequeño y, en consecuencia, la medida *KMO* será próxima a la unidad. Valores de *KMO* por debajo de 0,5 no serán aceptables, considerándose inadecuados los datos a un modelo de análisis factorial. Para valores superiores a 0,5 se considera aceptable la adecuación de los datos a un modelo de análisis factorial. Mientras más cerca estén de 1 los valores de *KMO*, mejor es la adecuación de los datos a un modelo factorial, considerándose ya excelente la adecuación para valores de *KMO* próximos a 0,9.

También existe una medida de adecuación muestral individual para cada una de las variables basada en la medida *KMO*. Esta medida se denomina *MSA* (*Measure of Sampling Adequacy*), se define de la siguiente forma:

$$MSA_j = \frac{\sum_{h \neq j} r_{jh}^2}{\sum_{h \neq j} r_{jh}^2 + \sum_{h \neq j} a_{jh}^2}$$

Si el valor de MSA_j se aproxima a la unidad, la variable X_j será adecuada para su tratamiento en el análisis factorial con el resto de las variables.

También en el modelo factorial pueden realizarse *contrastes después de la obtención de los factores con los que se pretende evaluar el modelo factorial una vez estimado*. Entre ellos tenemos el contraste para la bondad de ajuste del método de máxima verosimilitud y el contraste para la bondad de ajuste del método MINRES.

Rotación de los factores

El trabajo en el análisis factorial persigue que los factores comunes tengan una interpretación clara, porque de esa forma se analizan mejor las interrelaciones existentes entre las variables originales. Sin embargo, en muy pocas ocasiones resulta fácil encontrar una interpretación adecuada de los factores, iniciales, con independencia del método que se haya utilizado para su extracción. Precisamente los procedimientos de *rotación de factores* se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables.

© Alfaomega - RC Libros

Rotaciones ortogonales

- *Método Varimax.*
- *Método Quartimax.*
- *Métodos Ortomax: Ortomax general, Biquartimax y Equamax.*

A continuación se presentan las características de los métodos más importantes de rotación ortogonal.

- ***Método Varimax***. Método de rotación ortogonal que minimiza el número de variables que tienen saturaciones altas en cada factor. Simplifica la interpretación de los factores.

- ***Método Quartimax***. Método de rotación que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.

- ***Método Equamax***. Método de rotación que es combinación del método Varimax, que simplifica los factores, y el método Quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

Rotaciones oblicuas

- *Método Oblimax y método Quartimin.*
- *Métodos Oblimin: Covarimin, Oblimin directo (o general) y Biquartimin.*
- *Método Oblimin directo: Rotación Promax.*

A continuación se presentan las características de los métodos más importantes de rotación oblicua.

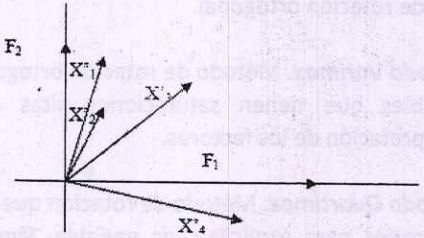
- *Criterio Oblimin directo*. Método para la rotación oblicua (no ortogonal). Cuando delta es igual a cero (el valor por defecto) las soluciones son las más oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor por defecto 0 para delta, introduzca un número menor o igual que 0,8.

© Alfaomega - RC Libros

- *Rotación Promax*. Rotación oblicua que permite que los factores estén correlacionados. Puede calcularse más rápidamente que una rotación Oblimin directa, por lo que es útil para conjuntos de datos grandes.

Interpretación gráfica de los factores

A continuación se presenta un gráfico relativo a cuatro variables X_1, X_2, X_3 y X_4 representadas por dos factores F_1 y F_2.



Como las saturaciones, pesos o cargas factoriales de cada variable en cada factor (elementos de la matriz factorial), se representan por las proyecciones ortogonales de cada variable en cada factor, la cuarta variable se explica fuertemente y de forma positiva por el primer factor (proyección positiva grande de X'_4 sobre F_1), mientras que se representa poco y en sentido negativo por el segundo factor (proyección negativa pequeña de X'_4 sobre F_2).

De la misma forma, la primera y segunda variables se explican fuertemente y de forma positiva por el segundo factor, y se explican poco y de forma positiva por el primer factor. La tercera variable se explica de igual forma por el primero y segundo factor.

Si la representación geométrica resulta difusa, se puede realizar una rotación de los factores que clarifique las proyecciones de las variables sobre ellos. Con una rotación factorial se transforma una solución factorial inicial en otro tipo de solución preferida. Tal transformación va encaminada a poner de manifiesto la solución de la manera más convincente y clara para su interpretación científica.

Puntuaciones o medición de los factores

El análisis factorial es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por los factores obtenidos. Por

© Alfaomega - RC Libros

ejemplo en el caso de estimación de modelos afectados de multicolinealidad. Por ello, es necesario conocer los valores que toman los factores en cada observación (puntuaciones factoriales). Sin embargo, es importante hacer constar que, salvo el caso de que se haya aplicado el análisis de componentes principales para la extracción de factores, no se obtienen unas puntuaciones exactas para los factores.

En su lugar, es preciso realizar estimaciones para obtenerlas. Estas estimaciones se pueden realizar por distintos métodos. Los procedimientos más conocidos, y que aparecen implementados en los paquetes de software, son los de *mínimos cuadrados, regresión, Anderson-Rubin y Barlett.*

Las características de los métodos más importantes de obtención de las componentes se presentan a continuación.

- *Método de regresión*. Método de estimación de los coeficientes de las puntuaciones factoriales. Las puntuaciones resultantes tienen de media 0 y varianza igual al cuadrado de la correlación múltiple entre las puntuaciones factoriales estimadas y los valores factoriales verdaderos. Las puntuaciones pueden estar correlacionadas incluso cuando los factores son ortogonales.

- *Puntuaciones de Barlett*. Método de estimación de los coeficientes para las puntuaciones factoriales. Las puntuaciones resultantes tienen una media de 0. Se minimiza la suma de cuadrados de los factores únicos sobre el rango de las variables.

- *Método de Anderson-Rubin*. Método de estimación de los coeficientes para las puntuaciones factoriales. Es una modificación del método de Barlett, que asegura la ortogonalidad de los factores estimados. Las puntuaciones resultantes tienen una media 0, una desviación típica de 1 y no están correlacionadas.

ANÁLISIS DE COMPONENTES PRINCIPALES A TRAVÉS DE R

El análisis en componentes principales se realiza fácilmente a través de los menús de R. Se comienza importando el archivo de datos (Figura 11-1). En este caso se trata de un archivo con variables económicas y de población de los barrios de Madrid de nombre *zonasmad.sav*. Las variables son población total (*pt*), población menor de 14 (*p14*), población mayor de 10 años (*p10*), jubilados (*p65*), analfabetos (*anal*), nivel de educación superior (*nes*), ocupados (*ocu*), ocupados en la industria (*ocuin*), ocupador en servicios (*ocuser*), técnicos (*tec*), personal directivo (pd) y trabajadores manuales (*tm*). Estas variables, inicialmente correlacionadas, hay que

© Alfaomega - RC Libros

- reducirlas a un grupo menor de variables incorrelacionadas mediante componentes principales. La Figura 11-2 muestra el fichero importado.

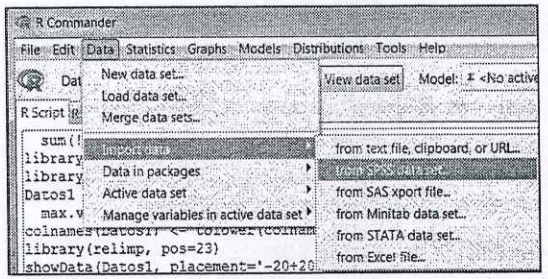
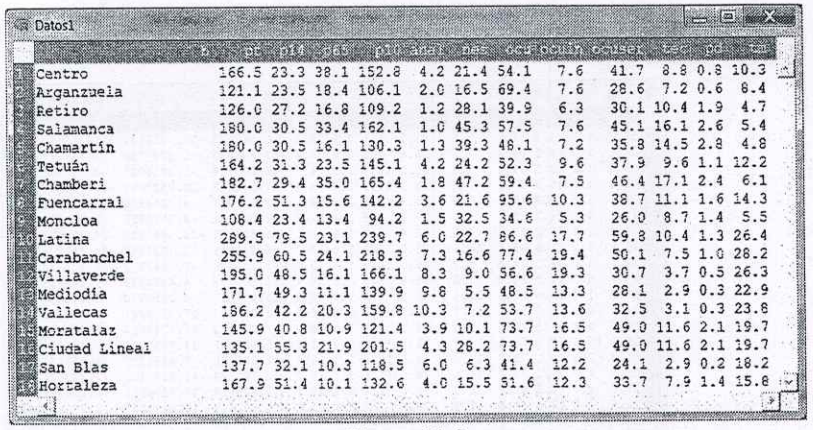


Figura 11-1



Datos1

		pt	p14	p65	p10	anal	nes	ocu	ocuin	ocuser	tec	pd	tm
1	Centro	166.5	23.3	38.1	152.8	4.2	21.4	54.1	7.6	41.7	8.8	0.8	10.3
2	Arganzuela	121.1	23.5	18.4	106.1	2.0	16.5	69.4	7.6	28.6	7.2	0.6	8.4
3	Retiro	126.0	27.2	16.8	109.2	1.2	28.1	39.9	6.3	30.1	10.4	1.9	4.7
4	Salamanca	180.0	30.5	33.4	162.1	1.0	45.3	57.5	7.6	45.1	16.1	2.6	5.4
5	Chamartín	180.0	30.5	16.1	130.3	1.3	39.3	48.1	7.2	35.8	14.5	2.8	4.8
6	Tetuán	164.2	31.3	23.5	145.1	4.2	24.2	52.3	9.6	37.9	9.6	1.1	12.2
7	Chamberí	182.7	29.4	35.0	165.4	1.8	47.2	59.4	7.5	46.4	17.1	2.4	6.1
8	Fuencarral	176.2	51.3	15.6	142.2	3.6	21.6	95.6	10.3	38.7	11.1	1.6	14.3
9	Moncloa	108.4	23.4	13.4	94.2	1.5	32.5	34.6	5.3	26.0	8.7	1.4	5.5
10	Latina	289.5	79.5	23.1	239.7	6.0	22.7	86.6	17.7	59.8	10.4	1.3	26.4
11	Carabanchel	255.9	60.5	24.1	218.3	7.3	16.6	77.4	19.4	50.1	7.5	1.0	28.2
12	Villaverde	195.0	48.5	16.1	166.1	8.3	9.0	56.6	19.3	30.7	3.7	0.5	26.3
13	Mediodía	171.7	49.3	11.1	139.9	9.8	5.5	48.5	13.3	28.1	2.9	0.3	22.9
14	Vallecas	186.2	42.2	20.3	159.8	10.3	7.2	53.7	13.6	32.5	3.1	0.3	23.8
15	Moratalaz	145.9	40.8	10.9	121.4	3.9	10.1	73.7	16.5	49.0	11.6	2.1	19.7
16	Ciudad Lineal	135.1	55.3	21.9	201.5	4.3	28.2	73.7	16.5	49.0	11.6	2.1	19.7
17	San Blas	137.7	32.1	10.3	118.5	6.0	6.3	41.4	12.2	24.1	2.9	0.2	18.2
18	Hortaleza	167.9	51.4	10.1	132.6	4.0	15.5	51.6	12.3	33.7	7.9	1.4	15.8

Figura 11-2

Para realizar el análisis en componentes principales utilizamos la subopción *Principal components analysis* de la opción *Dimensional analysis* del menú *Statistics* (Figura 11-3). En la pantalla *Data* (Figura 11-4) se seleccionan las variables a reducir (en nuestro caso todas) y en la pantalla *Options* (Figura 11-5) se elige el gráfico de sedimentación y añadir las puntuaciones de las componentes como variables adicionales al archivo. Cuando nos pregunten por el número de componentes las dejamos todas para obtener una salida rica en información.

© Alfaomega - RC Libros

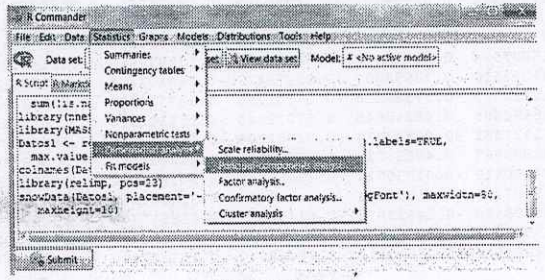


Figura 11-3

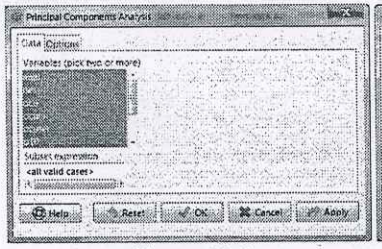


Figura 11-4

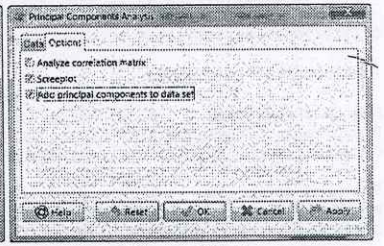


Figura 11-5

La salida, considerando todas las componentes, es la siguiente:

```
Rcmdr>  local({
Rcmdr+     .PC <-
princomp(~anal+nes+ocu+ocuin+ocuser+p10+p14+p65+pd+pt+tec+tm,
Rcmdr+     cor=TRUE, data=Datos1)
Rcmdr+     cat("\nComponent loadings:\n")
Rcmdr+     print(unclass(loadings(.PC)))
Rcmdr+     cat("\nComponent variances:\n")
Rcmdr+     print(.PC$sd^2)
Rcmdr+     cat("\n")
Rcmdr+     print(summary(.PC))
Rcmdr+  })

Component loadings:
           Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
anal    0.36132939  0.17534429  0.23628833 -0.13176183  0.12749884  0.05485120
nes    -0.28540796 -0.32820364  0.09703971 -0.28944474 -0.11113457 -0.41446910
ocu     0.20927928 -0.28583487 -0.29973566  0.73951979 -0.26818978 -0.11095715
ocuin   0.38986887 -0.04602526 -0.18112190 -0.10152847  0.47254174  0.05435351
ocuser  0.13787456 -0.44191228 -0.03535675  0.12830762  0.28485670  0.31547990
p10     0.28372312 -0.32707386  0.23077680 -0.14991710  0.07017480 -0.45379998
p14     0.35853098 -0.16834399 -0.26197238 -0.20903285 -0.27207686 -0.40232095
p65    -0.07566879 -0.31189537  0.69694229  0.27665804  0.20801953 -0.02862781
pd     -0.22567914 -0.34375241 -0.36230083 -0.30243018  0.24137662  0.14405080
pt      0.28253132 -0.26746403  0.21213014 -0.29551274 -0.60024926  0.53995004
```

© Alfaomega - RC Libros

```
tec     -0.23257631 -0.39817625 -0.15910925 -0.03176147  0.08063181  0.16982363
tm       0.41660848  0.03984481 -0.04076219 -0.05186551  0.21850077  0.02569982
              Comp.7      Comp.8      Comp.9     Comp.10     Comp.11     Comp.12
anal     0.73382579  0.32309548  0.15882562 -0.005977797 -0.08814635  0.26517055
nes      0.15987954 -0.23727872  0.54111136  0.174874669  0.32542657  0.14948716
ocu      0.31292510 -0.18558007  0.03978427 -0.099303404  0.08368619  0.05218196
ocuin   -0.10790079 -0.57586218 -0.07161260  0.257324667 -0.07762493  0.39483030
ocuser  -0.35561695  0.48240645  0.37071840 -0.112502510  0.16636495  0.22444295
p10     -0.11331381 -0.06882934 -0.04782204 -0.586652654 -0.38803624 -0.10646122
p14     -0.13096947  0.40856035 -0.32608865  0.442635518  0.07066255  0.05979831
p65      0.04400419 -0.01739724 -0.39861737  0.301573902  0.18652725 -0.06133131
pd       0.35397219  0.01502541 -0.44358687 -0.305366066  0.33019658 -0.06778193
pt      -0.01898488 -0.24523874 -0.03494604 -0.035517778  0.03499011  0.01604175
tec      0.19989627  0.04349334  0.11246593  0.377028499 -0.67979322 -0.25870784
tm       0.04863613 -0.07945910  0.24428249  0.113817220  0.28790766 -0.77967736

Component variances:
      Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6
5.5974741915 4.1270302140 1.0440274082 0.5211206527 0.3205494633 0.1546885170
      Comp.7       Comp.8       Comp.9      Comp.10      Comp.11      Comp.12
0.1119490553 0.0794364531 0.0243412119 0.0129085233 0.0055786045 0.0008957051

Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.3658982 2.0315093 1.02177659 0.72188687 0.56617088
Proportion of Variance 0.4664562 0.3439192 0.08700228 0.04342672 0.02671246
Cumulative Proportion  0.4664562 0.8103754 0.89737765 0.94080437 0.96751683
                           Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation     0.39330461 0.334587889 0.281844732 0.156016704
Proportion of Variance 0.01289071 0.009329088 0.006619704 0.002028434
Cumulative Proportion  0.98040754 0.989736625 0.996356330 0.998384764
                          Comp.10      Comp.11      Comp.12
Standard deviation     0.11361568 0.0746900564 2.992833e-02
Proportion of Variance 0.00107571 0.0004648837 7.464209e-05
Cumulative Proportion  0.99946047 0.9999253579 1.000000e+00
```

Se observa que las tres primeras componentes son las que tienen autovalor (*component variances*) mayor que la unidad. Esto nos lleva a repetir el procedimiento con 3 componentes (Figura 11-6). Al hacer clic en OK, se obtiene el gráfico de sedimentación (Figura 11-7) y se añaden las puntuaciones de las tres componentes al archivo de datos con nombres PC1, PC2 y PC3 (Figura 11-8).

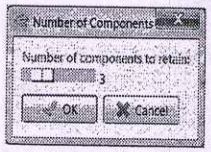


Figura 11-6

© Alfaomega - RC Libros

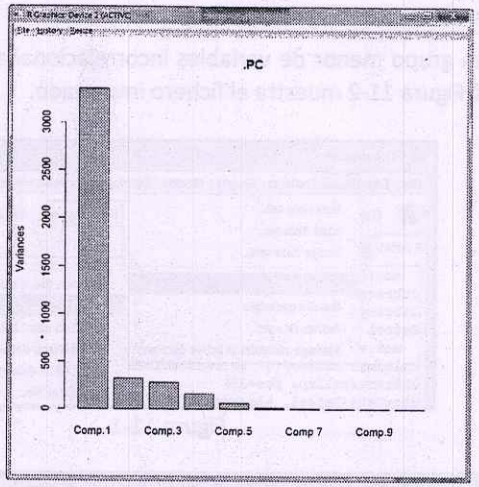
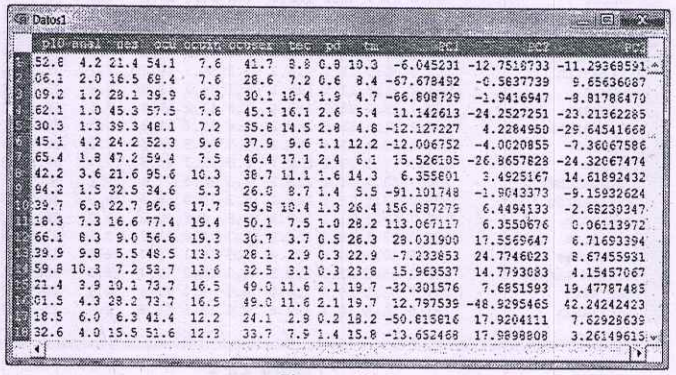


Figura 11-7



	p10	anal	nes	ocu	ocuin	ocuser	tec	pd	tm	pc1	pc2	pc3
1	52.8	4.2	21.4	54.1	7.6	41.7	8.8	6.8	10.3	-6.045231	-12.7518733	-11.29368591
2	06.1	2.0	16.5	69.4	7.6	28.6	7.2	0.6	8.4	-67.678452	-0.5837739	9.65636087
3	09.2	1.2	28.1	39.9	6.3	30.1	10.4	1.9	4.7	-66.808729	-1.9416947	-8.81786470
4	62.1	1.0	45.3	57.5	7.6	45.1	16.1	2.6	5.4	11.142613	-24.2527251	-23.21362285
5	30.3	1.3	39.3	48.1	7.2	35.8	14.5	2.8	4.8	-12.127227	4.2284950	-29.64541668
6	45.1	4.2	24.2	52.3	9.6	37.9	9.6	1.1	12.2	-12.006752	-4.0020855	-7.36067586
7	65.4	1.8	47.2	59.4	7.5	46.4	17.1	2.4	6.1	15.526105	-26.8657828	-24.32067474
8	42.2	3.6	21.6	95.6	10.3	38.7	11.1	1.6	14.3	6.355601	3.4925167	14.61892432
9	94.2	1.5	32.5	34.6	5.3	26.0	8.7	1.4	5.5	-91.101748	-1.9043373	-9.15932624
10	39.7	6.0	22.7	86.6	17.7	59.8	10.4	1.3	26.4	156.887279	6.4494133	-2.68230347
11	18.3	7.3	16.6	77.4	19.4	50.1	7.5	1.0	28.2	113.067117	6.3550676	0.06113972
12	66.1	8.3	9.0	56.6	19.2	30.7	3.7	0.5	26.3	28.031900	17.5569647	6.71693394
13	39.9	9.8	5.5	48.5	13.3	28.1	2.9	0.3	22.9	-7.233853	24.7746023	8.67455931
14	59.8	10.3	7.2	53.7	13.6	32.5	3.1	0.3	23.8	15.963537	14.7793083	4.15457067
15	21.4	3.9	10.1	73.7	16.5	49.0	11.6	2.1	19.7	-32.301576	7.6851593	19.47787485
16	01.5	4.3	28.2	73.7	16.5	49.0	11.6	2.1	19.7	12.797539	-48.9295465	42.24242423
17	18.5	6.0	6.3	41.4	12.2	24.1	2.9	0.2	18.2	-50.615016	17.9204111	7.62928639
18	32.6	4.0	15.5	51.6	12.3	33.7	7.9	1.4	15.8	-13.652468	17.9898808	3.26149615

Figura 11-8

Podemos realizar un gráfico de dispersión que represente la nube de puntos formada por las dos primeras componentes etiquetando los puntos del gráfico con los nombres de los barrios.

En primer lugar definimos la variable de etiquetado como un vector.

```
> v=c( "Centro", "Arganzuela", "Retiro", "Salamanca",
"Chamartín", "Tetuán", "Chamberi",          "Fuencarral",
"Moncloa", "Latina", "Carabanchel", "Villaverde", "Mediodía",
"Vallecas", "Moratalaz", "Ciudad Lineal" , "San Blas",
"Hortaleza")
```

© Alfaomega - RC Libros

En segundo lugar representamos el gráfico de dispersión (Figura 11-9).

```
> scatterplot(PC2~PC1, reg.line=FALSE, smooth=FALSE,
spread=FALSE, labels=v,boxplots=FALSE, span=0.5,
data=Datos1,id.n=18)
```

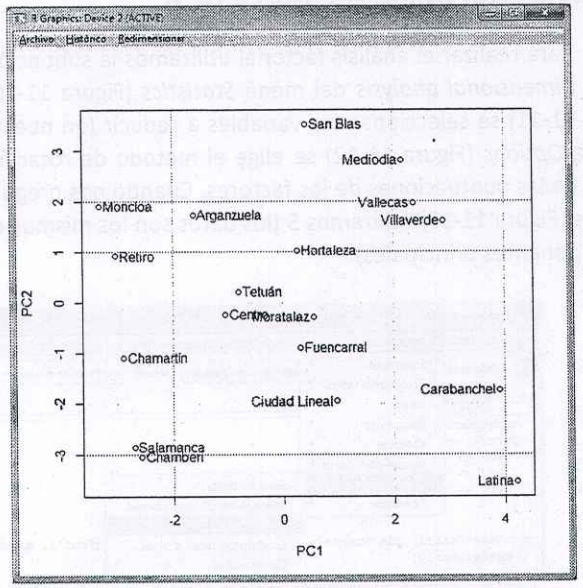


Figura 11-9

Se observa similitud en cuanto a desarrollo entre los barrios Salamanca y Chamberí, otro segmento lo forman Moncloa, Arganzuela, Retiro y quizá Chamartín, otro segmento lo forman San Blas, Mediodía, Vallecas y Villaverde, otro segmento lo forman Hortaleza, Tetuán, Centro, Moratalaz, Fuencarral y Ciudad Lineal. Finalmente, Carabanchel y Latina pueden formar otro segmento. De esta forma hemos realizado una segmentación de los barrios de Madrid por nivel de desarrollo.

COMPONENTES PRINCIPALES A TRAVÉS DE COMANDOS

El comando *princomp* permite realizar análisis en componentes principales mediante la siguiente sintaxis:

princomp(~v1+v2+...+vn, cor=TRUE, scores=TRUE, data=Conjunto de datos)

© Alfaomega - RC Libros

Las variables *v1, v2, ..., vn* son las variables a reducir, *cor=TRUE* permite obtener la matriz de correlaciones entre componentes y variables (matriz factorial), *scores=TRUE* permite calcular las puntuaciones de las componentes y data=Conjunto de datos permite declarar el conjunto de datos que contiene las variables.

En nuestro ejemplo anterior la sintaxis vía comandos sería la siguiente:

```
> componentes=princomp(~anal+nes+ocu+ocuin+ocuser+p10+p14+p65
+pd+pt+tec+tm, cor=TRUE, scores=TRUE, data=Datos1)

> summary(componentes)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.3658982 2.0315093 1.02177659 0.72188687 0.56617088
Proportion of Variance 0.4664562 0.3439192 0.08700228 0.04342672 0.02671246
Cumulative Proportion  0.4664562 0.8103754 0.89737765 0.94080437 0.96751683
                           Comp.6       Comp.7       Comp.8      Comp.9
Standard deviation     0.39330461 0.334587889 0.281844732 0.156016704
Proportion of Variance 0.01289071 0.009329088 0.006619704 0.002028434
Cumulative Proportion  0.98040754 0.989736625 0.996356330 0.998384764
                          Comp.10      Comp.11      Comp.12
Standard deviation     0.11361568 0.0746900564 2.992833e-02
Proportion of Variance 0.00107571 0.0004648837 7.464209e-05
Cumulative Proportion  0.99946047 0.9999253579 1.000000e+00
```

Para ver las puntuaciones utilizamos la sintaxis siguiente:

```
> componentes$scores
      Comp.1     Comp.2      Comp.3      Comp.4      Comp.5       Comp.6
1  -1.0583249 -0.2517475  2.33732497  1.02607277  0.30778512  0.295810697
2  -1.6275133  1.7205507 -0.12963914  1.70726329 -0.48831649 -0.162726086
3  -3.0462773  0.9173994 -0.35585201 -0.47702246 -0.04117624 -0.042739880
4  -2.7111251 -2.8531152  0.71525494 -0.26654232  0.12816107  0.014387564
5  -2.9050916 -1.0896091 -0.75788036 -1.19984495 -0.44374407  0.388387352
6  -0.8074149  0.2052439  0.72419066  0.09892785  0.07254999  0.108574846
7  -2.5801503 -3.0454383  0.99602206 -0.13986287  0.15015737  0.019671017
8   0.3024332 -0.9018130 -1.45279736  1.35182621 -1.05475948 -0.248976194
9  -3.3322879  1.9024023 -0.25377244 -0.55351576 -0.16573086 -0.372542743
10  4.2262072 -3.5484776 -0.05781957 -0.34707850 -0.90022151  0.006380601
11  3.9144767 -1.7361601  0.49995579 -0.12089343 -0.04545549  0.181306318
12  2.9101257  1.6140726  0.28684215 -0.45261259  0.31992669 -0.066180050
13  2.1631845  2.8179789  0.21335306 -0.52644559 -0.18212091 -0.079367703
14  2.3700093  1.9640946  1.22166406 -0.07666055  0.12341654  0.094840319
15  0.5443036 -0.2972883 -2.15305487  0.64309854  1.13432236  0.997767158
16  0.9616979 -1.9537882 -0.86868609  0.12590641  1.30353426 -1.120873931
17  0.4181049  3.5213894  0.10008249 -0.12732263  0.09200860 -0.007209674
18  0.2576423  1.0143055 -1.06518835 -0.66529340 -0.31033695 -0.006509610
          Comp.7      Comp.8      Comp.9      Comp.10      Comp.11      Comp.12
1  -0.273230422  0.28460511 -0.20936686 -0.016693221  0.04954580  0.002261065
2  -0.246001294 -0.31404592  0.04268136 -0.065133629 -0.04169829  0.047260661
3  -0.250387819  0.12119159 -0.25193360 -0.068990397  0.01326749  0.005391873
4   0.143454692 -0.13073527 -0.11614703  0.054866755  0.06414854 -0.010477964
5   0.323125843 -0.19885567 -0.03055060 -0.250001215 -0.04259783  0.023568108
6  -0.121586522  0.08026943  0.18392380  0.002292946 -0.13932098  0.018150859
7   0.336452333 -0.07541446  0.16603698  0.195997918 -0.09892113 -0.023148243
8   0.581791715  0.01457186 -0.06802199  0.042224264  0.03115285 -0.035994541
9  -0.275725625  0.05148293  0.33533566  0.075688347  0.18356461  0.003254183
```

© Alfaomega - RC Libros

```
10 -0.554112377  0.28277147  0.10574047 -0.002767937 -0.01557107  0.026417335
11 -0.171751287 -0.39030101  0.05503707 -0.110756125  0.08265677 -0.035455634
12  0.128455522 -0.74574578 -0.13208724  0.166941607  0.01325030  0.039761861
13  0.458475436  0.44965895  0.03682383  0.079736755 -0.03905199  0.029827960
14  0.601510808  0.25715270  0.02184608 -0.129477712  0.06696736 -0.005153316
15 -0.020510233  0.14051123  0.09873543  0.047525961  0.03100537  0.006751593
16  0.009238537  0.12895639 -0.03396946 -0.111674209 -0.02913066  0.005977207
17 -0.323121511 -0.16488599  0.08624009 -0.076470271 -0.10224167 -0.077867503
18 -0.346077796  0.20881243 -0.29032397  0.166690163 -0.02702545 -0.020525505
```

Para ver la matriz factorial utilizamos la siguiente sintaxis:

```
> print(unclass(loadings(componentes)))
         Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
anal   0.36132939  0.17534429  0.23628833 -0.13176183  0.12749884  0.05485120
nes   -0.28540796 -0.32820364  0.09703971 -0.28944474 -0.11113457 -0.41446910
ocu    0.20927928 -0.28583487 -0.29973566  0.73951979 -0.26818978 -0.11095715
ocuin  0.38986887 -0.04602526 -0.18112190 -0.10152847  0.47254174  0.05435351
ocuser 0.13787456 -0.44191228 -0.03535675  0.12830762  0.28485670  0.31547990
p10    0.28372312 -0.32707386  0.23077680 -0.14991710  0.07017480 -0.45379998
p14    0.35853098 -0.16834399 -0.26197238 -0.20903285 -0.27207686 -0.40232095
p65   -0.07566879 -0.31189537  0.69694229  0.27665804  0.20801953 -0.02862781
pd    -0.22567914 -0.34375241 -0.36230083 -0.30243018  0.24137662  0.14405080
pt     0.28253132 -0.26746403  0.21213014 -0.29551274 -0.60024926  0.53995004
tec   -0.23257631 -0.39817625 -0.15910925 -0.03176147  0.08063181  0.16982363
tm     0.41660848  0.03984481 -0.04076219 -0.05186551  0.21850077  0.02569982
         Comp.7      Comp.8      Comp.9     Comp.10     Comp.11     Comp.12
anal   0.73382579  0.32309548  0.15882562 -0.005977797 -0.08814635  0.26517055
nes    0.15987954 -0.23727872  0.54111136  0.174874669  0.32542657  0.14948716
ocu    0.31292510 -0.18558007  0.03978427 -0.099303404  0.08368619  0.05218196
ocuin -0.10790079 -0.57586218 -0.07161260  0.257324667 -0.07762493  0.39483030
ocuser -0.35561695  0.48240645  0.37071840 -0.112502510  0.16636495  0.22444295
p10   -0.11331381 -0.06882934 -0.04782204 -0.586652654 -0.38803624 -0.10646122
p14   -0.13096947  0.40856035 -0.32608865  0.442635518  0.07066255  0.05979831
p65    0.04400419 -0.01739724 -0.39861737  0.301573902  0.18652725 -0.06133131
pd     0.35397219  0.01502541 -0.44358687 -0.305366066  0.33019658 -0.06778193
pt    -0.01898488 -0.24523874 -0.03494604 -0.035517778  0.03499011  0.01604175
tec    0.19989627  0.04349334  0.11246593  0.377028499 -0.67979322 -0.25870784
tm     0.04863613 -0.07945910  0.24428249  0.113817220  0.28790766 -0.77967736
```

Para aislar las puntuaciones de las tres primeras componentes usamos la sintaxis siguiente:

```
> COMP1=componentes$scores[,1]
> COMP2=componentes$scores[,2]
> COMP3=componentes$scores[,3]
```

Ahora ya podemos representar las puntuaciones de la primera componente contra las puntuaciones de la segunda.

```
> v=c( "Centro", "Arganzuela", "Retiro", "Salamanca",
"Chamartín", "Tetuán", "Chamberi", "Fuencarral", "Moncloa",
"Latina", "Carabanchel", "Villaverde", "Mediodía","Vallecas",
"Moratalaz", "Ciudad Lineal" , "San Blas", "Hortaleza")
```


© Alfaomega - RC Libros


```
> scatterplot(COMP2~COMP1, reg.line=FALSE, smooth=FALSE,
spread=FALSE, labels=v,boxplots=FALSE, span=0.5,
data=Datos1,id.n=18)
```

Se obtiene el gráfico de la Figura 11-9.

ANÁLISIS FACTORIAL A TRAVÉS DE R

Para realizar el análisis factorial utilizamos la subopción *Factor analysis* de la opción *Dimensional analysis* del menú *Statistics* (Figura 11-10). En la pantalla *Data* (Figura 11-11) se seleccionan las variables a reducir (en nuestro caso todas) y en la pantalla *Options* (Figura 11-12) se elige el método de rotación y el método para el cálculo de las puntuaciones de los factores. Cuando nos pregunten por el número de factores (Figura 11-13) utilizamos 3 (los datos son los mismos que en el caso anterior de componentes principales).

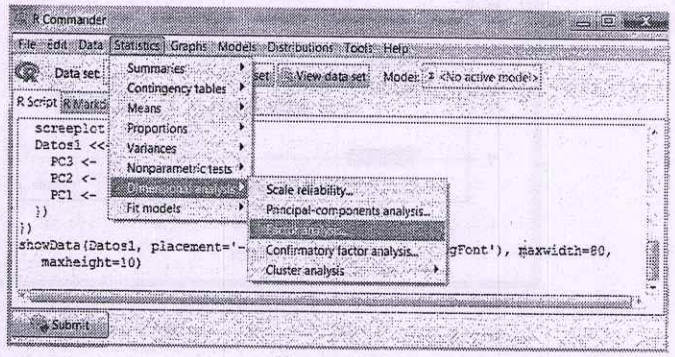


Figura 11-10

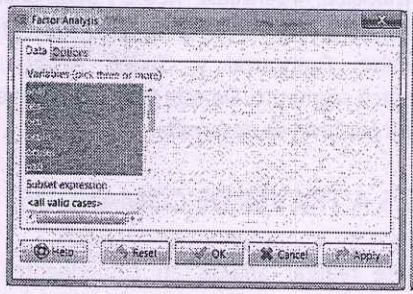


Figura 11-11

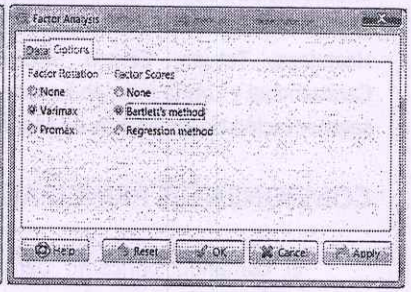


Figura 11-12

© Alfaomega - RC Libros

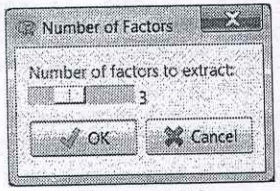


Figura 11-13

Al pulsar OK se obtiene la salida siguiente:

```
Rcmdr>  local({
Rcmdr+    .FA <-
factanal(~anal+nes+ocu+ocuin+ocuser+p10+p14+p65+pd+pt+tec+tm,
Rcmdr+    factors=3, rotation="varimax", scores="Bartlett",
data=Datos1)
Rcmdr+    print(.FA)
Rcmdr+    Datos1 <<- within(Datos1, {
Rcmdr+      F3 <- .FA$scores[,3]
Rcmdr+      F2 <- .FA$scores[,2]
Rcmdr+      F1 <- .FA$scores[,1]
Rcmdr+    })
Rcmdr+  })

Call:
factanal(x = ~anal + nes + ocu + ocuin + ocuser + p10 + p14 +
p65 + pd + pt + tec + tm, factors = 3, data = Datos1, scores =
"Bartlett",     rotation = "varimax")

Uniquenesses:
  anal   nes   ocu  ocuin ocuser   p10   p14   p65    pd    pt   tec
 0.116 0.115 0.482  0.043  0.084 0.086 0.194 0.108 0.050 0.316 0.005
    tm
 0.005

Loadings:
        Factor1 Factor2 Factor3
anal     0.433  -0.832
nes     -0.126   0.843   0.397
ocu      0.704   0.146
ocuin    0.836  -0.430  -0.269
ocuser   0.809   0.420   0.293
p10      0.875           0.380
p14      0.855  -0.216  -0.170
p65      0.150   0.287   0.887
pd       0.104   0.968
pt       0.764  -0.112   0.295
```


© Alfaomega - RC Libros


```
tec      0.105   0.966   0.227
tm       0.734  -0.647  -0.199

               Factor1 Factor2 Factor3
SS loadings      4.717   4.219   1.462
Proportion Var   0.393   0.352   0.122
Cumulative Var   0.393   0.745   0.866

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 68.58 on 33 degrees of freedom.
The p-value is 0.000271
RcmdrMsg: [7] NOTE: The dataset Datos1 has 18 rows and 19
columns.
```

En la salida se observa la matriz factorial, los valores propios asociados a cada factor (SS loadings), la proporción de varianza explicada por cada factor y la proporción acumulada (entre los tres factores se explica el 86,6% de la variabilidad de los datos). La reducción es bastante buena.

Las puntuaciones de los factores se sitúan como variables adicionales en la parte derecha del conjunto de datos (F1, F2, ...).

ANÁLISIS FACTORIAL A TRAVÉS DE COMANDOS

El comando *factanal* permite realizar análisis factorial mediante la siguiente sintaxis sencilla:

factanal(~v1+v2+...+vn, factors=n, rotation=METHOD, scores=METHOD, data=Conjunto de datos)

Las variables *v1, v2, ..., vn* son las variables a reducir, *factors =n* indica el número de factores a considerar, *rotation=METHOD* permite elegir el método de rotación, *scores=METHOD* permite calcular las puntuaciones de los factores a través del método especificado y *data=Conjunto de datos* permite declarar el conjunto de datos que contiene las variables.

En nuestro ejemplo anterior la sintaxis vía comandos sería la siguiente:

```
> factorial1=factanal(~anal+nes+ocu+ocuin+ocuser+p10+p14+p65+
pd+pt+tec+tm, factors=3, rotation="varimax", scores="Bartlett",
data=Datos1)
```

© Alfaomega - RC Libros

```
>
> summary(factorial1)
             Length Class      Mode
converged      1    -none-     logical
loadings      36    loadings   numeric
uniquenesses  12    -none-     numeric
correlation  144    -none-     numeric
criteria       3    -none-     numeric
factors        1    -none-     numeric
dof            1    -none-     numeric
method         1    -none-     character
rotmat         9    -none-     numeric
scores        54    -none-     numeric
STATISTIC      1    -none-     numeric
PVAL           1    -none-     numeric
n.obs          1    -none-     numeric
call           6    -none-     call
> factorial1

Call:
factanal(x = ~anal + nes + ocu + ocuin + ocuser + p10 + p14 +
p65 + pd + pt + tec + tm, factors = 3, data = Datos1, scores =
"Bartlett",      rotation = "varimax")

Uniquenesses:
  anal   nes   ocu  ocuin ocuser   p10   p14   p65    pd    pt   tec
 0.116 0.115 0.482  0.043  0.084 0.086 0.194 0.108 0.050 0.316 0.005
    tm
 0.005

Loadings:
       Factor1 Factor2 Factor3
anal    0.433  -0.832
nes    -0.126   0.843   0.397
ocu     0.704   0.146
ocuin   0.836  -0.430  -0.269
ocuser  0.809   0.420   0.293
p10     0.875           0.380
p14     0.855  -0.216  -0.170
p65     0.150   0.287   0.887
pd      0.104   0.968
pt      0.764  -0.112   0.295
tec     0.105   0.966   0.227
tm      0.734  -0.647  -0.199
```


© Alfaomega - RC Libros


```
               Factor1 Factor2 Factor3
SS loadings      4.717   4.219   1.462
Proportion Var   0.393   0.352   0.122
Cumulative Var   0.393   0.745   0.866

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 68.58 on 33 degrees of freedom.
The p-value is 0.000271
```

Para ver las puntuaciones de los factores utilizamos la siguiente sintaxis:

```
> factorial1$scores
       Factor1      Factor2     Factor3
1  -0.64901110 -0.572629411  2.3761608
2  -1.35617486 -0.395564730  0.2297504
3  -1.27445248  0.570143759 -0.3846389
4  -0.04457714  1.422842853  1.1347120
5  -0.51867804  1.535470459 -0.7084447
6  -0.35147699 -0.043352970  0.6582699
7   0.17115237  1.510830912  1.3970493
8   0.15999141  0.545089200 -0.5784566
9  -1.56067686  0.185911954 -0.4408794
10  2.02694800 -0.067270211  0.7246813
11  1.68152582 -0.678836218  0.4890759
12  0.59223374 -1.255573209 -0.5537121
13 -0.13747474 -1.395465935 -0.5520414
14  0.09534839 -1.589541366  0.4450031
15  0.95821451  0.932617739 -2.1015699
16  1.19686528  0.570099095 -0.2967166
17 -0.81596236 -1.279682860 -0.6664218
18 -0.17379492  0.004910936 -1.1718213
```

Podemos definir las puntuaciones para los tres primeros factores como variables mediante la sintaxis siguiente:

```
> F1=factorial1$scores[,1]
> F2=factorial1$scores[,2]
> F3=factorial1$scores[,3]
```

Para realizar un gráfico de dispersión del primer factor contra el segundo con etiquetas en los puntos (Figura 11-14) utilizaremos la siguiente sintaxis:

© Alfaomega - RC Libros

```
> attach(Datos2)
> b
 [1] "Centro                "  "Arganzuela            "  "Retiro                "
 [4] "Salamanca             "  "Chamartín             "  "Tetuán                "
 [7] "Chamberi              "  "Fuencarral            "  "Moncloa               "
[10] "Latina                "  "Carabanchel           "  "Villaverde            "
[13] "Mediodía              "  "Vallecas              "  "Moratalaz             "
[16] "Ciudad Lineal         "  "San Blas              "  "Hortaleza             "

> w=cbind(b)
> w
      b
 [1,] "Centro                "
 [2,] "Arganzuela            "
 [3,] "Retiro                "
 [4,] "Salamanca             "
 [5,] "Chamartín             "
 [6,] "Tetuán                "
 [7,] "Chamberi              "
 [8,] "Fuencarral            "
 [9,] "Moncloa               "
[10,] "Latina                "
[11,] "Carabanchel           "
[12,] "Villaverde            "
[13,] "Mediodía              "
[14,] "Vallecas              "
[15,] "Moratalaz             "
[16,] "Ciudad Lineal         "
[17,] "San Blas              "
[18,] "Hortaleza             "

> library("car", lib.loc="C:/Users/usuario/Documents/R/win-
library/3.2")

> scatterplot(F2~F1, reg.line=FALSE, smooth=FALSE,
spread=FALSE, labels=w,boxplots=FALSE, span=0.5,
data=Datos1,id.n=18)
```


© Alfaomega - RC Libros

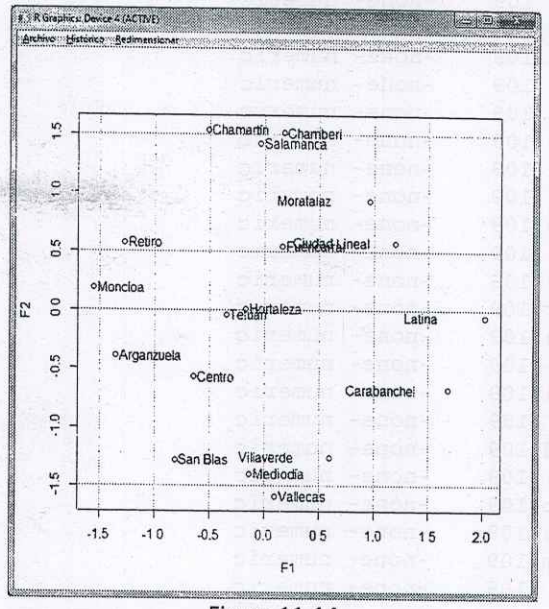


Figura 11-14

Ejercicio 11-1. El archivo mundo.sav contiene información sobre variables demográficas y de desarrollo de varios países del mundo altamente correlacionadas. Se trata de reducir el número inicial de variables correlacionadas a un número menor de variables incorrelacionadas con la mínima pérdida de información.

Comenzamos leyendo el fichero mundo.sav en formato SPSS que contiene los datos.

```
> library(foreign)
> mundo1=read.spss("H:/R/data/mundo.sav",
use.value.labels=FALSE)
```

A continuación observamos las variables del fichero.

```
> summary(mundo1)
         Length Class  Mode
país     109    -none- character
relig    109    -none- character
poblac   109    -none- numeric
densidad 109    -none- numeric
```

© Alfaomega - RC Libros

```
urbana    109    -none- numeric
espvidaf  109    -none- numeric
espvidam  109    -none- numeric
alfabet   109    -none- numeric
inc_pob   109    -none- numeric
mortinf   109    -none- numeric
pib_cap   109    -none- numeric
región    109    -none- numeric
calorías  109    -none- numeric
sida      109    -none- numeric
tasa_nat  109    -none- numeric
tasa_mor  109    -none- numeric
tasasida  109    -none- numeric
log_pib   109    -none- numeric
logtsida  109    -none- numeric
nac_def   109    -none- numeric
fertilid  109    -none- numeric
log_pob   109    -none- numeric
cregrano  109    -none- numeric
alfabmas  109    -none- numeric
alfabfem  109    -none- numeric
clima     109    -none- numeric
```

La siguiente tarea es realizar la reducción de la dimensión propuesta aplicando componentes principales.

```
> componentes=princomp(~poblac+densidad+urbana+espvidaf+
espvidam+alfabet+inc_pob+mortinf +pib_cap+calorías+sida+
tasa_nat +tasa_mor+tasasida+log_pib+logtsida+nac_def+fertilid
+log_pob+cregrano+alfabmas+alfabfem+clima, cor=TRUE,
scores=TRUE, data=mundo1)

> summary(componentes)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     3.2826846 1.7928472 1.47425240 1.23657441 1.07725328 0.91551886
Proportion of Variance 0.4685225 0.1397522 0.09449653 0.06648332 0.05045542 0.03644238
Cumulative Proportion  0.4685225 0.6082747 0.70277127 0.76925458 0.81971000 0.85615238
                           Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
Standard deviation     0.83449825 0.74676254 0.64787775 0.62022101 0.57376031 0.50628149
Proportion of Variance 0.03027771 0.02424584 0.01824981 0.01672496 0.01431308 0.01114439
Cumulative Proportion  0.88643009 0.91067593 0.92892574 0.94565070 0.95996378 0.97110817
                          Comp.13     Comp.14     Comp.15     Comp.16     Comp.17     Comp.18
Standard deviation     0.4844597 0.336806301 0.303175250 0.250579757 0.226287471 0.189284971
Proportion of Variance 0.0102044 0.004932108 0.003996314 0.002730009 0.002226349 0.001557774
Cumulative Proportion  0.9813126 0.986244681 0.990240995 0.992971005 0.995197353 0.996755127
                           Comp.19      Comp.20      Comp.21      Comp.22     Comp.23
Standard deviation     0.168781391 0.1340427268 0.1101137396 0.1063401049 0.06887823
Proportion of Variance 0.001238572 0.0007811936 0.0005271755 0.0004916616 0.00020627
Cumulative Proportion  0.997993699 0.9987748929 0.9993020683 0.9997937300 1.00000000
```


© Alfaomega - RC Libros

Se observa que los *p*-valores (*Standard deviation*) mayores que la unidad corresponden a las 5 primeras componentes que suman entre todas prácticamente el 80% de la proporción de la varianza total (muy buena reducción).

Para aislar las puntuaciones de las cinco primeras componentes usamos la sintaxis siguiente:

```
> COMP1=componentes$scores[,1]
> COMP2=componentes$scores[,2]
> COMP3=componentes$scores[,3]
> COMP4=componentes$scores[,4]
> COMP5=componentes$scores[,5]
```

Ahora ya podemos representar las puntuaciones de la primera componente contra las puntuaciones de la segunda etiquetando los puntos (Figura 11-15). Los países ricos aparecen cercanos en la gráfica y los pobres también.

```
> attach(mundo1)
> v=cbind(país)
> v
      país
 [1,] "Azerbaján      "
 [2,] "Afganistán     "
 [3,] "Alemania       "
 [4,] "Arabia Saudí   "
 [5,] "Argentina      "
 [6,] "Armenia        "
 [7,] "Australia      "
 [8,] "Austria        "
 [9,] "Bahrein        "
[10,] "Bangladesh     "
[11,] "Barbados       "
[12,] "Bélgica        "
[13,] "Bielorrusia    "
[14,] "Bolivia        "
[15,] "Bosnia         "
[16,] "Botswana       "
[17,] "Brasil         "
[18,] "Bulgaria       "
[19,] "Burkina Faso   "
[20,] "Burundi        "
[21,] "Camboya        "
[22,] "Camerún        "
[23,] "Canadá         "
```

© Alfaomega - RC Libros

```
 [24,] "Colombia        "
 [25,] "Corea del Norte "
 [26,] "Corea del Sur   "
 [27,] "Costa Rica      "
 [28,] "Croacia         "
 [29,] "Cuba            "
 [30,] "Chile           "
 [31,] "China           "
 [32,] "Dinamarca       "
 [33,] "Ecuador         "
 [34,] "Egipto          "
 [35,] "El Salvador     "
 [36,] "Emiratos Árabes "
 [37,] "España          "
 [38,] "Estados Unidos  "
 [39,] "Estonia         "
 [40,] "Etiopía         "
 [41,] "Filipinas       "
 [42,] "Finlandia       "
 [43,] "Francia         "
 [44,] "Gabón           "
 [45,] "Gambia          "
 [46,] "Georgia         "
 [47,] "Gran Bretaña    "
 [48,] "Grecia          "
 [49,] "Guatemala       "
 [50,] "Haití           "
 [51,] "Honduras        "
 [52,] "Hong Kong       "
 [53,] "Hungría         "
 [54,] "India           "
 [55,] "Indonesia       "
 [56,] "Irán            "
 [57,] "Iraq            "
 [58,] "Irlanda         "
 [59,] "Islandia        "
 [60,] "Israel          "
 [61,] "Italia          "
 [62,] "Japón           "
 [63,] "Jordania        "
 [64,] "Kenia           "
 [65,] "Kuwait          "
 [66,] "Letonia         "
 [67,] "Líbano          "
 [68,] "Liberia         "
```


© Alfaomega - RC Libros


```
 [69,] "Libia           "
 [70,] "Lituania        "
 [71,] "Malasia         "
 [72,] "Marruecos       "
 [73,] "México          "
 [74,] "Nicaragua       "
 [75,] "Nigeria         "
 [76,] "Noruega         "
 [77,] "Nueva Zelanda   "
 [78,] "Omán            "
 [79,] "Países Bajos    "
 [80,] "Pakistán        "
 [81,] "Panamá          "
 [82,] "Paraguay        "
 [83,] "Perú            "
 [84,] "Polonia         "
 [85,] "Portugal        "
 [86,] "Rep. C. Africana"
 [87,] "Rep. Checa      "
 [88,] "Rep. Dominicana "
 [89,] "Ruanda          "
 [90,] "Rumanía         "
 [91,] "Rusia           "
 [92,] "Senegal         "
 [93,] "Singapur        "
 [94,] "Siria           "
 [95,] "Somalia         "
 [96,] "Sudáfrica       "
 [97,] "Suecia          "
 [98,] "Suiza           "
 [99,] "Tailandia       "
[100,] "Taiwan          "
[101,] "Tanzania        "
[102,] "Turquía         "
[103,] "Ucrania         "
[104,] "Uganda          "
[105,] "Uruguay         "
[106,] "Uzbekistán      "
[107,] "Venezuela       "
[108,] "Vietnam         "
[109,] "Zambia          "

> windows()
> scatterplot(COMP2~COMP1, reg.line=FALSE, smooth=FALSE,
spread=FALSE, labels=v,boxplots=FALSE, span=0.5,
data=mundo1,id.n=109)
```

© Alfaomega - RC Libros

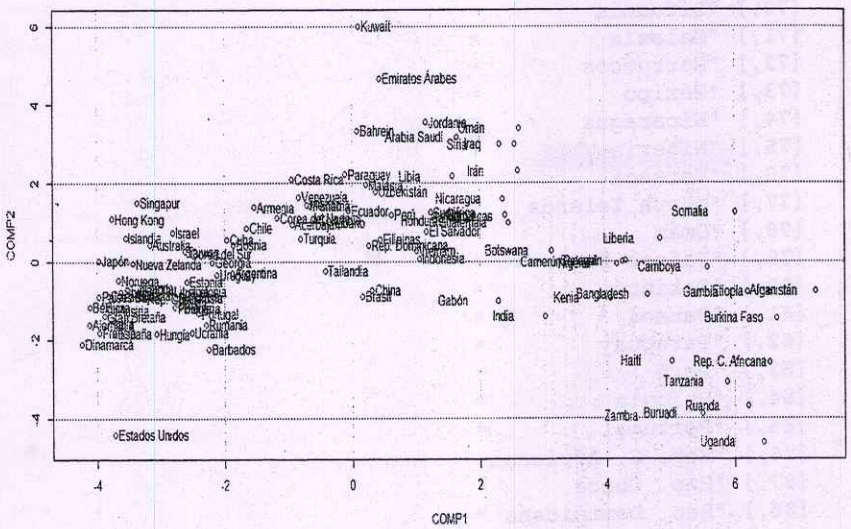


Figura 11-15

© Alfaomega - RC Libros

CAPÍTULO 12

TÉCNICAS DE CLASIFICACIÓN Y SEGMENTACIÓN A TRAVÉS DE R

EL ANÁLISIS CLÚSTER

El análisis de conglomerados, también llamado análisis clúster, constituye uno de los procedimientos estadísticos más utilizados hoy en día para la segmentación. La definición de la propia técnica se asemeja, de hecho, a los fines genéricos que persigue la segmentación: identificar grupos de sujetos lo más heterogéneos posible entre sí y lo más homogéneos posible dentro de cada grupo.

El análisis establece dichos grupos basándose en la similitud que presentan un conjunto de entidades (por ejemplo, turistas) respecto a una serie de características que el investigador ha de especificar previamente (motivaciones, necesidades, beneficios buscados, etc.). No obstante, es el análisis, y no el analista, el que finalmente extrae los grupos de sujetos y sus características definitorias: número de segmentos, número de integrantes de cada segmento, etc. El análisis de conglomerados constituye, por tanto, el ejemplo paradigmático del enfoque de *segmentación post hoc*. Asimismo se trata de un método descriptivo de segmentación.

© Alfaomega - RC Libros

El análisis clúster es una técnica de *Data Mining* de clasificación automática de datos. Su finalidad esencial es revelar concentraciones en los datos (casos o variables) para su agrupamiento eficiente en clústeres (o conglomerados) según su homogeneidad. El agrupamiento puede realizarse tanto para casos como para variables, pudiendo utilizarse variables cualitativas o cuantitativas. Los grupos de casos o variables se realizan basándose en la proximidad o lejanía de unos con otras, por lo tanto es esencial el uso adecuado del concepto de distancia. Es fundamental que los elementos dentro de un clúster sean homogéneos y lo más diferentes posibles de los contenidos en otros clústeres.

El análisis clúster es por tanto una técnica de clasificación, conociéndose también con el nombre de *taxonomía numérica*. Otros nombres asignados al mismo concepto son análisis de *conglomerados, análisis tipológico, clasificación automática* y otros. El número de clústeres no es conocido de antemano y los grupos se crean en función de la naturaleza de los datos. Se trata por tanto de una técnica de clasificación *post hoc*. Podíamos definir el análisis clúster como un método estadístico multivariante de clasificación automática que a partir de una tabla de datos (casos-variables), trata de situarlos en grupos homogéneos, conglomerados o clústeres, no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que los individuos que puedan ser considerados similares sean asignados a un mismo clúster, mientras que individuos diferentes (disimilares) se localicen en clústeres distintos.

La diferencia esencial con el análisis discriminante estriba en que en este último es necesario especificar previamente los grupos por un camino objetivo (técnica de clasificación *ad hoc*), ajeno a la medida de las variables en los casos de la muestra.

El análisis clúster define grupos tan distintos como sea posible en función de los propios datos sin especificación previa de los citados grupos (técnica de clasificación *post hoc*). Si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto de la media. Es necesario observar también los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clústeres unitarios. También es nocivo para el análisis clúster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad. Si es necesario se realiza un análisis factorial previo y posteriormente se aglomeran las puntuaciones factoriales. La solución del análisis clúster no tiene por qué ser única, pero no deben encontrarse soluciones contradictorias por distintos métodos. El número de observaciones en cada clúster debe ser relevante, ya

© Alfaomega - RC Libros

que en caso contrario puede haber valores atípicos que difuminen la construcción de los clústeres. Los conglomerados deben tener sentido conceptual y no variar mucho al variar la muestra o el método de aglomeración. Los grupos finales serán tan distintos como permitan los datos. Con estos grupos se podrán realizar otros análisis: descriptivos, discriminante, regresión logística, diferencia...

CLÚSTERES JERÁRQUICOS, SECUENCIALES, AGLOMERATIVOS Y EXCLUSIVOS (S.A.H.N.)

Los *métodos de análisis de conglomerados* que más se usan son los que son a la vez secuenciales, aglomerativos, jerárquicos y exclusivos, y que reciben el acrónimo, en lengua inglesa, de S.A.H.N. (*Sequential, Agglomerative, Hierarchic* y *Nonoverlaping*). En todos los *métodos de tipo S.A.H.N.* se siguen dos pasos fundamentales en el proceso de elaboración de los conglomerados.

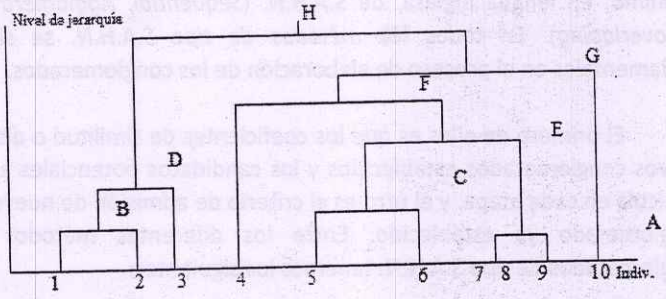
El primero de ellos es que los coeficientes de similitud o disimilitud entre los nuevos conglomerados establecidos y los candidatos potenciales a ser admitidos se recalcula en cada etapa, y el otro es el criterio de admisión de nuevos miembros a un conglomerado ya establecido. Entre los diferentes métodos de análisis de conglomerados de tipo S.A.H.N. tenemos los siguientes:

- Método de unión simple (*Single Linkage Clustering*), entorno o vecino más cercano (*Nearest Neighbour*) o método del mínimo (*Minimum Method*)
- Método de la distancia máxima o método del máximo (*Complete Linkage Clustering, Furthest Neighbour* o *Maximum Method*)
- Método de la media o de la distancia promedio no ponderado (*Weighted Pair Groups Method Using Arithmetic Averages* WPGMW)
- Método de la media ponderada o de la distancia promedio ponderado (*Group Average o Unweighted Pair Groups Method Using Arithmetic Averages* UPGMA)
- Método de la mediana o de la distancia mediana (*Weighted Pair Group Centroid Method* WPGMC)
- Método del Centroide o de la Distancia Prototipo (*Unweighted Pair Group Centroid Method* UPGMC)
- Método de Ward o de mínima varianza

© Alfaomega - RC Libros

EL DENDOGRAMA EN EL ANÁLISIS CLÚSTER JERÁRQUICO

Es habitual en la investigación la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clústeres, los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al clúster final que contiene todos los casos analizados. La representación gráfica de estas etapas de formación de grupos, a modo de árbol invertido, se denomina *dendograma* y se representa a continuación:



La figura, que corresponde a un estudio de los individuos, muestra cómo el 8 y el 9 se agrupan en un primer clúster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 (clúster B); y enseguida los 5, 6 y 7 (C). Un paso siguiente engloba el clúster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el clúster total (H) que reúne los 10 casos.

ANÁLISIS CLÚSTER NO JERÁRQUICO

La clasificación de todos los casos de una tabla de datos en grupos separados que configura el propio análisis proporciona clústeres no jerárquicos. Esta denominación alude a la no existencia de una estructura vertical de dependencia entre los grupos formados y, por consiguiente, estos no se presentan en distintos niveles de jerarquía. El análisis precisa que el investigador fije de antemano el número de clústeres en que quiere agrupar sus datos.

© Alfaomega - RC Libros

Como puede no existir un número definido de grupos o, si existe, generalmente no se conoce, la prueba debe ser repetida con diferente número a fin de tantear la clasificación que mejor se ajuste al objetivo del problema, o la de más clara interpretación.

Los métodos no jerárquicos también se conocen como *métodos partitivos* o de optimización, dado que, como hemos visto, tienen por objetivo realizar una sola partición de los individuos en *K* grupos. Esto implica que el investigador debe especificar a priori los grupos que deben ser formados. Esta es, posiblemente, la principal diferencia respecto de los métodos jerárquicos. La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia está en que estos métodos trabajan con la matriz de datos original y no requieren su conversión en una matriz de proximidades. Pedret agrupa los métodos no jerárquicos en las cuatro familias siguientes: *reasignación, búsqueda de la densidad, directos y reducción de dimensiones.*

Los *métodos de reasignación* permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior si esto optimiza el criterio de selección. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Algunos de los algoritmos más conocidos dentro de estos métodos son el *método K-means (o K-medias)* de MeQueen (1967), el *Quick Cluster Analysis* y el *método de Forgy*, los cuales se suelen agrupar bajo el nombre de *métodos centroides o centros de gravedad*. Por otra parte está el *método de las nubes dinámicas*, debido a Diday.

Los *métodos de búsqueda de la densidad* presentan una aproximación tipológica y una aproximación probabilística. En la primera aproximación, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre los algoritmos más conocidos dentro de estos métodos están el *análisis modal de Wishart*, el *método de Taxmap de Carmichael y Sneath*, y el *método de Fortin*. En la segunda aproximación, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Destaca en esta aproximación el *método de las combinaciones de Wolf*.

Los *métodos directos* permiten clasificar simultáneamente a los individuos y a las variables. Las entidades agrupadas, ya no son los individuos o las variables, sino que son las observaciones, es decir, los cruces que configuran la matriz de datos.

© Alfaomega - RC Libros

Los *métodos de reducción de dimensiones*, como el análisis factorial de tipo Q, guardan relación con el análisis clúster. Este método consiste en buscar factores en el espacio de los individuos, correspondiendo cada factor a un grupo. La interpretación de los grupos puede ser compleja dado que cada individuo puede corresponder a varios factores diferentes.

Resulta muy intuitivo suponer que una clasificación correcta debe ser aquella en que la dispersión dentro de cada grupo formado sea la menor posible. Esta condición se denomina *criterio de varianza,* y lleva a seleccionar una configuración cuando la suma de las varianzas dentro de cada grupo (varianza residual) sea mínima.

Se han propuesto diversos algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza, que difieren en la elección de los clústeres provisionales que necesita el arranque del proceso y en el método de asignación de individuos a los grupos. Aquí se describen los dos más utilizados.

El *algoritmo de las H-medias* parte de una primera configuración arbitraria de grupos con su correspondiente media, eligiendo un primer individuo de arranque de cada grupo y asignando posteriormente cada caso al grupo cuya media es más cercana. Una vez que todos los casos han sido ubicados, calcula de nuevo las medias o centroides y las toma en lugar de los primeros individuos como una mejor aproximación de los mismos, repitiendo el proceso mientras la varianza residual vaya disminuyendo. La partición de arranque define el número de clústeres que, lógicamente, puede disminuir si ningún caso es asignado a alguno de ellos.

El *algoritmo de las K-medias*, el más importante desde los puntos de vista conceptual y práctico, parte también de unas medias arbitrarias y, mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los casos a cada uno de los grupos. El valor mínimo de varianza determina una configuración de nuevos grupos con sus respectivas medias. Se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia puede ya disminuir la varianza residual; o se alcance otro criterio de parada: un número limitado de pasos de iteración o, simplemente, que la diferencia obtenida entre los centroides de dos pasos consecutivos sea menor que un valor prefijado. El procedimiento configura los grupos maximizando, a su vez, la distancia entre sus centros de gravedad. Como la varianza total es fija, minimizar la residual hace máxima la factorial o intergrupos. Y puesto que minimizar la varianza residual es equivalente a conseguir que sea mínima la suma de distancias al cuadrado desde los

© Alfaomega - RC Libros

casos a la media del clúster al que van a ser asignados, es esta distancia euclídea al cuadrado la utilizada por el método.

Como se comprueban los casos secuencialmente para ver su influencia individual, el cálculo puede verse afectado por el orden de los mismos en la tabla; pese a lo cual es el algoritmo que mejores resultados produce. Otras variantes propuestas a este método llevan a clasificaciones muy similares.

Como cualquier otro método de clasificación no jerárquica, proporciona una solución final única para el número de clústeres elegido, a la que se llegará con menor número de iteraciones cuanto más cerca estén las "medias" de arranque de las que van a ser finalmente obtenidas. Los programas automáticos seleccionan generalmente estos primeros valores, tantos como grupos se pretenda formar, entre los puntos más separados de la nube. Los clústeres no jerárquicos están indicados para grandes tablas de datos, y son también útiles para la detección de casos atípicos: Si se elige previamente un número elevado de grupos, superior al deseado, aquellos que contengan muy escaso número de individuos servirían para detectar casos extremos que podrían distorsionar la configuración.

Es aconsejable realizar el análisis definitivo sin ellos, ya con el número deseado de grupos para después, opcionalmente, asignar los atípicos al clúster adecuado que habrá sido formado sin su influencia distorsionante. Un problema importante que tiene el investigador para clasificar sus datos en grupos es, como se ha dicho, la elección de un número adecuado de clústeres. Puesto que siempre será conveniente efectuar varios tanteos, la selección del más apropiado al fenómeno que se estudia ha de basarse en criterios tanto matemáticos como de interpretabilidad. Entre los primeros, se han definido numerosos indicadores de adecuación como el Criterio cúbico de clústeres y la Pseudo F que se describen en el ejemplo de aplicación práctica. El uso inteligente de estos criterios, combinado con la interpretabilidad práctica de los grupos, constituye el arte de la decisión en la clasificación multivariante de datos.

Matemáticamente, un método de clasificación no jerarquizado consiste en formar un número prefijado *K* de clases homogéneas excluyentes, pero con máxima divergencia entre las clases. Las *K* clases o clústeres forman una única partición (*clústering*) y no están organizadas jerárquicamente ni relacionadas entre sí. La clasificación no jerárquica o de reagrupamiento tiene una estructura matemática menos precisa que la clasificación jerárquica. El número de métodos existentes ha crecido excesivamente en los últimos años y algunos problemas derivados de su utilización todavía no han sido resueltos.

© Alfaomega - RC Libros

ANÁLISIS CLÚSTER JERÁRQUICO A TRAVÉS DE R

Para realizar el análisis clúster utilizamos la subopción *Factor analysis* de la opción *Dimensional analysis* del menú *Statistics* (Figura 12-1). En la pantalla *Data* (Figura 12-2) se seleccionan las variables para la clasificación. En nuestro caso queremos clasificar los barrios de Madrid por nivel de desarrollo y utilizaremos como variables de clasificación los tres factores hallados en el capítulo anterior. Utilizar reducción de la dimensión antes de la clasificación es una práctica habitual, ya que las variables iniciales de clasificación suelen estar correlacionadas, lo que introduce ruido en la clasificación. Como los factores son incorrelacionados e involucran a todas las variables iniciales, son buenas variables de clasificación. En la pantalla *Options* (Figura 12-3) se elige el método clúster y la distancia a utilizar. Al pulsar OK se obtiene el dendograma (Figura 12-4).

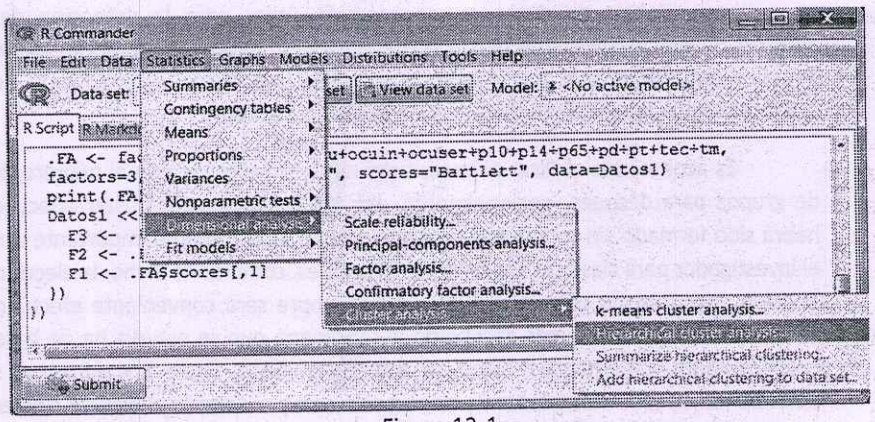


Figura 12-1

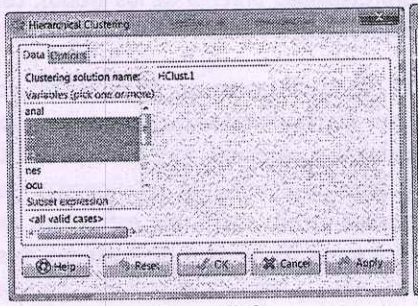


Figura 12-2

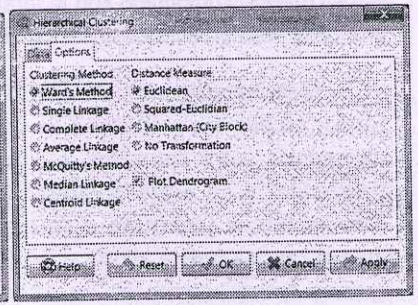


Figura 12-3

© Alfaomega - RC Libros

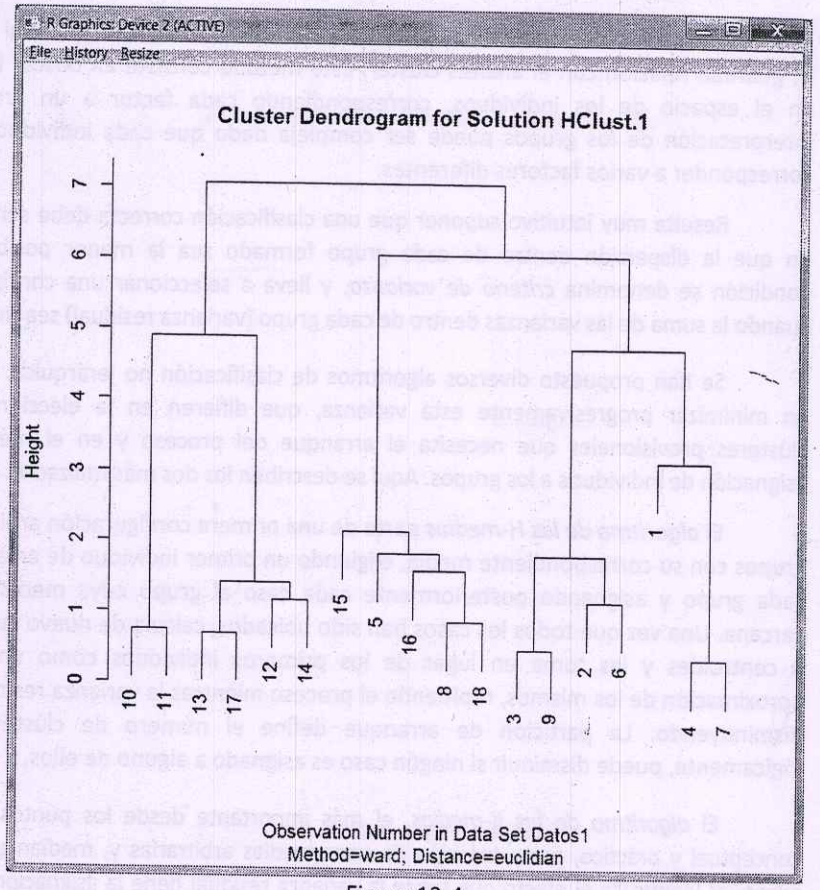


Figura 12-4

ANÁLISIS CLÚSTER A TRAVÉS DE COMANDOS

El comando *hclust* permite realizar análisis clúster mediante la siguiente sintaxis sencilla:

hclust(dist(model.matrix(~ -1+v1+v2+...+vn, labels = variable con etiquetas, data=Conjunto de datos, method=METODO)))

© Alfaomega - RC Libros

Las variables *v1, v2, ..., vn* son las variables de clasificación y *data=Conjunto de datos* permite declarar el conjunto de datos que contiene las variables.

En nuestro ejemplo anterior la sintaxis vía comandos sería la siguiente:

```
> HClust.1 <- hclust(dist(model.matrix(~-1 + F1+F2+F3,
Datos2)) , method= "ward")

RcmdrMsg: [8] NOTE: The "ward" method has been renamed to
"ward.D"; note new
RcmdrMsg+ "ward.D2"

> plot(HClust.1, main= "Clúster Dendrogram for Solution
HClust.1", xlab= "Observation Number in Data Set Datos2",
sub="Method=ward; Distance=euclidian")
```

Si queremos etiquetar los elementos del dendograma, utilizaremos la siguiente sintaxis:

```
> w=cbind("b")
> w
     [,1]
[1,] "b"
> attach(Datos2)
> b
 [1] "Centro          " "Arganzuela       " "Retiro          "
 [4] "Salamanca       " "Chamartín       " "Tetuán         "
 [7] "Chamberi        " "Fuencarral      " "Moncloa        "
[10] "Latina          " "Carabanchel     " "Villaverde     "
[13] "Mediodía        " "Vallecas        " "Moratalaz      "
[16] "Ciudad Lineal   " "San Blas        " "Hortaleza      "
> w=cbind(b)
> w
      b
[1,] "Centro          "
[2,] "Arganzuela      "
[3,] "Retiro          "
[4,] "Salamanca       "
[5,] "Chamartín       "
[6,] "Tetuán          "
[7,] "Chamberi        "
[8,] "Fuencarral      "
[9,] "Moncloa         "
[10,] "Latina         "
[11,] "Carabanchel    "
[12,] "Villaverde     "
```


© Alfaomega - RC Libros


```
[13,] "Mediodía       "
[14,] "Vallecas       "
[15,] "Moratalaz      "
[16,] "Ciudad Lineal  "
[17,] "San Blas       "
[18,] "Hortaleza      "

> plot(HClust.1, labels=w, main= "Clúster Dendrogram for
Solution HClust.1", xlab= "Observation Number in Data Set
Datos1", sub="Method=ward; Distance=euclidian")
```

La Figura 12-5 muestra el dendograma etiquetado.

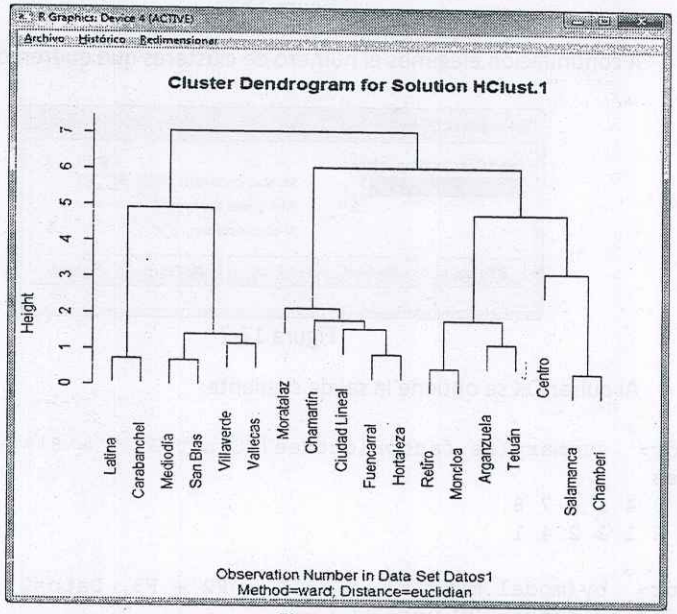


Figura 12-5

Para obtener un resumen del clúster jerárquico utilizamos la subopción *Summarize hierarchical clustering* de la opción *Cluster analysis* (Figura 12-6).

© Alfaomega - RC Libros

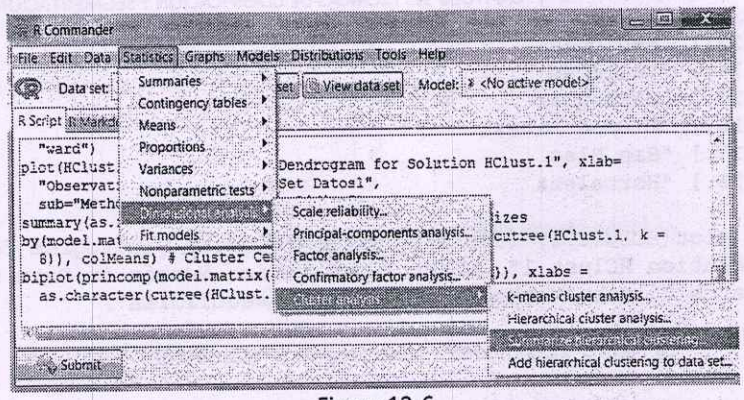


Figura 12-6

A continuación elegimos el número de clústeres que queremos (Figura 12-7).

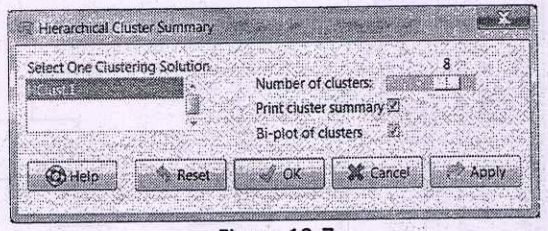


Figura 12-7

Al pulsar OK se obtiene la salida siguiente:

```
Rcmdr>  summary(as.factor(cutree(HClust.1, k = 8))) # Clúster
Sizes
1 2 3 4 5 6 7 8
1 4 2 1 3 2 4 1

Rcmdr>  by(model.matrix(~-1 + F1 + F2 + F3, Datos2),
as.factor(cutree(HClust.1, k =
Rcmdr+     8)), colMeans) # Clúster Centroids
INDICES: 1
       F1          F2          F3
-0.6490111 -0.5726294  2.3761608
------------------------------------------------------------
INDICES: 2
        F1           F2           F3
-1.13569530  0.07928450  0.01562548
------------------------------------------------------------
```


© Alfaomega - RC Libros


```
INDICES: 3
        F1          F2          F3
0.06328761 1.46683688 1.26588063
------------------------------------------------------------
INDICES: 4
        F1          F2          F3
-0.5186780  1.5354705 -0.7084447
------------------------------------------------------------
INDICES: 5
        F1          F2          F3
 0.3943539  0.3733664 -0.6823315
------------------------------------------------------------
INDICES: 6
        F1          F2          F3
 1.8542369 -0.3730532  0.6068786
------------------------------------------------------------
INDICES: 7
         F1           F2           F3
-0.06646374 -1.38006584 -0.33179303
------------------------------------------------------------
INDICES: 8
        F1          F2          F3
 0.9582145  0.9326177 -2.1015699
```

ANÁLISIS CLÚSTER NO JERÁRQUICO A TRAVÉS DE R

Para realizar el análisis clúster no jerárquico utilizamos la subopción *k-means cluster analysis* de la opción *Cluster analysis* del menú *Statistics* (Figura 12-8). En la pantalla *Data* (Figura 12-9) se seleccionan las variables para la clasificación. En nuestro caso queremos clasificar los barrios de Madrid por nivel de desarrollo y utilizaremos como variables de clasificación las tres componentes principales halladas en el capítulo anterior. Utilizar reducción de la dimensión antes de la clasificación es una práctica habitual, ya que las variables iniciales de clasificación suelen estar correlacionadas, lo que introduce ruido en la clasificación. Como los factores son incorrelacionados e involucran a todas las variables iniciales, son buenas variables de clasificación. En la pantalla *Options* (Figura 12-10) se elige el número de clúster y otras características. Al pulsar OK se obtiene la salida del análisis.

© Alfaomega - RC Libros

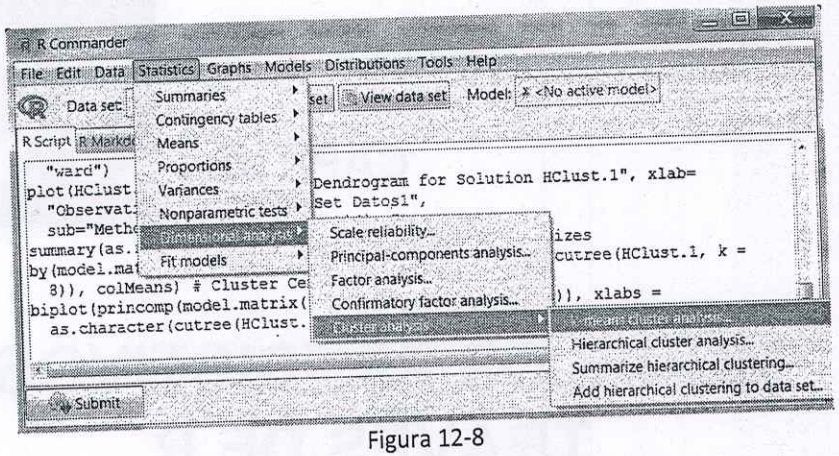


Figura 12-8

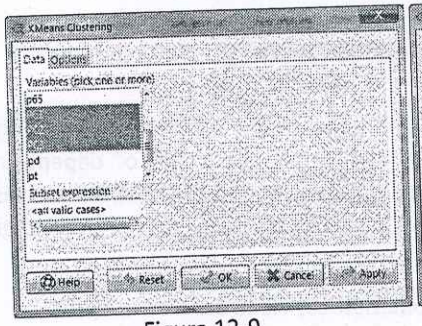


Figura 12-9

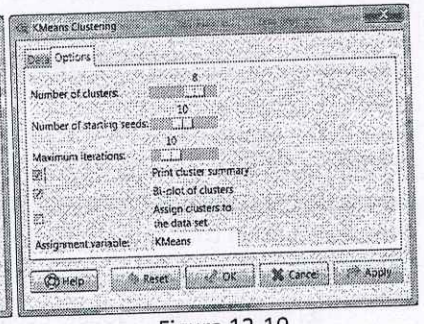


Figura 12-10

Se obtiene la salida siguiente:

```
Rcmdr>  .clúster <-  KMeans(model.matrix(~-1 + PC1 + PC2 + PC3,
Datos1), centers = 8, iter.max = 10, num.seeds = 10)

Rcmdr>  .clúster$size # Clúster Sizes
[1] 3 3 2 2 2 1 3 2

Rcmdr>  .clúster$centers # Clúster Centroids
  new.x.PC1  new.x.PC2  new.x.PC3
1 -10.05974  -4.175155 -16.099926
2 -75.19632  -1.476602  -2.773610
3  13.33436 -25.559254 -23.767149
4 -41.55870  12.802785  13.553581
5 -10.44316  21.382242   5.968028
```


© Alfaomega - RC Libros


```
6  12.79754 -48.929547  42.242424
7  16.78375  11.942930   8.496810
8 134.97720   6.402240  -1.310582

Rcmdr>  .clúster$withinss # Within Clúster Sum of Squares
[1] 451.34778 612.86303  13.63432 293.96326  58.26616   0.00000
406.41096
[8] 963.87100

Rcmdr>  .clúster$tot.withinss # Total Within Sum of Squares
[1] 2800.357

Rcmdr>  .clúster$betweenss # Between Clúster Sum of Squares
[1] 68626.71
RcmdrMsg: [9] WARNING: Warning in arrows(0, 0, y[, 1L] * 0.8,
y[, 2L] * 0.8, col
RcmdrMsg+ = col[2L], length = arrow.len) :
RcmdrMsg+ zero-length arrow is of indeterminate angle and so
skipped

Rcmdr>  biplot(princomp(model.matrix(~-1 + PC1 + PC2 + PC3,
Datos1)), xlabs =
Rcmdr+     as.character(.clúster$clúster))

Rcmdr>  remove(.clúster)
```

Ejercicio 12-1. El archivo mundo.sav contiene información sobre variables demográficas y de desarrollo de varios países del mundo altamente correlacionadas. Se trata de reducir el número inicial de variables correlacionadas a un número menor de variables incorrelacionadas con la mínima pérdida de información. Con las variables reducidas realizar una segmentación clúster de los países del mundo según su nivel de desarrollo.

La primera parte de este ejercicio ya se realizó en el ejercicio del capítulo anterior. Utilizaremos las 5 componentes obtenidas como variables de clasificación en el análisis clúster COMP1, COMP2, COMP3, COMP4 y COMP5 y las etiquetas *v* de los países. La sintaxis de R para clúster jerárquico sería la siguiente:

```
Clúster.1 <- hclust(dist(model.matrix(~-1 +
COMP1+COMP2+COMP3+COMP4+COMP5, mundo1)) , method= "ward")
```

Ahora se representa el dendograma (Figura 12-11) mediante la siguiente sintaxis:

© Alfaomega - RC Libros

```
> windows()
> plot(Clúster.1, labels=v, main= "Dendograma", xlab= "Número
de observación en mundo1", sub="Método=ward;
Distance=euclidian")
```

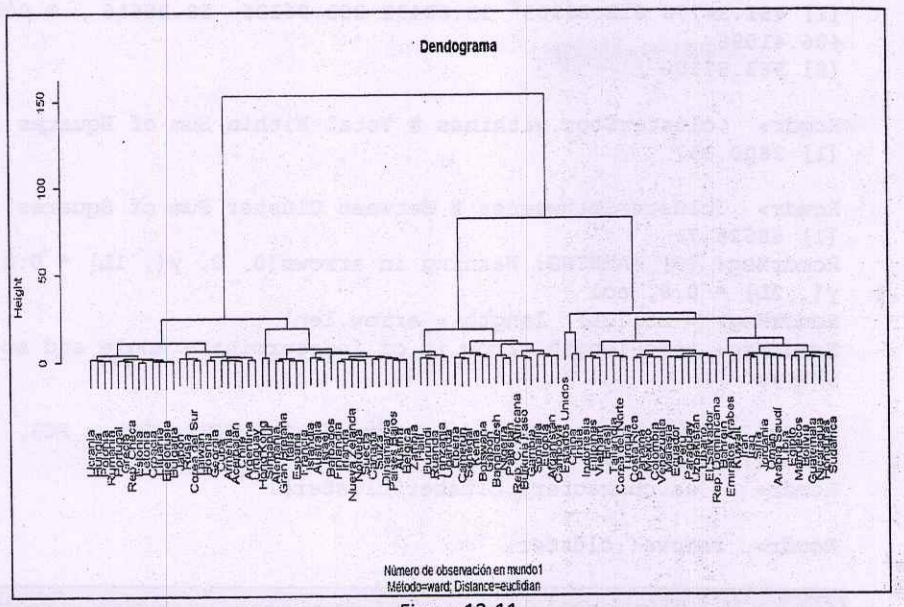


Figura 12-11

© Alfaomega - RC Libros

CAPÍTULO 13

MODELOS PREDICTIVOS A TRAVÉS DE R

MODELO LINEAL DE REGRESIÓN MÚLTIPLE

El modelo de regresión múltiple tiene como objetivo explicar el comportamiento de una variable endógena, explicada o dependiente, que designaremos como Y, utilizando la información proporcionada por los valores tomados por un conjunto de variables explicativas, exógenas o independientes, que designaremos por $X_1, X_2, ..., X_k$.

El modelo lineal (modelo econométrico) viene dado de la forma:

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+....+\beta_k X_k+u$$

Los coeficientes (parámetros) $\beta_1, \beta_2, ..., \beta_k$ denotan la magnitud del efecto que las variables explicativas (exógenas o independientes) $X_1, X_2, ..., X_k$ tienen sobre la variable explicada (endógena o dependiente) Y. El coeficiente β_0 se denomina término constante (o independiente) del modelo. El término ε se denomina término de error del modelo.

Disponemos de un conjunto de T observaciones para cada una de las variables endógena y exógenas, una por cada período temporal. Entonces, podremos escribir el modelo de la forma:

$$Y_t=\beta_0+\beta_1 X_{1t}+\beta_2 X_{2t}+....+\beta_k X_{kt}+u_t \quad t=1,2,3,...,T$$

© Alfaomega - RC Libros

La aparición (no necesaria) de un término independiente en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor sea siempre 1.

El problema fundamental que se aborda es el siguiente: suponiendo que la relación entre la variable Y y el conjunto de variables $X_1, X_2, \ldots, X_k$ es como se ha descrito en el modelo, y que se dispone de un conjunto de T observaciones en el tiempo para cada una de las variables, la endógena y las exógenas, ¿cómo pueden asignarse valores numéricos a los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ basándonos en la información muestral? Estos valores se llamarán estimaciones de los parámetros.

Una vez encontradas las estimaciones de los parámetros del modelo, podremos hacer predicciones acerca del comportamiento futuro de la variable Y.

HIPÓTESIS EN EL MODELO LINEAL

Las hipótesis básicas que debe cumplir cualquier modelo lineal pueden clasificarse en cuatro grandes grupos según las componentes del propio modelo. El primer grupo de hipótesis lo constituyen las relativas a la perturbación aleatoria u. El segundo grupo de hipótesis lo forman las relativas a los regresores. Un tercer grupo de hipótesis lo forman las relativas a la forma funcional. Un último grupo lo forman las hipótesis relativas al vector de parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$.

Hipótesis relativas a la perturbación aleatoria

Formulamos el modelo lineal bajo las siguientes hipótesis clásicas relativas a la perturbación aleatoria u.

- La variable u (término de error) es una variable aleatoria con esperanza nula y matriz de covarianzas constante y diagonal (matriz escalar). Es decir que, para todo instante temporal t, la variable ε_t tiene media cero y varianza σ^2 no dependiente de t, y además $Cov(u_i, u_j) = 0$ para todo i y para todo j instantes temporales distintos entre sí. El hecho de que la varianza de ε_t sea constante para todo t (que no dependa de t), se denomina hipótesis de *homoscedasticidad* y puede también expresarse como $V(u/X_1,X_2, \ldots, X_k) = \sigma^2$ y $V(Y/X_1,X_2, \ldots, X_k) = \sigma^2$. El hecho de que $Cov(u_i, u_j) = 0$ para todo instante temporal i distinto de j se denomina hipótesis de *no autocorrelación*.

© Alfaomega - RC Libros

- El término de error u es una variable aleatoria no observable, lo que implica que la variable Y es aleatoria, ya que depende de la variable aleatoria u.
- También se considera la hipótesis de *normalidad* de los residuos, consistente en que las variables ε_t sean normales para todo t. Dicho de otro modo, el vector de perturbaciones aleatorias del modelo tiene una distribución normal multivariante de media cero $E(\boldsymbol{u}) = \mathbf{0}$ y matriz de varianzas covarianzas escalar $E(\boldsymbol{uu}') = \sigma^2 I$. Podemos escribir $\boldsymbol{u} \rightarrow \boldsymbol{N}(\mathbf{0}, \sigma^2 I)$.

Hipótesis relativas a los regresores

Se tendrán en cuenta las siguientes hipótesis relativas a los regresores (variables independientes o exógenas) del modelo:

- Las variables $X_1, X_2, \ldots, X_k$ son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de *independencia*, y cuando no se cumple, decimos que el modelo presenta *multicolinealidad*. Es equivalente decir que la matriz de regresores tiene rango k.
- Las variables $X_1, X_2, \ldots, X_k$ son deterministas (no son variables aleatorias), ya que su valor es constante proveniente de una muestra tomada en el tiempo, y además no están correlacionadas con el término de error u, es decir, $E(u/X_1,X_2, \ldots, X_k) = 0$ (hipótesis de *exogeneidad*).
- Los regresores no tienen errores de observación o de medida.

Hipótesis relativas al vector de parámetros

- El vector de parámetros β es un vector fijo.

Esta hipótesis asegura la estabilidad en el tiempo de las estimaciones. El estudio profundo de esta hipótesis nos llevará a la teoría de la cointegración, que es un tema que se tratará posteriormente.

Hipótesis relativas a la forma funcional

- La relación entre Y y $X_1, X_2, \ldots, X_k$ es efectivamente lineal (hipótesis de *linealidad*).

© Alfaomega - RC Libros

- También se supone la ausencia de errores de especificación, es decir, que suponemos que todas las variables *X* que son relevantes para la explicación de la variable *Y*, están incluidas en la definición del modelo lineal.

ESTIMACIÓN DEL MODELO LINEAL POR MÍNIMOS CUADRADOS ORDINARIOS MCO

Supongamos que queremos ajustar el modelo de regresión lineal múltiple siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Disponemos de un conjunto de *T* observaciones para cada una de las variables endógena y exógenas (una para cada instante temporal). Entonces, podremos escribir el modelo de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \quad t = 1,2,3,\ldots,T$$

El criterio de mínimos cuadrados ordinarios MCO considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error *u*, lo que es equivalente a minimizar:

$$S(\beta_0,\beta_1,\cdots,\beta_k) = \sum_{t=1}^{T} u_t^2 = \sum_{t=1}^{T} (y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))^2$$

Derivando respecto de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$, e igualando a cero tenemos:

$$\left.\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= 2\sum_{t=1}^{T}(y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \phi_k x_{kt}))(-1) = 0 \\
\frac{\partial S}{\partial \beta_1} &= 2\sum_{t=1}^{T}(y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))(-x_{1t}) = 0 \\
&\vdots \\
\frac{\partial S}{\partial \beta_k} &= 2\sum_{t=1}^{T}(y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))(-x_{kt}) = 0
\end{aligned}\right\} \Rightarrow$$

© Alfaomega - RC Libros

$$\Rightarrow \begin{cases}
\sum_{t=1}^{T} y_t = T\beta_0 + \beta_1 \sum_{t=1}^{T} x_{1t} + \cdots + \beta_k \sum_{t=1}^{T} x_{kt} \\
\sum_{t=1}^{T} y_t x_{1t} = \beta_0 \sum_{t=1}^{T} x_{1t} + \beta_1 \sum_{t=1}^{T} x_{1t}^2 + \cdots + \beta_k \sum_{t=1}^{T} x_{1t} x_{kt} \\
\vdots \\
\sum_{t=1}^{T} y_t x_{kt} = \beta_0 \sum_{t=1}^{T} x_{kt} + \beta_1 \sum_{t=1}^{T} x_{kt} x_{1t} + \cdots + \beta_k \sum_{t=1}^{T} x_{kt}^2
\end{cases}$$

Estas ecuaciones forman un sistema denominado sistema de ecuaciones normales, que puede resolverse para $\beta_0, \beta_1, \ldots, \beta_k$ mediante cualquier método apropiado para resolver sistemas de ecuaciones lineales. De esta forma se obtiene la estimación del modelo.

Alternativamente, podemos partir de la notación matricial del modelo $y = X\beta + u$, donde *y* es el vector $T \times 1$, *X* es una matriz $T \times k$, β es un vector $k \times 1$ y *u* es un vector $T \times 1$.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1T} & X_{2T} & \cdots & X_{kT} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_t \end{pmatrix}$$

Abreviadamente podemos poner: $Y = X\beta + u$

Denotando el modelo estimado por $\hat{y} = X\hat{\beta}$, el vector de residuos se puede expresar, por tanto, como $\hat{u} = y - \hat{y} = y - X\hat{\beta}$. Si designamos con *S* a la suma de los cuadrados de los residuos, tendremos:

$$S = \hat{u}'\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_T) \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \sum_{t=1}^{T} \hat{u}_t^2$$

© Alfaomega - RC Libros

Que también puede escribirse como:

$$S=(y-X\hat{\beta})'(y-X\hat{\beta})=y'y-\hat{\beta}'X'y-y'X\hat{\beta}+\hat{\beta}'X'X\hat{\beta}=y'y-2\hat{\beta}'X'y+\hat{\beta}'X'X\hat{\beta}$$

ya que $\hat{\beta}'X'y=y'X\hat{\beta}$ (se trata del mismo escalar en ambas expresiones).

Aplicar el criterio de mínimos cuadrados es equivalente a minimizar el escalar *S*. Para minimizar *S* se calcula la primera derivada de *S* con respecto al vector de coeficientes mínimo cuadráticos, $\hat{\beta}$, y se tiene:

$$\frac{\delta S}{\delta\hat{\beta}}=-2X'y+2X'X\hat{\beta}$$

Al igualar a cero se obtiene el sistema de ecuaciones normales:

$$X'X\hat{\beta}=X'y$$

Para poder resolver el sistema respecto a $\hat{\beta}$ unívocamente, se debe cumplir que el rango de la matriz *X'X* sea igual a *k*. Si se cumple esta condición, se pueden premultiplicar ambos miembros del sistema por $[X'X]^{-1}$:

$$[X'X]^{-1}[X'X]\hat{\beta}=[X'X]^{-1}X'y\Rightarrow\hat{\beta}=[X'X]^{-1}X'y$$

Ya tenemos la expresión del vector de estimadores mínimo cuadráticos:

$$\hat{\beta}=[X'X]^{-1}X'y$$

Además, *S* presenta un mínimo en $\hat{\beta}$, ya que la matriz de segundas derivadas, 2*X'X*, es definida positiva.

ANÁLISIS DE LOS RESIDUOS

Una vez construido el modelo de regresión con series temporales, tendremos que contrastar, entre otras, las hipótesis de linealidad, normalidad, homoscedasticidad,

© Alfaomega - RC Libros

no autocorrelación e independencia. Los residuos van a presentar una primera información sobre estas hipótesis.

Si el histograma de frecuencias de los residuos no se ajusta al de una normal, pueden existir valores atípicos. Eliminando los pares (X_i Y_i) que producen los valores atípicos, se puede conseguir *normalidad* en los residuos.

Si graficamos los valores de *t* contra los valores de $\hat{u}_t$ (o sea, si hacemos la gráfica cuyos puntos son los pares (t, $\hat{u}_t$) y detectamos una tendencia creciente o decreciente en el grafo) puede existir *autocorrelación o correlación serial*.

Si graficamos los valores de $\hat{Y}_t$ contra los valores de $\hat{u}_t$, o sea, si hacemos la gráfica cuyos puntos son los pares ($\hat{Y}_t,\hat{u}_t$) y detectamos una tendencia de cualquier tipo en el grafo, puede existir autocorrelación, ya que habrá correlación entre los residuos. También puede haber en este caso *heteroscedasticidad,* o también falta de *linealidad*.

Si graficamos los valores de $\hat{Y}_t$ contra los valores de $\hat{u}_t^2$, o sea, si se hace la gráfica cuyos puntos son los pares ($\hat{Y}_t,\hat{u}_t^2$) y detectamos una tendencia de cualquier tipo en el grafo, puede existir heteroscedasticidad.

Si graficamos los valores de X_t contra los valores de $\hat{u}_t$, o sea, si se hace la gráfica cuyos puntos son los pares (X_t, $\hat{u}_t$) y detectamos una tendencia creciente o decreciente en el grafo, puede existir autocorrelación, ya que los residuos no estarán incorrelados con las variables explicativas. También puede haber heteroscedasticidad, o falta de linealidad.

Si graficamos los valores de X_t contra los valores de $\hat{u}_t^2$, o sea, si se hace la gráfica cuyos puntos son los pares (X_t, $\hat{u}_t^2$) y detectamos cualquier tendencia en el grafo, puede existir heteroscedasticidad o falta de linealidad (habrá relación entre la varianza del término del error y las variables explicativas).

Estos análisis pueden realizarse también utilizando residuos estandarizados o residuos estudentizados, que suelen ser más efectivos para detectar deficiencias en el modelo.

© Alfaomega - RC Libros

Los residuos estudentizados, cuya distribución es una *t* de Student con *T–k–2* grados de libertad, se usan también para detectar valores atípicos en los residuos (análisis de la normalidad de los residuos o de la mala especificación del modelo).

Un contraste muy importante para detectar la autocorrelación es el contraste de Durbin-Watson, según el cual el estadístico:

$$d=\frac{\sum_{t=2}^{T}(\hat{u}_t-\hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^2}$$

nos permite adoptar la regla no demasiado rigurosa de que si d vale 0 hay autocorrelación perfecta positiva; si *d* se aproxima a 2 no hay autocorrelación, y si *d* se aproxima a 4 hay autocorrelación perfecta negativa. No obstante, *d* se encuentra tabulado, y según la franja en la que caiga su valor, se acepta o rechaza la hipótesis de autocorrelación.

Si la matriz *X'X* tiene determinante cero, no podrá calcularse su inversa (ni podrán hacerse las estimaciones de los parámetros), en cuyo caso hay multicolinealidad. Este problema suele resolverse sustituyendo las variables del modelo por un conjunto grande de sus componentes principales para hacer de nuevo la regresión, pero esta vez con las componentes principales como variables independientes del modelo. La multicolinealidad también puede abordarse mediante la matriz de correlaciones de las variables independientes, ya que las variables con correlación alta pueden ser candidatas a provocar multicolinealidad, problema que puede intentar solucionarse eliminando una de las variables fuertemente correlacionadas, siempre que dicha variable no sea relevante.

EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE EN R

Es posible obtener la siguiente información relativa a un modelo:

© Alfaomega - RC Libros

r=lm(y~x1+x2+x3+....)	*Efectúa la regresión lineal por mínimos cuadrados con variables independientes x_i y con variable dependiente y*
summary(r)	*Muestra resultados detallados del modelo ajustado*
coefficients(r)	*Lista los parámetros ajustados del modelo*
confint(r)	*Lista los intervalos de confianza al 95% para los parámetros estimados del modelo*
fitted(r)	*Lista los valores predichos en el modelo ajustado*
residuals(r)	*Lista los residuos del modelo ajustado*
anova(r)	*Lista la tabla ANOVA del modelo ajustado*
vcov(r)	*Lista la matriz de covarianzas de los parámetros estimados del modelo*
AIC(r)	*Devuelve el estadístico de AKAIKE*
plot(r)	*Muestra diagnósticos gráficos para el modelo ajustado*
predict(r)	*Predice valores con el modelo ajustado*

Como ejemplo vamos a considerar la regresión lineal que haga depender el consumo de los automóviles de su potencia, su peso y su motor.

```
> r1=lm(consumo~cv+peso+motor)
> r1

Call:
lm(formula = consumo ~ cv + peso + motor)

Coefficients:
(Intercept)           cv         peso        motor
  0.8762787    0.0412840    0.0051719    0.0003043

> summary(r1)

Call:
lm(formula = consumo ~ cv + peso + motor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1803 -1.1196 -0.0541  0.8530 20.0023
```

© Alfaomega - RC Libros

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.8762787  0.5427383   1.615    0.107
cv          0.0412840  0.0057376   7.195 3.07e-12 ***
peso        0.0051719  0.0009563   5.408 1.10e-07 ***
motor       0.0003043  0.0001819   1.674    0.095 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.938 on 402 degrees of freedom
Multiple R-squared:  0.7609,   Adjusted R-squared:  0.7591
F-statistic: 426.3 on 3 and 402 DF,  p-value: < 2.2e-16
```

Observamos que el R^2 ajustado es bastante alto, los *p*-valores de la T de Student para los parámetros estimados son lo suficientemente bajos para que todos sean significativos individualmente al menos al 90%. Además, como el *p*-valor de la F de Fisher es muy pequeño, también se cumple la significatividad conjunta de los parámetros estimados.

Para obtener intervalos de confianza al 95% para los parámetros estimados del modelo utilizamos la siguiente sintaxis:

```
> confint(r1)
                  2.5 %       97.5 %
(Intercept) -1.906812e-01 1.9432385272
cv           3.000464e-02 0.0525633722
peso         3.291844e-03 0.0070519317
motor       -5.316155e-05 0.0006618361
```

Si queremos guardar los residuos del modelo en una variable utilizamos la siguiente sintaxis:

```
> residuos=residuals(r1)
```

Si queremos guardar los valores predichos en una variable utilizamos la siguiente sintaxis:

```
> predichos=fitted(r1)
```

Si queremos ver la tabla ANOVA del modelo utilizamos la siguiente sintaxis:

```
> anova(r1)
```


© Alfaomega - RC Libros


```
Analysis of Variance Table

Response: consumo
          Df Sum Sq Mean Sq   F value Pr(>F)
cv         1 4454.1  4454.1 1186.0506 <2e-16 ***
peso       1  338.7   338.7   90.1766 <2e-16 ***
motor      1   10.5    10.5    2.8008  0.095 .
Residuals 402 1509.7     3.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esta tabla suele utilizarse para probar la significatividad conjunta de los parámetros del modelo a través del *p*-valor de la F de Fisher (debe ser pequeño).

Para obtener la matriz de covarianzas de los parámetros estimados utilizamos la siguiente sintaxis:

```
> vcov(r1)
            (Intercept)            cv          peso         motor
(Intercept)  2.945649e-01 -9.591434e-04 -4.275433e-04  7.482507e-05
cv          -9.591434e-04  3.291957e-05 -7.696842e-07 -5.401682e-07
peso        -4.275433e-04 -7.696842e-07  9.145781e-07 -1.248696e-07
motor        7.482507e-05 -5.401682e-07 -1.248696e-07  3.306999e-08
```

Como tiene valores pequeños el modelo está bien estimado.

A continuación evaluamos la capacidad predictiva del modelo con el estadístico de AIC de Akaike.

```
> AIC(r1)
[1] 1695.379
```

Un modelo predice mejor que otro cuando su estadístico AIC es menor.

Ahora generamos diagnósticos gráficos para realizar la diagnosis del modelo mediante la siguiente sintaxis. Obtenemos el gráfico de los residuos contra los valores predichos (como es aleatorio no hay problemas de heteroscedasticidad ni linealidad). La Figura 13-1 muestra un valor atípico. La Figura 13-2 muestra el gráfico QQ de normalidad (se ajusta bien a la diagonal del primer cuadrante salvo el atípico). La Figura 13-3 muestra los valores predichos contra los residuos estudentizados (su aleatoriedad refuerza la linealidad y la homoscedasticidad). La Figura 13-4 muestra los residuos atípicos resultando influyente la observación 35 (se sale de las bandas del estadístico de Cook). Lo mejor sería eliminar esta observación y volver a reestimar el modelo.

© Alfaomega - RC Libros

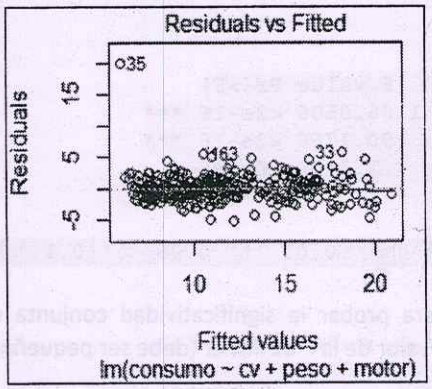


Figura 13-1

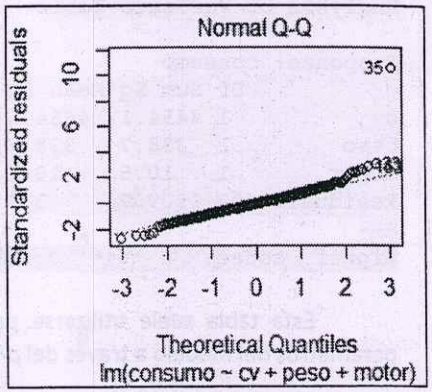


Figura 13-2

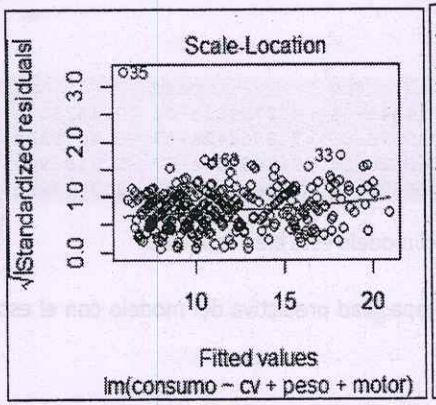


Figura 13-3

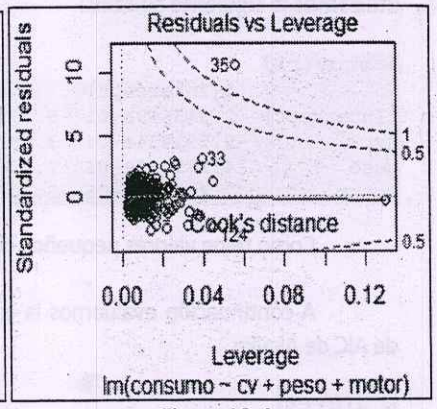


Figura 13-4

Si realizamos el histograma de los residuos, veremos que se ajusta a una normal (Figura 13-5). La gráfica de los residuos estimados contra el número de fila es aleatoria (Figura 13-6).

```
> hist(residuos,xlim=c(-10,10))
```


© Alfaomega - RC Libros

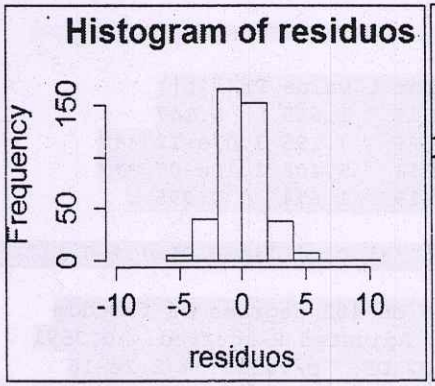


Figura 13-5

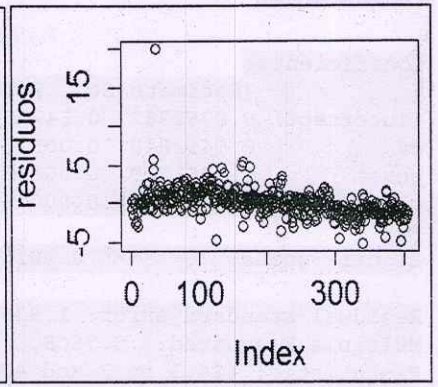


Figura 13-6

REGRESIÓN POLINÓMICA

En R se puede realizar la regresión polinómica mediante la siguiente sintaxis:

lm(y~x+I(x^2)+I(x^3)+...)

Como ejemplo hacemos la regresión del consumo contra las tres primeras potencias de la potencia.

```
> rp=lm(consumo~cv+I(cv^2)+I(cv^3))
> summary(rp)

Call:
lm(formula = consumo ~ cv + I(cv^2) + I(cv^3))

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2530 -1.2606 -0.1914  1.1002 15.7371 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.880e+00  2.094e+00   2.330  0.02027 *
cv          -1.651e-02  5.422e-02  -0.304  0.76093
I(cv^2)      1.126e-03  4.368e-04   2.579  0.01027 *
I(cv^3)     -3.510e-06  1.099e-06  -3.195  0.00151 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

© Alfaomega - RC Libros

```
Residual standard error: 2.073 on 402 degrees of freedom
Multiple R-squared:  0.7263,   Adjusted R-squared:  0.7242
F-statistic: 355.6 on 3 and 402 DF,  p-value: < 2.2e-16
```

REGRESIÓN CON INTERACCIONES

En R se puede realizar la regresión que contiene interacciones de las variables mediante la siguiente sintaxis:

lm(y~x+z+x:z+...)

Como ejemplo hacemos una regresión del consumo contra la potencia, el motor y la interacción de potencia y motor.

```
> rin=lm(consumo~cv+motor+cv:motor)
> summary(rin)

Call:
lm(formula = consumo ~ cv + motor + cv:motor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9681 -1.1109 -0.0549  0.9243 18.2565

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.439e-01  7.401e-01   0.194    0.846
cv           8.097e-02  9.466e-03   8.553 2.54e-16 ***
motor        1.732e-03  1.999e-04   8.663  < 2e-16 ***
cv:motor    -7.239e-06  1.544e-06  -4.689 3.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.954 on 402 degrees of freedom
Multiple R-squared:  0.7568,   Adjusted R-squared:  0.7549
F-statistic: 416.9 on 3 and 402 DF,  p-value: < 2.2e-16
```

REGRESIÓN SIN CONSTANTE

La regresión sin constante se realiza situando como primera variable independiente la variable cero. La sintaxis será la siguiente:

r=lm(y~0+x1+x2+x3+....)

© Alfaomega - RC Libros

Como ejemplo vamos a ajustar el modelo de regresión del ejemplo anterior sin constante, ya que la constante no resultó significativa por tener un valor muy alto. La sintaxis será la siguiente:

```
> rin=lm(consumo~0+cv+motor+cv:motor)
> summary(rin)

Call:
lm(formula = consumo ~ 0 + cv + motor + cv:motor)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0099 -1.1064 -0.0644  0.9205 18.2441

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
cv          8.265e-02  3.833e-03   21.56   <2e-16 ***
motor       1.755e-03  1.619e-04   10.84   <2e-16 ***
cv:motor   -7.512e-06  6.467e-07  -11.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.952 on 403 degrees of freedom
Multiple R-squared:  0.9735,   Adjusted R-squared:  0.9733
F-statistic:  4932 on 3 and 403 DF,  p-value: < 2.2e-16
```

Observamos que la significatividad individual de todos los parámetros estimados es muy buena (*p*-valores de la T muy bajos) y la significatividad conjunta también es muy buena (*p*-valor de la F muy pequeño).

© Alfaomega - RC Libros

ÍNDICE ANALÍTICO

A

Algoritmos de cálculo numerico 128
Análisis clúster 395
Análisis de componentes principales 359
Análisis de los residuos 416
Análisis en componentes principales 359, 360, 362, 364, 368, 369, 375, 376, 380
Análisis exploratorio de datos 147
Análisis exploratorio de datos a través de R 147
Análisis exploratorio formal 195
Análisis exploratorio y gráfico de los datos 147
Análisis factorial..... .. 359, 367, 368, 369, 370, 371, 372, 374, 383, 385, 396, 400
Análisis factorial a través de comandos 385
Análisis factorial a través de R..... 383
Archivos en formato ASCII.....9
Archivos en formato r.....7
Arrays 65
Autocorrelación..... 364, 412, 416, 417, 418

B

Break 126
Bucle For 117
Bucle While..... 119

C

Características de las distribuciones continuas 242
Círculos de correlación 365, 366
Clúster a través de comandos..... 403
Clúster jerárquico a través de R 402
Clúster no jerárquico 398
Clúster no jerárquico a través de R.....407
Clústeres jerárquicos 397
Coeficiente absoluto de asimetría.....208
Coeficiente de asimetría de Bowley.....208
Coeficiente de asimetría de Flsher 206, 207
Coeficiente de asimetría de Pearson.....208
Coeficiente de confianza 153, 276, 285, 307
Coeficiente de contingencia..... 290, 292, 293
Coeficiente de correlación 106, 115, 116, 157, 209, 210, 211, 212, 215, 220, 222, 263, 290, 291, 299, 300, 302, 319, 365
Coeficiente de correlación lineal 209, 212
Coeficiente de correlación lineal entre dos variables.....209
Coeficiente de correlación por rangos 212
Coeficiente de curtosis 208, 209
Coeficiente de determinación 211
Coeficiente de variación de Pearson 205
Coeficientes de regresión.....210
Comando contour.....342
Comando curve.....336
Comando filled.contour 343
Comando image.....344
Comando persp 340, 341, 342
Comando plot 329
Comandos para gráficos exploratorios en R 161, 171
Comparación de dos poblaciones normales con datos apareados 281
Componentes principales 375
Componentes principales a retener.....364
Componentes principales a través de comandos.....380
Componentes principales a través de R.....375

Comunalidad 365, 366, 368, 369
Condicional if else 121
Condicional ifelse 124
Contrastar la igualdad de varianzas..... 285
Contraste chi-cuadrado de independencia291
Contraste de Kolmogorov-Smirnov de la bondad de ajuste 296
Contraste de la igualdad de medias 287
Contraste de la igualdad de proporciones 289
Contrastes de aleatoriedad 298
Contrastes de hipótesis para poblaciones normales 279
Contrastes de la bondad de ajuste 294
Contrastes en el modelo factorial..... 370
Contrastes múltiples de homogeneidad... 293
Contrastes no paramétricos..... 289, 300
Control de flujo..... 117
Covarianza 212
Cuadrado de la contingencia..... 292
Cuantiles 199
Cuantiles, percentiles y momentos 195
Cuasivarianza, desviación típica y error estándar 203
Curvas planas en coordenadas explícitas.. 333
Curvas planas en coordenadas paramétricas..... 333
Curvas planas en coordenadas polares..... 339

D

Data frame.....72
Definición de funciones 93
Definir funciones a medida 93
Dendograma en el análisis clúster jerárquico..... 398
Densidad de probabilidad 240
Desigualdad de Tchebichev..... 248
Desviación típica y error estándar 203
Desviaciones medias..... 203
Diagrama de tallo y hojas 150
Distribución beta..... 247, 260
Distribución binomial 232, 233, 234, 237, 239, 244, 250, 255, 302, 303, 321, 322
Distribución chi-cuadrado..... 247, 251, 371
Distribución de Bernoulli.....232
Distribución de Poisson 112, 233, 234, 238, 246, 254, 259, 321
Distribución de probabilidad de tipo discreto 227
Distribución de Weibull.....248
Distribución exponencial..... 246, 251, 264
Distribución f de Fisher Snedecor..... 247, 277
Distribución gamma 246
Distribución hipergeométrica 233, 256
Distribución hipergeométrica generalizada 235
Distribución log-normal.....248
Distribución multinomial.....234
Distribución normal 250
Distribución t de Student 247, 252, 277
Distribución uniforme 250, 253
Distribución uniforme discreta 236
Distribuciones bidimensionales de probabilidad 230, 243
Distribuciones binomial, de Poisson, geométrica e hipergeométrica 231
Distribuciones continuas..... 244, 250
Distribuciones de frecuencias... 158, 195, 198, 213, 229, 241, 242
Distribuciones de más de dos dimensiones 213
Distribuciones de probabilidad continuas .241
Distribuciones de Weibull y Erlang.....248
Distribuciones discretas de una y dos dimensiones 229
Distribuciones discretas y continuas 225

E

Editor de R.....94
El lenguaje R y los intervalos de confianza y contrastes de hipótesis.....284
Elementos de variables matriciales.....29
Elementos de variables vectoriales 22
Elementos del lenguaje de R.....15
Entorno de trabajo de R 2

© Alfaomega - RC Libros

Entorno de trabajo de R Studio.....5
Error estándar 203, 204
Escalamiento óptimo 359
Esperanza matemática de una variable aleatoria 229
Estadística básica.....4
Estadística descriptiva..... 195, 215
Estadística descriptiva con R 195
Estimación del modelo lineal por mínimos cuadrados ordinarios MCO 413
Estimación por intervalos 275
Estructura ifelse..... 124
Estructura if-else 121
Estructura repeat 127
Estructura switch..... 125
Estructuras vectoriales y matriciales 61
Exogeneidad 413
Exploración formal 148
Exploración visual..... 148, 158
Exploratorio gráfico..... 147
Exportando datos.....7

F

Factores 26
For 118, 122, 135
Fórmula de Foster 201
Función de densidad. 161, 170, 240, 241, 243, 244, 246, 250, 251, 252, 253, 260, 261
Función de distribución 110, 111, 228, 229, 230, 231, 237, 238, 239, 240, 241, 243, 244, 246, 250, 251, 252, 253, 257, 258, 260, 261, 277, 296, 300
Función de distribución conjunta 230, 231, 243, 244
Función de fiabilidad..... 241
Función de riesgo acumulada 242
Función de riesgo promedio 242
Funciones básicas de r para estadística descriptiva..... 215
Funciones de análisis estadístico básico con argumento matricial 103
Funciones de cadena 44
Funciones de densidad condicionadas 243, 244
Funciones de densidad marginales... 243, 244
Funciones de distribución condicionadas.231, 243
Funciones de distribución marginales.....230, 231, 243, 244
Funciones de información.....42
Funciones de selección y manipulación.....47
Funciones de variable entera.....16
Funciones de variable real 16
Funciones de variables matriciales.....34
Funciones de variables vectoriales.....24
Funciones específicas de variable numérica 19
Funciones exponenciales y logarítmicas.....18
Funciones exponenciales, logarítmicas y otras funciones de variable vectorial.....25, 35
Funciones hiperbólicas.....18
Funciones hiperbólicas y sus inversas.....18
Funciones matemáticas predefinidas en r.101
Funciones matemáticas elementales de variable real.....101
Funciones matriciales específicas.....32
Funciones predefinidas 93
Funciones trigonométricas 17
Funciones trigonométricas y sus inversas....17

G

Gráfico de caja y bigotes 151
Gráfico de dispersión..... 156, 161
Gráfico de sedimentación..... 365, 376, 378
Gráfico de simetría y normalidad 156
Gráfico múltiple de caja y bigotes 153
Graficos básicos en R.....329
Gráficos de barras.....175
Gráficos de caja y bigotes.....171
Gráficos de puntos.....184
Gráficos de sectores 179
Gráficos de secuencia y dispersión.....329
Gráficos de superficies 340
Gráficos para variables cualitativas 158

© Alfaomega - RC Libros

H

Heteroscedasticidad 417, 421
Hipótesis en el modelo lineal 412
Histograma de frecuencias..... 148
Histogramas..... 181
Homegeneidad de poblaciones 301
Homoscedasticidad.....412, 416, 421

I

Ifelse 138
Importación de archivos de stata 84
Importación de archivos Excel 80
Importación de archivos SPSS 81
Importación de datos estadísticos como data frames 78
Importando y exportando datos.....7
Importar ficheros ASCII externos 9
Importar ficheros externos en formato ASCII y ficheros web 9
Impotación de archivos de SAS.....85
Importación de archivos de texto.....78
Independencia.....226, 243, 289, 291
Independencia: teorema de Bayes..... 226
Índice complejo 227
Índice de dispersión respecto a la mediana 205
Índice de dispersión respecto a la mediana 205
Instalacion de paquetes estadísticos adicionales 3
Integración numérica..... 130
Interpretación gráfica de los factores..... 374
Intervalos de confianza.....275, 276, 277, 278, 284
Intervalos de confianza para comparar poblaciones normales..... 277
Intervalos de confianza para parámetros de distribuciones normales 276
Intervalos de confianza y test de hipótesis para la media y la diferencia de medias 284
Introducción a R 1

K

Kaiser-Meyer y Olkin 371

L

Leer archivos en formatos SPSS, SAS, Stata, S, Minitab, Systat, Octave 11
Leer datos de ficheros externos 7
Leer ficheros ASCII, SPSS, Stata.....7
Lenguaje r y los contrastes no paramétricos 300
Ley de probabilidad ...228, 229, 230, 231, 233, 234, 235, 236, 237, 238, 239, 240, 250, 255, 257, 258, 322, 399
Ley de probabilidad conjunta 230, 231
Leyes de probabilidad condicionadas 230
Leyes de probabilidad marginales..... 230, 231
Linealidad 413, 416, 417, 421
Listas 69

M

Marcos de datos 72
Matriz de cargas factoriales 365
Matriz de correlaciones.....113, 115, 214, 215, 217, 221, 222, 368, 369, 370, 381, 418
Matriz de covarianzas106, 113, 115, 214, 217, 221, 222, 412, 419, 421
Matriz de varianzas covarianzas 234, 236, 412
Media aritmética 104, 196, 201, 216
Media aritmética ponderada..... 104, 196, 216
Media armónica..... 197, 201
Media cuadrática 197
Media geométrica 197, 201
Media, mediana, moda, cuantiles, percentiles y momentos 195
Mediana 198
Mediana, moda, cuantiles, percentiles y momentos.....195
Medición de las componentes 364
Medida *KMO*..... 371, 372
Medidas de asimetría..... 206, 208
Medidas de centralización 195

© Alfaomega - RC Libros

Medidas de curtosis o apuntamiento 206
Medidas de dispersión.....195, 201, 202, 203, 204, 205
Medidas de dispersión absolutas 202
Medidas de forma..... 205
Medidas de posición de una distribución de frecuencias..... 196
Método de Newton 128
Método de Schroder's..... 130
Método de Simpson 133
Método del trapecio..... 130
Método equamax 373
Método oblimax..... 373
Método oblimin directo 373
Método quartimax..... 373
Método quartimin 373
Método varimax..... 373
Métodos multivariantes de la dependencia 360
Métodos multivariantes de la interdepedencia 359
Métodos multivariantes de reducción de la dimensión 359
Métodos oblimin..... 373
Minería de datos 360
Moda..... 198
Moda, cuantiles, percentiles y momentos 195
Modelo de regresion lineal multiple en R . 418
Modelo lineal de regresión múltiple 411
Modelos de probabilidad 231, 244
Modelos predictivos a través de R..... 411
Momento de orden *k* centrado en el origen 229, 242
Momento de orden *k* centrado en la media 230, 242
Momentos potenciales..... 200
MSA (Measure of Sampling Adequacy) 372
Multicolinealidad 364, 375, 396, 413, 418

N

Newton 128
Next..... 125
Nivel de confianza..... 113, 276
Normalidad de los residuos 412, 417
Números enteros..... 16
Números reales..... 16
Números y operaciones aritméticas en R.... 15

O

Operaciones aritméticas en R..... 15
Operadores aritméticos 38
Operadores lógicos..... 39
Operadores relacionales 40

P

Paquetes adicionales 3, 6
Percentiles y momentos..... 195
Probabilidad 225, 226, 227
Probabilidad condicionada 226
Programación estructurada..... 121
Puntuaciones o medición de las componentes 364
Puntuaciones o medición de los factores .. 374

R

R Studio..... 5
R y las distribuciones discretas 237
Recta de regresión..... 209, 210
Regla de Laplace 225
Regresión con interacciones..... 424
Regresión polinómica 423
Regresión simple mínimo cuadrática..... 209
Regresión sin constante 424
Relación entre intervalos de confianza y contrastes de hipótesis..... 278
Relacionar todos los promedios 201
Repeat..... 127
Rotación de las componentes 367
Rotación de los factores..... 372
Rotación promax 373, 374
Rotaciones oblicuas..... 373
Rotaciones ortogonales..... 373

© Alfaomega - RC Libros

S

Schroder 130
Secuencias 61
Sentencia for 117
Sentencia repeat 127
Sentencia while 119
Series de tiempo 85
Sistema de ecuaciones normales..... 210, 415, 416
Simpson 133
Switch 125

T

Técnicas de clasificación y segmentación . 395
Técnicas de reducción de la dimensión..... 359
Teorema central del límite 244, 249, 250, 259, 270, 271, 272, 273
Teorema de Bayes 226
Teorema de de Moivre 244
Teorema de Moivre 249
Teorema de Tchebichev 248, 270, 271
Test de Daniel 298, 299
Test de homogeneidad de varias muestras 300
Test de independencia de caracteres..... 300
Test de Kolmogrov-Smirnov para la homogeneidad de dos muestras 301
Test de la bonda de ajuste a una distribución 300
Test de la chi-cuadrado 294
Test de rachas 298
Test de rachas de Wald-Wolfowitz 298
Test de Wilcoxon-Mann-Whitney para la igualdad de medianas 301
Test de Wilcoxon-Mann-Whitney para la media de una población.....301
Test del cuadrado medio de diferencias sucesivas 298, 299
Test múltiple de Kruskal-Wallis para la igualdad de medias de varias poblaciones 301
Tests de hipótesis.....275
Transformación de datos 359

V

Variable aleatoria discreta 227
Variables aleatorias continuas..... 240, 243
Variables matriciales.....28
Variables simples 19
Variables y funciones de cadena 44
Varianza residual..... 211, 400
Varianza, cuasivarianza, desviación típica y error estándar 203
Vector esperanza matemática..... 234, 236
Ventanas gráficas en r 345

W

While.....136

Esta edición se terminó de imprimir en **julio** *de* **2017.** *Publicada por*
ALFAOMEGA GRUPO EDITOR, S.A. de C.V.
Dr. Isidoro Olvera (Eje 2 sur) No. 74, Col. Doctores, C.P. 06720,
Del. Cuauhtémoc, Ciudad México
La impresión y encuadernación se realizó en
CARGRAPHICS, S.A. de C.V. *Calle Aztecas No.27*
Col. Santa Cruz Acatlán, Naucalpan, Estado de México, C.P. 53150. México.