



Estadística Básica
con
R y R–Commander

2ª Edición Revisada
(Versión Marzo 2013)

Autores:

A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas



UCA | Universidad
de Cádiz

Servicio de Publicaciones

Copyright ©2008, 2013 Universidad de Cádiz. Se concede permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.3 o cualquier otra versión posterior publicada por la Free Software Foundation. Una traducción de la licencia está incluida en la sección titulada "Licencia de Documentación Libre de GNU".

Copyright ©2008, 2013 Universidad de Cádiz. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation. A copy of the license is included in the section entitled "GNU Free Documentation License".

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

<http://www.uca.es/publicaciones>

ISBN:

Depósito legal:



Apéndice A

Ficheros de datos

Puede accederse a los ficheros documentados en esta sección en la dirección http://knuth.uca.es/repos/ebrcmdr/bases_datos.

caracoles.dat Conjunto de datos que recoge las medidas del diámetro y la separación entre espirales (μm) de las conchas de 20 caracoles adultos.

cebada.dat Contiene información sobre la producción de cuatro variedades de cebada A, B, C y D.

chickwts Datos contenidos en el paquete “datasets” de R. Peso de 71 pollos sometidos a distintos tipos de alimentación. Contiene dos variables, una numérica *weight*: peso y un factor *feed*: tipo de alimentación, con 6 niveles.

eolico_apilado.dat Los datos del fichero **parque_eolico.dat** apilados según las variables *velocidad* y *parque*. Estos datos permiten trabajar más cómodamente en Rcmdr.

fenofibrato.dat Niveles de fibrinógeno de 32 pacientes, antes y después de ser tratados durante un año con fenofibrato.

iris Datos contenidos en el paquete “datasets” de R. Proviene del famoso estudio realizado por el estadístico y genetista Sir Ronald A.

Fisher. sobre la clasificación de 3 especies de iris (setosa, versicolor y virginica). Las variables de estudio son la longitud y el ancho del sépalo y, la longitud y el ancho del pétalo de las 3 especies mencionadas.

neumaticos.dat Vidas medias en rodaje de 5 tipos de neumáticos A, B, C, D y E, medidas en miles de kilómetros, probados en distintos coches de similares características.

niv_estudios_cadiz.dat Nivel académico de la población gaditana. Fuente: Instituto Estadístico de Andalucía.

peso_altura.dat Fichero en el que se proporcionan peso, altura y presión arterial inicial y final de un grupo de 100 pacientes sometidos a cierto fármaco (Ca Antagonista + diurético, IECA o placebo).

reproduccion_vir.dat Número de virus reproducidos en función del tiempo (minutos) y de la temperatura (grados), según el tipo de cultivo (ácido, básico o neutro).

titanic.dat Recoge información sobre el naufragio del buque Titanic (estatus económico, sexo, edad y supervivientes). Éste es el fichero incluido en el paquete “datasets” de R y está modificado para que se cargue correctamente en Rcmdr.

parque_eolico.dat Mediciones de la velocidad del viento (m/s) en dos localizaciones alternativas (Parque1 y Parque2) registradas de forma simultánea durante 730 horas.

Apéndice B

Tabla de medidas estadísticas



En la siguiente tabla se ofrece un resumen de las medidas más usadas en estadística descriptiva con sus correspondientes instrucciones en R

Medidas de posición	Instrucciones en R
Cuantiles	<code>> quantile(datos,p)</code> con p vector de cuantiles deseados. <code>> quantile(datos)</code> obtenemos todos los cuantiles.
Medidas de centralización	
Media	<code>> mean(datos)</code>
Mediana	<code>> median(datos)</code>
Medidas de dispersión	
Cuasivarianza	<code>> var(datos)</code>
Cuasidesviación típica	<code>> sd(datos)</code>
Varianza	<code>> var(datos)*</code> <code>(length(datos)-1)/length(datos)</code>

Medidas de dispersión	Instrucciones en R
Desviación típica	<code>>sqrt(var(datos)* (length(datos)-1)/length(datos))</code>
Rango muestral	<code>>max(datos)-min(datos)</code>
Rango intercuartílico	<code>>quantile(datos,.75) -quantile(datos,.25)</code>
Coefficiente de variación	<code>>sd(datos)/abs(mean(datos))</code>
Medidas de forma	En el paquete fBasics
Coefficiente de curtosis	<code>>kurtosis(datos)</code>
Coefficiente de asimetría	<code>>skewness(datos)</code>

Estadística Básica con R y R-commander
(Versión Marzo 2013)

*Autores: A. J. Arriaza Gómez, F. Fernández Palacín,
M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza,
A. Sánchez Navas*

©2008, 2013 Servicio de Publicaciones de la Universidad de Cádiz
<http://knuth.uca.es/ebrcmdr>



Apéndice C

Tabla de modelos

Modelo	Instrucción	Ecuación
Lineal	<code>>lm(Y ~ X, data=Datos)</code>	$Y = a + b \cdot X$
Lineal sin término independiente	<code>>lm(Y ~ 0 + X, data=Datos)</code>	$Y = a \cdot X$
Polinomial	<code>>lm(Y ~ X + I(X²) + + I(X³) + ... + I(Xⁿ), data=Datos)</code>	$Y = a_0 + a_1 \cdot X +$ $+ \dots + a_n \cdot X^n$
Polinomial sin término independiente	<code>>lm(Y ~ 0 + X + I(X²) + + I(X³) + ... + I(Xⁿ), data=Datos)</code>	$Y = a_1 \cdot X +$ $+ \dots + a_n \cdot X^n$
Potencial	<code>>lm(log(Y) ~ log(X), data=Datos)</code>	$Y = a' \cdot X^b, (1)$
Exponencial	<code>>lm(log(Y) ~ X, data=Datos)</code>	$Y = e^{a+b \cdot X}$
Logarítmico	<code>>lm(Y ~ log(X), data=Datos)</code>	$Y = a + b \cdot \log(X)$
Hiperbólico	<code>>lm(Y ~ I(1/X), data=Datos)</code>	$Y = a + \frac{b}{X}$
Doble inverso	<code>>lm(I(1/Y) ~ I(1/X), data=Datos)</code>	$Y = \frac{1}{a + \frac{b}{X}}$
Lineal generalizado	<code>>glm(fórmula, family= =familia(link), data=Datos)</code>	(2)

(1) Los coeficientes a y b obtenidos en Rcmdr corresponden a la ecuación $\log(Y) = a + b \cdot \log(X)$, con lo que el modelo potencial sería $Y = e^a \cdot X^b$.

(2) `familia` puede tomar los valores `gaussian`, `binomial`, `poisson`, `Gamma`, `inverse.gaussian`, `quasibinomial` y `quasipoisson`. La función de enlace (`link`) puede tomar distintos valores según la familia seleccionada. Podemos ver las distintas opciones consultando en la ayuda de R la función `family` (`help(family)` o `?family`).

Índice alfabético

Análisis

- de bondad del ajuste, 62
- de residuos, 63
- inferencial, 109

ANOVA, 141

Atributo, 27

clase, 2

Codificar, 44

Coefficiente

- de contingencia, 55
- de correlación lineal, 62
- de Cramer, 55
- de determinación, 62

Conjunto soporte, 27

Contraste de hipótesis, 110

- bilateral, 112
- unilateral, 112

Corrección de Yates, 132

Cuantiles, 87

Cuartil, 42, 44

data.frame, 25

Datos aleatorios, 100

Diagrama

- de barras, 42, 45
- de caja, 47
- de dispersión, 59
- de sectores, 40, 42

Distancia de Cook, 71

Distribución

- Beta, 86
- Binomial, 86, 87
- Binomial negativa, 86, 91
- Cauchy, 86
- Chi-cuadrado, 86, 100
- de probabilidad
 - continua, 86
 - discreta, 86
- discreta, 86
- Exponencial, 86, 97
- F-Snedecor, 86, 100
- Gamma, 86
- Geométrica, 86, 91
- Gumbel, 86
- Hipergeométrica, 86, 90
- Logística, 86
- Lognormal, 86
- marginal, 53
- Normal, 86, 93
- Poisson, 86, 89
- t-Student, 86, 98
- Uniforme, 86, 95
- Weibull, 86

Error

- de tipo I, 111
- de tipo II, 111

Escala de Medida, 40

Experimento aleatorio, 85

- Función
 - de densidad, 86
 - de distribución, 86
 - de probabilidad, 85
- GNU-GPL, *véase* Licencia GNU-GPL
- Gráfica
 - de barras, 58
 - de dispersión, 75
 - de los residuos, 63
 - estudentizados, 70
 - indexados, 70
 - de mosaico, 56
 - de potenciales hat, 71
- Hipótesis
 - alternativa, 111
 - nula, 110
- Histograma, 47
- Homocedasticidad, 141
- Imputación, 25
- Inferencia
 - estadística, 83
 - no paramétrica, 111
 - paramétrica, 111
- Intervalos de confianza, 110
- Licencia GNU-GPL, 2
- Mínimos cuadrados, 60
- Medidas
 - centrales, 40
 - de dispersión, 40
 - de posición, 42
- Modelo de datos, 24
- Muestra
 - aleatoria, 86
 - de la distribución, 87
- Nivel de confianza, 111
- objeto, 2
- Población, 23
- Predicciones, 61, 68
- Probabilidades
 - acumuladas, 87
- R-Commander, *véase* Rcmdr
- Rango
 - intercuartílico relativo, 45
- Rcmdr, 5, 7
- Recorrido
 - intercuartílico, 42
- Representaciones gráficas, 40
- Residuos indexados, 70
- Suceso elemental, 85
- Tabla
 - de doble entrada, 53
 - de frecuencias
 - conjunta, 53
- Teoría
 - Bayesiana, 109
 - de probabilidad, 83
- Test
 - de autocorrelación de Ljung-Box, 126
 - de Barlett, 143
 - de comparaciones múltiples, 142
 - de Kolmogorov-Smirnov, 128, 130
 - de Kruskal Wallis, 142, 147
 - de la χ^2 , 128
 - de la F, 141, 144
 - de rachas, 127
 - de Scheffé, 142
 - de Shapiro-Wilk, 128

- de Welch, 142
- de Wilcoxon, 135, 137
- HSD de Tukey, 142

Valores

- ajustados, 68
- influyentes, 63
- missing, 25
- omitidos, 25

Variable, 27

- absolutamente continua, 27
- aleatoria, 83, 85, 86
- continua, 27
- discreta, 27
- etiqueta, 24