Package 'StatR'

April 12, 2024

Version 2.0

Title Estadística Con R

Author Cástor Guisande González, Antonio Vaamonde Liste & Aldo Barreiro Felpeto

Maintainer Cástor Guisande González <castor@uvigo.es>

Description Aplicación para facilitar la realización de análisis y gráficos estadísticos con R.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.1.1)

Repository CRAN

R topics documented:

II1 4
III1 12
III2 18
III3 20
III4 22
IV1 24
IV2 26
IV3 29
IV4 31
IX1 33
IX10 38
IX11 42
IX12 45
IX13 48
IX2 52
IX3 57
IX4 66
IX5 74
IX6 78
IX7 84
IX8 89

IX9 100
PAR 105
VI1 108
VI2 115
VI3 120
VII1 125
VII2 134
VII3 141
VII4 145
VII5 150
VIII1 154
VIII2 162
X1 170
X2 176
X3 181
XI1 186
XI10 196
XI11 200
XI2 202
XI3 220
XI4 224
XI5 226
XI6 235
XI7 242
XI8 243
XI9 248
XII1 250
XII2 255
XII3 258
XII4 261
XII5 264
XIII1 267
XIII2 275
XIII3 283
XIII4 287
XIV1 293
XIV2 297
XIV4 307
XIV5 310
XV1 313
XV2 328
XV3 334
XV4 337
XV5 340
XV6 341
XV7 353
XV8 355
XV9 365

XVI1 . . . 368
XVI2 . . . 377
ZII1 . . . 383
ZIV1 . . . 384
ZIV3 . . . 384
ZIX1 . . . 384
ZIX10 . . . 385
ZIX11 . . . 385
ZIX3 . . . 386
ZIX5 . . . 386
ZIX6 . . . 387
ZIX7 . . . 387
ZIX9 . . . 388
ZVI1 . . . 388
ZVII1 . . . 389
ZVII2 . . . 389
ZVII3 . . . 390
ZVII4 . . . 390
ZVII5 . . . 391
ZVIII1 . . . 391
ZVIII2 . . . 392
ZX1 . . . 392
ZX3 . . . 393
ZXI1 . . . 393
ZXI10 . . . 394
ZXI11 . . . 394
ZXI12 . . . 395
ZXI2 . . . 396
ZXI3 . . . 396
ZXI4 . . . 397
ZXI5 . . . 397
ZXI6 . . . 398
ZXI7 . . . 398
ZXI8 . . . 399
ZXI9 . . . 399
ZXII1 . . . 400
ZXIII1 . . . 400
ZXIII2 . . . 401
ZXIII3 . . . 401
ZXIII4 . . . 402
ZXIV1 . . . 402
ZXIV2 . . . 403
ZXIV3 . . . 403
ZXIV4 . . . 403
ZXV1 . . . 404
ZXV2 . . . 404
ZXV3 . . . 405

Index **406**

II1 *ESTADISTICA DESCRIPTIVA*

Description

Estadística descriptiva que incluye medidas de posición como las medias aritmética, geométrica, armónica, ponderada, la moda, la mediana, etc., medidas de dispersión o variabilidad de la muestra como la varianza, la desviación típica, el coeficiente de variación, etc., y medidas de distribución como la asimetría y la curtosis.

Usage

```
II1(data, variables, agrupa=NULL, recorte=0.05, cuantiles=c(0.025,0.05,0.95,0.975),
variablespon=NULL, variablepeso=NULL, agrupapon=NULL, file1="Output1.csv",
file2="Output2.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
variables	Variables de las que se quiere calcular las medidas estadísticas.
agrupa	Variables con las que se agrupan los datos para hacer los cálculos estadísticos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna.
recorte	Punto de corte de la media acotada.
cuantiles	Valores de los cuantiles que se quieren calcular.
variablespon	Si interesa la media ponderada, hay que especificar en primer lugar cual es la variable o variables para las que se va obtener el promedio ponderado. Si se selecciona NULL, entonces no se calcula la media ponderada.
variablepeso	Variable o variables que se usan para ponderar. Al igual que antes, si se selecciona NULL no se calcula la media ponderada.
agrupapon	Variables con las que se agrupan los datos para hacer los cálculos de media ponderada. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna.
file1	ARCHIVOS CSV. Nombre del archivo de la estadística descriptiva.
file2	ARCHIVOS CSV. Nombre del archivo de la media ponderada.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

II. ESTADÍSTICA DESCRIPTIVA

II.1. MEDIDAS DE POSICIÓN

El primer paso a la hora de tratar unos datos es intentar buscar algún tipo de medida que nos permita caracterizar, diferenciar y distinguir las series de datos. Esto se puede hacer determinando la posición de los datos. Dentro de este grupo se diferencian las llamadas medidas de posición central y otras medidas que estiman diferentes posiciones de los datos. Unas u otras se utilizan dependiendo del tipo de datos con el que se esté trabajando.

II.1.1. Medidas de posición central

II.1.1.1. Media aritmética

La media aritmética ($\bar{x}$), que también se conoce simplemente como media o promedio, se calcula por medio de la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde *x* son cada uno de los valores *i* de la variable y *n* el número de datos.

Cuando la media es del conjunto de toda la población se suele denotar por el símbolo μ y cuando es la media de una muestra de la población, entonces se utiliza $\bar{x}$.

En el caso de datos agrupados (tabla de frecuencias) la media se calcula por medio de la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

donde *m* y *f* son el valor medio y la frecuencia (número de datos) de la clase *i*, respectivamente, y *k* el número de intervalos o clases.

La media aritmética es la que se utiliza más frecuentemente porque es la que tiene un error estándar más pequeño, es la más fácil de estimar, tiende hacia una distribución Normal incluso si los datos originales no presentan esta distribución y, por último, es más sensible que otras medidas de posición a cambios en la distribución de los datos, lo cual es muy importante en estadística para poder determinar diferencias entre series de datos (Sokal & Rohlf, 1981). El problema que plantea la media aritmética también deriva de su sensibilidad, ya que está más afectada por los datos raros que se salen del rango.

II.1.1.2. Media geométrica

La media geométrica *MG* se utiliza en algunos casos con datos relativos como porcentajes o tasas de incremento y se calcula por medio de la siguiente fórmula:

$$MG = \sqrt[n]{x_1 x_2 .... x_n}$$

II.1.1.3. Media armónica

La media armónica *MA* se utiliza, por ejemplo, en algunos casos en los que es necesario promediar variaciones con respecto al tiempo. Se calcula de la siguiente forma:

$$MA = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} .... \frac{1}{x_n})}$$

La media armónica es siempre inferior o igual a la media geométrica y, por tanto, también inferior o igual a la media aritmética.

II.1.1.4. Moda

A veces es necesario determinar la posición de variables que no son numéricas sino cualitativas como, por ejemplo, las especies a las que pertenecen diferentes individuos. En este caso no podemos utilizar ninguna de las medias citadas anteriormente y, por lo tanto, debemos recurrir a medidas como la moda. También puede ser utilizada con variables cuantitativas.

La moda de un conjunto de observaciones se define como el valor que se presenta con mayor frecuencia, es decir, un mayor número de veces. La moda se puede denotar por *Mo*.

Si solamente es un valor el que se repite más veces, en ese caso se dice que la distribución es unimodal, porque solo tiene una moda. En el caso de que ningún valor se repita, entonces por definición se considera que no existe moda. Se puede dar el caso de que dos observaciones tengan la misma frecuencia máxima, siendo en este caso la distribución bimodal. La distribución multimodal sería en aquellas situaciones en las que existen más de dos valores con la misma frecuencia máxima. Por último, en situaciones raras en las que la distribución de los datos tenga forma de U, el punto medio de la distribución se denomina antimoda.

II.1.1.5. Mediana

La mediana de un conjunto de observaciones es el valor para el cual, cuando todas las observaciones se ordenan de manera creciente, la mitad de éstas son menores que este valor y la otra mitad son mayores. Sea $x_1, x_2, ...x_n$ una muestra aleatoria de *n* observaciones ordenadas de forma creciente, la mediana de estos datos se calcula de la siguiente manera:

Si *n* es un número impar.

$$X_{(\frac{n+1}{2})}$$

Si *n* es un número par.

$$\frac{X_{\frac{n}{2}} + X_{(\frac{n}{2}+1)}}{2}$$

La mediana se suele utilizar en series de datos que tienen una distribución muy asimétrica, ya que en estos casos no siempre es adecuado utilizar la media aritmética. Por último, no se ve afectada por valores raros que se salgan mucho del rango normal, como le ocurre a la media aritmética.

II.1.1.6. Media acotada o recortada

Otra medida de tendencia central que se suele utilizar con fines exploratorios es la media acotada (en inglés, *trimmed mean*), porque es más resistente a valores extremos que la media y no es tan insensible como lo podría ser la mediana. La media acotada se define como el promedio de los valores en donde se ha excluido un porcentaje determinado de la población a cada lado de los valores extremos. Para su cálculo hay que seguir tres pasos: 1) Se ordenan los datos; 2) Se descarta el porcentaje de datos a cada extremo de la población o muestra y 3) Se calcula la media con los datos restantes.

Esta medida sirve como un parámetro exploratorio, ya que muestra el impacto de los valores extremos sobre la estimación de los promedios, porque se calcula a diferentes porcentajes. Además facilita la detección de valores inusuales, los cuales advierten al investigador de posibles errores, de datos que podrían no ser parte de la población objetivo del estudio o de valores anómalos que indiquen nuevos rumbos en la investigación.

II.1.1.7. Media ponderada

La media ponderada ($\bar{x}_w$) se utiliza en aquellos casos en que cada valor es a su vez un promedio, que debe ser ponderado por el tamaño de la muestra a la que corresponde. También puede utilizarse cuando algunos datos son más fiables que otros, simplemente porque unos datos han sido calculados con un mayor esfuerzo o por cualquier otra razón y, por ello, es necesario darle más peso a estos valores a la hora de estimar la media. Se calcula por medio de la siguiente fórmula:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

donde *x* son cada uno de los valores *i* de la variable, *w* el peso que se le da al dato y *n* el número de datos.

Una aplicación de la media ponderada que merece una mención especial, porque puede dar una gran información, es la obtención del centro de gravedad de una distribución de datos de acuerdo a una determinada variable. En Ecología es un parámetro que permite cuantificar el nicho ecológico de una especie, ya que da información sobre el valor medio de su distribución para una determinada variable. Por ejemplo, se puede obtener la temperatura, la salinidad, concentración de oxígeno, etc. media en que suele aparecer una especie. En este caso, *w* son cada uno de los valores *i* de los datos de abundancia de la especie, *x* el valor de la variable para el valor *w* y *n* el número de datos.

II.1.2. Otras medidas de posición

Se definen los cuantiles de orden *k* como los valores de la variable, supuesta ésta ordenada de menor a mayor, que la dividen en *k* partes con la misma frecuencia de observaciones. Por lo tanto existirán *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* de frecuencia de observaciones. Por ejemplo, el cuantil 15 de orden 100 deja por debajo el 15% de los valores del total de la serie completa de valores.

Los cuantiles más utilizados son los **percentiles**, **cuartiles** y **deciles**, los cuales se describen a continuación.

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.

Los **cuartiles** son los tres valores de la variable que dividen la distribución en 4 partes iguales, es decir, en 4 intervalos, dentro de cada cual está incluido el 25% de los valores de la distribución. El percentil 25 *(*P_{25}*)* sería igual al cuartil 1 *(*Q_1*)*, el percentil 50 *(*P_{50}*)* sería igual al cuartil 2 (Q_2, igual también a la mediana), etc.

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. El percentil 10 *(*P_{10}*)* sería igual al decil 1 *(*D_1*)*, el percentil 20 *(*P_{20}*)* sería igual al decil 2 *(*D_2*)*, etc.

El procedimiento para encontrar el valor del cuantil *j* de orden *k* a partir de datos ordenados de menor a mayor, es el siguiente:

1. Encontrar la posición *i* del cuantil *j* -ésimo mediante el cálculo de *nj/k*.

2. Si *nj/k* no es un entero, entonces la posición *i* es el siguiente entero más grande y el valor del cuantil es el dato ordenado en la posición de este entero más grande.

3. Si *nj/k* es un entero, entonces la posición del cuantil será *i=nj/k* + 0,5 y así el valor del cuantil es el promedio de las observaciones ordenadas *nj/k* y *nj/k* + 1.

II.2. MEDIDAS DE DISPERSIÓN

Además de la posición, es también importante la dispersión o variabilidad de los datos. Dos series de datos pueden tener la misma media, pero la variabilidad puede ser diferente entre ambas.

Las medidas de dispersión tienen como finalidad estudiar hasta qué punto, para una determinada distribución de datos, las medidas de posición representan bien el conjunto de datos de la distribución. Por ejemplo, si se quiere determinar si 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, éste será representativo de ellos. Es decir, será más representativa la media aritmética de una variable cuanto más agrupados en torno a ella estén los valores promediados y, 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.

Por lo tanto, para complementar la información que se obtiene a partir de la media es necesario otro tipo de parámetros que midan la dispersión o variabilidad de los datos, los cuales se muestran a continuación.

II.2.1. Amplitud

El método más simple de estimar la dispersión de los datos es por medio de la amplitud, también conocida como rango, es decir, la diferencia entre el valor mínimo y el máximo de la serie de datos.

II.2.2. Varianza y cuasivarianza

La mejor forma de medir la dispersión de una serie de datos es comparar cada uno de ellos con la media de la serie, y esto es exactamente lo que hace la varianza (σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

A menudo queremos utilizar la varianza de una muestra como valor aproximado o estimado de la varianza desconocida de la población de la que procede esa muestra. En estos casos, el error cometido es generalmente más pequeño si en vez de considerar como estimador la varianza de la muestra, se utiliza lo que se denomina cuasivarianza (s^2), que se calcula como la anterior, pero cambiando el denominador por *n* - 1:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

En aquellos casos en los que disponemos de toda la población se debe emplear *n* y no *n* - 1. Sin embargo, como en general nuestros datos son una muestra de una población mayor, lo correcto es usar la cuasivarianza en vez de la varianza.

Muchas veces se hace referencia a la varianza cuando en realidad se está calculando la cuasivarianza (varianza estimada). La mayoría de los programas de estadística utilizan solo la cuasivarianza y no la varianza. Es frecuente también observar que σ^2 y s^2 se usan indistintamente para indicar varianza o cuasivarianza sin un criterio definido.

II.2.3. Desviación típica y cuasidesviación típica

El problema que presenta la varianza es que, para evitar valores negativos, las diferencias se elevan al cuadrado. Para poder tener la medida de dispersión en las mismas unidades que la media, se suele utilizar, en vez de la varianza, la desviación típica (σ), también llamada desviación estándar, y que se calcula simplemente como la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Al igual que ocurría con la varianza, salvo en el caso de que el tamaño de la muestra sea idéntico al de la población, es preferible utilizar la cuasidesviación típica (*s*), en vez de la desviación típica. Para calcular la cuasidesviación típica se divide por el número de grados de libertad (*n* - 1) en vez de dividir por el total de datos (*n*).

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Es necesario mencionar que la mayoría de los programas estadísticos solo utilizan la cuasidesviación típica y, al igual que ocurría con la varianza, es frecuente observar que se hace mención a la desviación típica cuando en realidad se ha calculado la cuasidesviación típica. También se observa con frecuencia que σ y s se usan indistintamente para desviación o cuasidesviación típica sin un criterio definido.

II.2.4. Desviación absoluta respecto a la media y la mediana

Una medida de dispersión que se suele utilizar es la desviación media absoluta que se define en el caso de la media como:

$$DAM = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

y para la mediana como:

$$DAMed = \frac{1}{n}\sum_{i=1}^{n}|x_i - Mediana|$$

Esta medida indica el promedio de las desviaciones ya sean estas negativas o positivas con respecto al promedio general y facilita la interpretación del grado de dispersión de los datos analizados.

II.2.5. Coeficiente de variación

Cuando se comparan series de datos que tienen medias diferentes, la desviación típica no permite comparar adecuadamente cual de las dos series de datos presenta una mayor variación, ya que normalmente la serie de datos con una mayor media también suele tener una mayor variabilidad. En este caso es más aconsejable el uso del coeficiente de variación *CV*, que se calcula como el porcentaje que representa la desviación típica con respecto a la media. En el caso de que la desviación típica sea desconocida porque nuestros datos son de una muestra, se utiliza la cuasidesviación típica como estimador de la desviación típica poblacional.

$$CV = \frac{s * 100}{\bar{x}}$$

II.2.6. Error estándar de la media

Proporciona una medida de la precisión de la estimación de la media poblacional a partir de una muestra, mientras que la desviación típica mide la variabilidad de los datos respecto de la media en la muestra. El error estándar se calcula a partir de la desviación típica. Cuando ésta es desconocida, se utiliza la cuasidesviación típica para obtener el error estándar estimado

$$ES = \frac{s}{\sqrt{n}}$$

II.2.7. Recorrido intercuartílico

El recorrido intercuartílico Q se calcula a partir de los percentiles 75 (P_{75}) y 25 (P_{25}) de la siguiente forma:

$$Q = P_{75} - P_{25} = Q_3 - Q_1$$

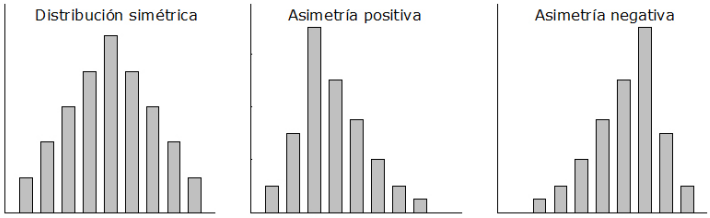
Esta medida de dispersión se utiliza cuando se expresa la posición central por medio de la mediana.

II.3. MEDIDAS DE DISTRIBUCIÓN

II.3.1. Asimetría

Las medidas de asimetría tienen como finalidad el determinar si una variable se distribuye de forma simétrica con respecto a un valor central, o bien si la distribución de los datos tiene una forma diferente del lado derecho que del lado izquierdo. La distribución es simétrica si al lado derecho y al lado izquierdo del valor central la distribución de los datos es la misma. La distribución tiene asimetría positiva si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en el derecho hay frecuencias más pequeñas. La asimetría es negativa cuando las frecuencias más pequeñas están en el lado izquierdo (Figura II.1). Como valor central de referencia se suele usar la media aritmética, aunque también se puede utilizar la mediana.

Figura II.1. Diferentes tipos de asimetría



Existen muchas formas de medir la simetría y una de ellas es la fórmula que se muestra a continuación (coeficiente de asimetría de Fisher), la cual utiliza como valor central la media aritmética:

$$A = \frac{n\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{(n-1)(n-2)\sigma^3}$$

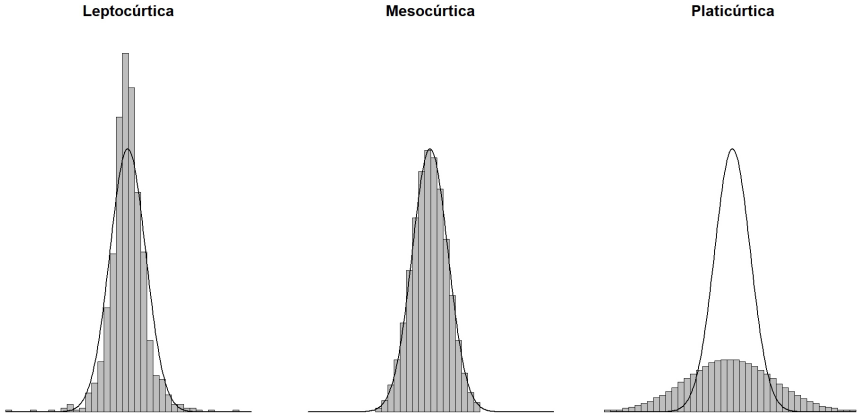
siendo σ la desviación típica estimada o cuasidesviación típica muestral. Cuando el valor es cercano a cero la distribución es simétrica, la asimetría es positiva cuando el valor es mayor que cero y la asimetría es negativa cuando el valor es menor que cero.

II.3.2. Apuntamiento o curtosis

La medida de curtosis caracteriza las colas (o el potencial atípico) de la distribución de los datos. Históricamente, se pensó que una curtosis más alta implica una distribución más o menos "puntiaguda". Por esta razón, históricamente, la curtosis se llamó una medida de apuntamiento, sesgo o concentración central. Sin embargo, este concepto es erróneo. La curtosis no informa sobre la forma del apuntamiento; su única interpretación inequívoca es en términos del extremo de la cola, es decir, ya sea valores atípicos existentes (para la curtosis de muestra) o propensión a producir valores atípicos (para la curtosis de una distribución de probabilidad) (ver Westfall, 2014).

Cuando el valor de Kurtosis es positivo, la distribución es leptocúrtica, lo que significa que el carácter atípico de la distribución es más extremo que el de una distribución normal (Figura II.2, panel izquierdo). Cuando el valor de curtosis es cero o cercano a cero, la distribución tiene una característica similar a la distribución Normal, y se denomina mesocúrtica (Figura II.2, panel central). Finalmente, cuando el valor de curtosis es negativo, la distribución es platicúrtica, lo que significa que el carácter atípico de la distribución es menos extremo que el de una distribución normal (Figura II.2, panel derecho).

Figura II.2. Tipos de curtosis (barras) en comparación con la distribución Normal (línea continua).



Al igual que ocurría con la asimetría, existen diversas formas de estimar la curtosis de una distribución de datos, pero una de las más usadas es el coeficiente de apuntamiento de Fisher que se muestra a continuación:

$$C = \frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4 - 3\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)(n-1)}{(n-1)(n-2)(n-3)\sigma^4}$$

EJEMPLO

Se estiman medidas de posición como las medias aritmética, geométrica, armónica, la moda, la mediana, etc., y medidas de dispersión o variabilidad de la muestra como la varianza, la desviación típica, el coeficiente de variación, etc., de la abundancia de varias especies de peces en dos zonas (S1 y S2). Estos parámetros fueron estimados realizando los cálculos para cada familia y género. Se calcula también la media ponderada de la temperatura para los dos sitios de muestreo en base a la abundancia de las especies.

Value

Se exporta un archivo CSV con medidas de posición como las medias aritmética, geométrica, armónica, la moda, la mediana, etc. y medidas de dispersión o variabilidad de la muestra como la varianza, la desviación típica, el coeficiente de variación, etc. Opcionalmente se exporta un segundo archivo CSV con la media ponderada.

References

Sokal RR & Rohlf FJ (1981) *Biometry*. WH Freeman and Company, New York.

Westfall, R.H. (2014) Kurtosis as Peakedness, 1905. *RIP* (2014) *American Statistician*, 68: 191-195.

Examples

```
## Not run:

data(ZII1)

II1(data=ZII1, variables=c("S1","S2"),agrupa=c("Family","Genus"),
variablespon=c("Temperature"), variablepeso=c("S1","S2") )


## End(Not run)
```

III1 *DISTRIBUCION NORMAL*

Description

Calcula la probabilidad de una distribución Normal indicando la media y la desviación.

Usage

```
III1(valor, media, desvi, prob.mayor=FALSE)
```

Arguments

valor	Valor para el que se quiere calcular la probabilidad de la distribución Normal.
media	Media de la distribución Normal.
desvi	Desviación estándar de la distribución Normal.
prob.mayor	Se indica si se quiere obtener la probabilidad mayor "Si" o menor "No" que el valor.

Details

III. DISTRIBUCIÓN

Otra información importante es la relacionada con la distribución que tiene una variable. Se puede dar el caso de que dos variables tengan exactamente la misma media y la misma dispersión, pero que el tipo de distribución sea completamente diferente. Por lo tanto, además de la información sobre las medidas de posición y dispersión, que se explicaron en la función II1, es necesario también conocer la distribución de los valores de la variable.

Para estudiar la distribución de una variable, e incluso comparar medias y varianzas entre variables, lo que se hace es comparar las frecuencias de los valores de la variable con las probabilidades resultantes de modelos teóricos de distribuciones. El modelo teórico de distribución a utilizar variará en función del tipo de variable con la que estemos trabajando.

III.1. TIPOS DE VARIABLES

Existen básicamente dos tipos de variables: *cualitativas* y *cuantitativas*. Las primeras no se expresan numéricamente (sexo, especie a la que pertenece un individuo, provincia de nacimiento) aunque pueden utilizarse códigos numéricos para representar sus valores (por ejemplo en la variable sexo, en vez de macho y hembra se pueden nombrar como 1 y 2). Las cuantitativas se expresan directamente en términos numéricos (número de hojas de una planta, edad, longitud, temperatura, etc.).

Las variables cualitativas a su vez pueden ser de dos tipos: *nominales*, si sus valores no están ordenados de modo natural (lugar de nacimiento, especie), y ordinales si sus valores tienen un orden (por ejemplo una variable toxicidad que toma los valores nada, poco, bastante y muy tóxico).

Las variables cuantitativas pueden ser también de dos tipos: *discretas* y *continuas*. Son discretas cuando solo pueden tomar valores específicos, y no es posible ningún valor entre dos consecutivos (número de hijos que puede tener una familia, número de hojas de una planta, etc.). En el caso de variables discretas a cada valor de la variable le corresponde una probabilidad, de la cual dependerá el número de veces que se repita ese valor en relación con los valores restantes. A la correspondencia entre los valores y sus respectivas probabilidades se le llama *función de probabilidad*.

Las variables cuantitativas continuas son aquellas que pueden tomar cualquier valor a lo largo de un continuo, de modo que no hay valores consecutivos, ya que entre dos valores cualesquiera siguen existiendo infinitos valores posibles (temperatura, longitud). Las variables continuas se pueden agrupar en categorías, pero de una forma arbitraria. Por ejemplo, la variable altura se puede dividir en categorías como pequeño, normal y alto, y los límites de cada una de estas categorías se pueden establecer de forma arbitraria. A diferencia de lo que ocurría con las variables discretas, para las variables continuas no es útil establecer la correspondencia entre valores y probabilidades. Lo que se hace es calcular la probabilidad contenida en un determinado segmento o rango de valores, que dividida por la amplitud del segmento es la densidad media de probabilidad del segmento, a partir de la cual se determina la densidad de probabilidad para cada valor. La correspondencia entre los valores y su respectiva densidad de probabilidad se llama *función de densidad*.

La distinción entre estos cuatro tipos de variables es importante por varios motivos:

1. El cálculo de algunas medidas de posición o de dispersión no tiene sentido con variables cualitativas, por ejemplo en el caso de la variable sexo.

2. Para la aplicación correcta de técnicas de análisis estadísticos: así la mayoría de las pruebas no paramétricas requieren que la variable sea al menos ordinal, y muchos métodos de análisis multivariante exigen que las variables sean cuantitativas (por ejemplo análisis factorial o análisis discriminante).

III.2. DISTRIBUCIONES PARA VARIABLES CONTINUAS

III.2.1. Normal

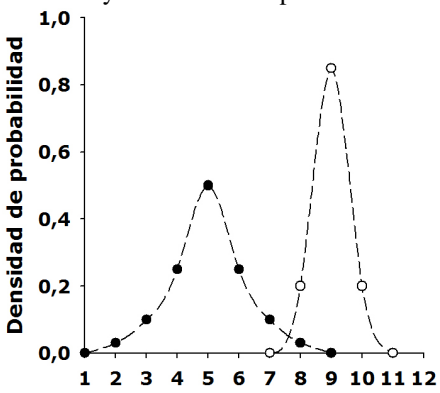
La distribución Normal es una de las más importantes porque se observa en muchas series de datos y, además, distintos tipos de análisis estadísticos presentan como condición para poder aplicarse a la serie de datos, que ésta presente este tipo de distribución. Varios matemáticos intervinieron en su desarrollo, entre los que figura el matemático y astrónomo Carl Friedrich Gauss (1777-1885), por lo cual a veces es llamada en su honor la campana o distribución de Gauss. La función de densidad de la distribución Normal se describe por medio de la siguiente fórmula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

donde *f(x)* es la densidad de probabilidad del valor *x*, μ es la media y σ es la desviación típica.

La forma de la distribución Normal varía en función de la media μ y la desviación típica σ como se observa en la Figura III.1.

Figura III.1. Distribuciones normales con diferentes medias y desviaciones típicas.



III.2.1.1. Aplicaciones de la distribución Normal

La tabla de la distribución Normal N(0,1), es decir con μ = 0 yσ = 1 (Tabla 1 del Apéndice I en Guisande et al., 2011), nos permite calcular probabilidades relativas a cualquier otra distribución Normal con diferente μ y σ. Para ello, basta tipificar la variable, es decir, calcular el valor *Z* (unidades de desviación con respecto a la media) correspondiente a los valores *x* indicados mediante la operación:

$$Z = \left(\frac{x-\mu}{\sigma}\right)$$

Este valor *Z* que se calcula a partir de la variable *X* nos permite obtener en las tablas las probabilidades correspondientes a cualquier intervalo (véase ejemplo).

III.2.2. t de Student

Cuando una variable sigue una distribución Normal, la media de una muestra aleatoria de esa variable también tiene distribución Normal, y su media es la media poblacional desconocida μ. Esto

puede ser utilizado para estimar μ. Sin embargo, a menudo no se conoce la desviación típica de la población σ (solo se trabaja con una muestra de individuos del total de la población) y, además, puede ocurrir que el número de observaciones de la muestra sea pequeño (menor de 30). En estos casos, se puede utilizar la cuasidesviación típica de la muestra (*S*) junto con la distribución *t* de Student:

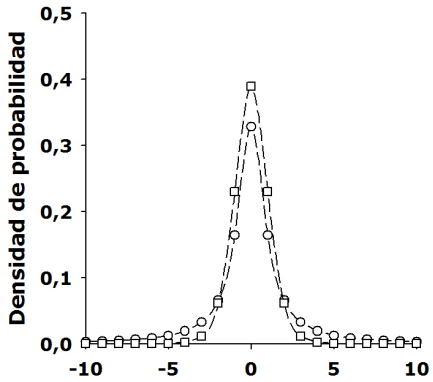
$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

La función de densidad de probabilidad de la distribución *t* de Student viene dada por la siguiente expresión:

$$f(x) = \frac{1}{\sqrt{\upsilon\pi}} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\frac{\upsilon}{2}} \left(1 + \frac{x^2}{\upsilon}\right)^{-\left(\frac{\upsilon+1}{2}\right)}$$

La distribución *t* de Student puede tener diferentes formas dependiendo de los grados de libertad (υ) (Figura III.2). La apariencia general de la distribución *t* es similar a la de la distribución Normal estándar. Sin embargo, la distribución *t* tiene colas más amplias que la Normal, es decir, la probabilidad de las colas es mayor que en la distribución Normal. La distribución *t* se transforma en una distribución Normal cuando el número de datos tiende a infinito. Los valores críticos para distintos niveles de significación y distintos grados de libertad se pueden ver en la Tabla 2 del Apéndice I en Guisande et al. (2011).

Figura III.2. Funciones de densidad de la distribución de Student para 1 (círculo) y 10 (cuadrado) grados de libertad.



Las principales aplicaciones de la distribución *t* de Student en la inferencia estadística son:

1. Para estimar, mediante intervalos de confianza, la media poblacional.

2. Estimar y probar hipótesis sobre una diferencia de medias.

Las hipótesis o supuestos para poder aplicar la *t* de Student son que en cada grupo la variable estudiada siga una distribución Normal y que la dispersión en ambos grupos sea homogénea (hipótesis de homocedasticidad = igualdad de varianzas) aunque este estadístico también se puede usar sin asumir igualdad de varianzas.

III.2.3. Chi-cuadrado

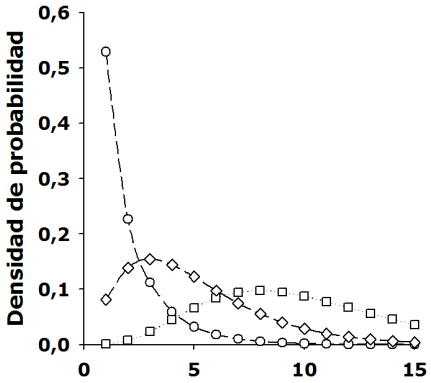
La función de densidad de la distribución Chi-cuadrado (χ^2) se describe por medio de la siguiente expresión:

$$f(x) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)} x^{\left(\frac{\upsilon}{2}-1\right)} e^{-\frac{x}{2}}$$

donde υ son los grados de libertad y *x* no es negativo.

A diferencia de lo que ocurría con la distribución Normal, debido a que la distribución χ^2 depende de los grados de libertad, no existe una curva típica sino que la distribución χ^2 puede tener diferentes formas dependiendo de los grados de libertad (Figura III.3).

Figura III.3. Curvas de densidad de la distribución de χ^2 para 1 (círculo), 5 (rombo) y 10 (cuadrado) grados de libertad.



El valor de la variable que deja a su derecha un área α bajo la curva de densidad se llama el punto crítico correspondiente al nivel de significación α. Están tabulados los puntos críticos de distintos niveles de significación y distintos grados de libertad (Tabla 3 del Apéndice I en Guisande et al., 2011).

Son tres las aplicaciones principales que tiene la distribución χ^2: test de bondad de ajuste, test de independencia y test de homogeneidad. Además, esta distribución juega un papel importante en otras muchas pruebas estadísticas.

1. El test de bondad de ajuste consiste en el planteamiento de hasta qué punto una muestra se puede considerar como perteneciente a una población con una distribución teórica ya conocida. Es un método que se utiliza frecuentemente para determinar si una serie de datos presenta una distribución Normal, de Poisson, etc.

2. El test de independencia determina si dos caracteres *X* e *Y* de una población son dependientes o independientes. Por ejemplo, queremos deteminar si la supervivencia de los descendientes de las hembras de una población es independiente o dependiente de la cantidad diaria que reciben de alimento las hembras.

3. El test de homogeneidad permite determinar si varias muestras que estudian el mismo carácter *A* han sido tomadas o no de la misma población, respecto de dicha característica *A*. Por ejemplo, hemos seleccionado varios grupos de individuos de una población a los que hemos sometido a la misma dosis de diferentes metales y queremos determinar si los metales afectan de forma diferente a la supervivencia de los individuos.

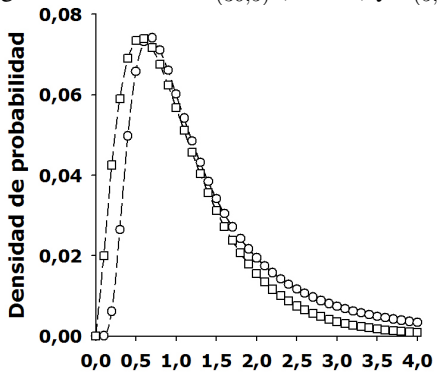
III.2.4. *F* de Fisher-Snedecor

La función de densidad de probabilidad de la distribución *F* de Fisher-Snedecor viene dada por la siguiente expresión:

$$f(x) = \frac{\Gamma\left(\frac{\upsilon+\omega}{2}\right)}{\Gamma\frac{\upsilon}{2}\Gamma\frac{\omega}{2}}\upsilon^{\frac{\upsilon}{2}}\omega^{\frac{\omega}{2}}\frac{x^{\frac{\upsilon}{2}-1}}{(\omega+x\upsilon)^{\frac{\upsilon+\omega}{2}}}$$

donde υ y ω son los grados de libertad del numerador y denominador respectivamente, siendo *x* no negativo. Al depender de dos tipos de grados de libertad, la función de densidad puede tener diversas formas (Figura III.4).

Figura III.4. Funciones de densidad de la distribución de *F* Fisher-Snedecor para distintos grados de libertad $F_{(30,5)}$ (círculo) y $F_{(5,30)}$ (cuadrado).



Los valores críticos de la distribución *F* Fisher-Snedecor con distintos niveles de significación y distintos grados de libertad se muestran en la Tabla 4 del Apéndice I en Guisande et al. (2011).

Esta distribución se usa principalmente en dos tipos de situaciones, requiriéndosele en ambos casos que la distribución de las variables sea Normal:

1. Para probar si dos muestras provienen de poblaciones que poseen varianzas iguales. Esta prueba es útil para determinar si una población Normal tiene una mayor variación que la otra y es importante ya que, a la hora de comparar medias, varios estadísticos presentan como requisito la homogeneidad de varianzas.

2. También se aplica cuando se trata de comparar simultáneamente varias medias poblacionales.

EJEMPLOS

La longitud del cuerpo de una especie de una determinada población se distribuye según una Normal de μ = 10,8 cm y σ = 3,7 cm.

Caso 1. Calcular la probabilidad de que un individuo tenga un tamaño menor de 8,9 cm.

Resultado: 0,3038

Caso 2. Calcular la probabilidad de que un individuo tenga un tamaño entre 8,9 y 12,4 cm.

Resultado: 0,3635

Value

Se obtiene el valor de probabilidad, para la media y desviación indicada, considerando una distribución Normal

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Examples

```
## Not run:

#Caso 1
III1(valor=8.9, media=10.8, desvi=3.7, prob.mayor=FALSE)

#Caso 2

III1(valor=12.4, media=10.8, desvi=3.7, prob.mayor=FALSE)-III1(valor=8.9,
media=10.8, desvi=3.7, prob.mayor=FALSE)


## End(Not run)
```

III2 *DISTRIBUCION BINOMIAL*

Description

Calcula la probabilidad de una distribución Binomial.

Usage

```
III2(n=NULL, x=NULL, p=NULL, opcion=1)
```

Arguments

n	Número total de individuos.
x	Número de individuos.
p	Probabilidad de éxito.
opcion	1. Exactamente la probabilidad p(X=x). 2. Probabilidad de que sea mayor o igual p(X>=x). 3. Probabilidad de que sea estrictamente mayor p(X>x). 4. Probabilidad de que sea menor o igual p(X<=x). 5. Probabilidad de que sea estrictamente menor p(X<x).

Details

III. DISTRIBUCIÓN

III.3. DISTRIBUCIONES PARA VARIABLES DISCRETAS

III.3.1. Distribución binomial

Una variable presenta una distribución binomial cuando sólo tiene dos posibles resultados: «éxito» y «fracaso», siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones, es decir, ni la probabilidad de éxito ni la de fracaso cambian de una prueba a otra, y además el resultado de cada prueba es independiente del resultado de las demás pruebas. La probabilidad de éxito está representada por *p* y la probabilidad de fracaso se representa por *q* = 1 - *p*.

En el caso de las variables discretas en lugar de la función de densidad se utiliza la función de probabilidad o de cuantía, que da una probabilidad para cada valor de la variable. La función de probabilidad binomial viene expresada por la siguiente fórmula:

$$f(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

EJEMPLOS

La aplicación de un determinado tratamiento a un grupo de individuos que presentan un parásito produce una mejoría en un 67% de los casos. Si se aplica el tratamiento a 8 individuos:

El valor de *p* = 0,67 y, por tanto, el valor de *q* es 0,33.

Caso 1. ¿Cuál es la probabilidad de que mejoren 7 individuos en una muestra de 8?

$$\frac{8!}{7!(8-7)!}0,67^7 0,33^{8-7} = 0,16$$

Resultado: 0,16

Caso2. ¿Cuál es la probabilidad de que al menos mejoren 3 individuos?

La probabilidad de que al menos mejoren 3 individuos será 1 menos la probabilidad que mejoren 2 individuos menos la probabilidad de que mejore 1 individuo menos la probabilidad de que no mejore ninguno:

$$1 - \frac{8!}{2!(8-2)!}0,67^2 0,33^{8-2} - \frac{8!}{1!(8-1)!}0,67^1 0,33^{8-1} - \frac{8!}{0!(8-0)!}0,67^0 0,33^{8-0} = 0,981$$

Resultado: 0,981

Value

Se obtiene el valor de probabilidad considerando una distribución Binomial.

Examples

```
## Not run:

#Caso 1
III2(n=8, x=7, p=0.67, opcion=1)

#Caso 2
1-III2(n=8, x=2, p=0.67, opcion=1)-III2(n=8, x=1, p=0.67, opcion=1)- III2(n=8,
x=0, p=0.67, opcion=1)


## End(Not run)
```

III3 *DISTRIBUCION HIPERGEOMETRICA*

Description

Calcula la probabilidad de una distribución Hipergeométrica.

Usage

```
III3(N=NULL, N.p=NULL, n=NULL, x=NULL, opcion=1)
```

Arguments

N	Número total de elementos en la población.
N.p	Valor de Np, número de elementos con probabilidad inicial p.
n	Tamaño de muestra.
x	Valor de la variable aleatoria para el que se quiere calcular la probabilidad.
opcion	1. Exactamente la probabilidad p(X=x). 2. Probabilidad de que sea mayor o igual p(X>=x). 3. Probabilidad de que sea estrictamente mayor p(X>x). 4. Probabilidad de que sea menor o igual p(X<=x). 5. Probabilidad de que sea estrictamente menor p(X<x).

Details

III. DISTRIBUCIÓN

III.3. DISTRIBUCIONES PARA VARIABLES DISCRETAS

III.3.2. Distribución hipergeométrica

En la distribución hipergeométrica la variable también es aleatoria y dicotómica como en la distribución binomial, pero se diferencia de esta última en dos características importantes: la población es finita, mientras que en la binomial puede ser infinita y, además, la probabilidad *p* cambia, no es

constante, ya que el resultado de cada prueba depende del resultado de las anteriores. La función de probabilidad viene expresada por la siguiente fórmula:

$$f(x) = \frac{\frac{N_p!}{x!(N_p-x)!}\frac{N_q!}{(n-x)!(N_q-n+x)!}}{\frac{N!}{n!(N-n)!}}$$

donde N_p y N_q son el número de elementos con probabilidad inicial p y q, respectivamente (es decir, $p = N_p/N$ y $q = N_q/N$), N el número total de elementos y n el número de elementos de la muestra extraída de los N de la población.

EJEMPLOS

En un lote de 30 vacunas se ha comprobado que 8 están en mal estado. Ya se han utilizado 4 vacunas del lote.

Caso 1. ¿Cuál es la probabilidad de que al menos una de las vacunas suministradas esté en malas condiciones?

Se calcula la probabilidad de que ninguna de las vacunas suministradas esté en mal estado

$$1 - \frac{\frac{22!}{4!(22-4)!}\frac{8!}{(4-4)!(8-4+4)!}}{\frac{30!}{4!(30-4)!}} = 0,733$$

Resultado: 0,733

Caso 2. ¿Cuál es la probabilidad de que 3 de las vacunas suministradas estén en malas condiciones?

$$\frac{\frac{22!}{1!(22-1)!}\frac{8!}{(4-1)!(8-4+1)!}}{\frac{30!}{4!(30-4)!}} = 0,045$$

Resultado: 0,045

Value

Se obtiene el valor de probabilidad considerando una distribución Hipergeométrica.

Examples

```
## Not run:

#Caso 1.
1- III3(N=30, N.p=8, n=4, x=0, opcion=1)

#Caso 2
III3(N=30, N.p=8, n=4, x=3, opcion =1)


## End(Not run)
```

III4 *DISTRIBUCION DE POISSON*

Description

Calcula la probabilidad de una distribución de Poisson.

Usage

```
III4(lambda=NULL, x=NULL, opcion=1)
```

Arguments

`lambda`	Valor del parámetro lambda.
`x`	Valor sobre el cual se quiere calcular la probabilidad.
`opcion`	1. Exactamente la probabilidad p(X=x). 2. Probabilidad de que sea mayor o igual p(X>=x). 3. Probabilidad de que sea estrictamente mayor p(X>x). 4. Probabilidad de que sea menor o igual p(X<=x). 5. Probabilidad de que sea estrictamente menor p(X<x).

Details

III. DISTRIBUCIÓN

III.3. DISTRIBUCIONES PARA VARIABLES DISCRETAS

III.3.3. Distribución de Poisson

Un proceso de Poisson es un proceso de sucesos independientes que se caracteriza por:

1. El número de sucesos en dos intervalos distintos siempre es independiente.

2. La probabilidad de que un suceso ocurra en un intervalo infinitesimal es proporcional a la longitud del intervalo.

3. La probabilidad de que ocurra más de un suceso en un intervalo muy pequeño *h* es 0.

4. Los sucesos son expresados por unidad de área, tiempo, etc.

La distribución de Poisson describe el número de sucesos en una unidad de tiempo de un proceso Poisson. Muchos fenómenos se modelan como un proceso de Poisson, por ejemplo el número de accidentes en una determinada zona de una carretera.

Las diferencias más importantes con respecto a la distribución binomial son que esta distribución se aplica a sucesos que pueden tener una probabilidad muy baja y, además, el tamaño de *n* es infinito. En algunos casos la distribución de Poisson se utiliza como aproximación a la Binomial cuando *n* es muy grande y, por tanto, es difícil el cálculo de la binomial y, además, cuando la probabilidad de alguno de los eventos es muy baja. La función de probabilidad de la distribución de Poisson se expresa por la siguiente fórmula:

$$f(x) = \frac{\lambda^x}{x! e^{\lambda}}$$

donde λ es la media o promedio de sucesos por unidad de tiempo y x es la variable que indica el número de sucesos.

EJEMPLOS

Caso 1. La abundancia media de una especie es de 23 individuos por 100 m^2. Como se trata de sucesos por unidad de área se utiliza una Poisson. ¿Cuál es la probabilidad de no encontrar ningún individuo en 25 m^2?

$$\lambda = \frac{23 individuos * 25 m^2}{100 m^2} = 5,75$$

$$\frac{5,75^0}{0! e^{5,75}} = 0,0031$$

Resultado: 0,0031

Caso 2. El número nedio de tigres observados es de 120 en 30 días. Como se trata de sucesos por unidad de tiempo se utiliza una Poisson. ¿Cuál es la probabilidad de ver 5 tigres en 10 días?

$$\lambda = \frac{120 tigres * 10 d}{30 d} = 40$$

$$\frac{40^5}{5! e^{40}} = 3,63^{-12}$$

Resultado: $3,63^{-12}$

Value

Se obtiene el valor de probabilidad considerando una distribución de Poisson.

Examples

```
## Not run:

#Caso 1
III4(lambda=5.75, x=0, opcion=1)

#Caso 2
III4(lambda=40, x=5, opcion=1)


## End(Not run)
```

IV1 *INTERVALO DE CONFIANZA DE LA MEDIA DE UNA POBLACION NORMAL*

Description

Calcula el intervalo de confianza de la media de una población Normal.

Usage

```
IV1(data, varInteres, varAgrupacion=NULL, alfa=0.05, total=FALSE,
file="Output.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere calcular el intervalo de confianza.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
alfa	Nivel de confianza del intervalo.
total	Valor lógico que si es FALSE significa que no se ha muestreado toda la población, que será en la mayoría de los estudios, y si es TRUE significa que se ha muestreado toda la población.
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

IV. INTERVALOS DE CONFIANZA

En la función II1 se describió la forma de estimar determinados parámetros que nos permiten tener información sobre las variables que estamos estudiando, como por ejemplo los distintos tipos de media, varianza, desviación típica, etc. En la mayoría de los casos estos parámetros se estiman a partir de los datos de una muestra, y no del conjunto total de la población. Para medir la fiabilidad de la estimación es necesario calcular el intervalo de confianza del parámetro.

El intervalo de confianza de un parámetro al nivel de confianza 1 - α es aquel que cumple la propiedad de que la probabilidad de que sus extremos tomen valores tales que el parámetro esté comprendido entre ellos es igual a 1 - α (Viedma, 1989). Los extremos del intervalo de confianza de un parámetro se denominan límites de confianza.

IV.1. INTERVALO DE CONFIANZA DE LA MEDIA DE UNA POBLACIÓN NORMAL

IV.1.1. Desviación típica conocida

En el supuesto de que se conozca la desviación típica de la población, el intervalo de confianza de la media al nivel de $1-\alpha$ $\left(I_{\mu}^{1-\alpha}\right)$ se calcula por medio del intervalo estimador siguiente:

$$I_{\mu}^{1-\alpha}=\left(\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I en Guisande et al., 2011).

IV.1.2. Desviación típica desconocida

Lo más frecuente es que estemos trabajando con unos cuantos elementos del total que forma la población. Esto quiere decir que no conocemos la desviación típica del conjunto de toda la población. En este tipo de situación, el intervalo de confianza se calcula de dos formas diferentes dependiendo del tamaño de la muestra.

IV.1.2.1. Tamaño de muestra pequeño (<30)

Si el tamaño de la muestra es pequeño, en vez de utilizar la distribución Normal y, por tanto, el estadístico *Z*, lo que se utiliza es la distribución *t* de Student con *n* - 1 grados de libertad. El intervalo de confianza de la media al nivel de confianza $1-\alpha$ $\left(I_{\mu}^{1-\alpha}\right)$ es:

$$I_{\mu}^{1-\alpha}=\left(\bar{x}-t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}},\bar{x}+t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}\right)$$

Se utilizan los valores críticos de la distribución *t* de Student (Tabla 2 del Apéndice I en Guisande et al., 2011).

IV.1.2.2. Tamaño de muestra grande (≥30)

Cuando el tamaño de la muestra tiende a infinito, la *t* de Student tiende a una distribución Normal estándar. Por ello, la mayoría de las tablas de la distribución *t* tienen solamente los valores correspondientes a un número de grados de libertad entre 1 y 30; cuando es mayor se utiliza directamente la tabla de la Normal como aproximación, ya que las diferencias que se obtienen son muy pequeñas.

Los programas estadísticos no necesitan utilizar esta aproximación Normal, ya que pueden calcular el valor exacto de *t* cualquiera que sea el número de grados de libertad. Las diferencias que se pueden observar entre los intervalos que se obtienen con STATISTICA o SPSS y la metodología que se explica a continuación se deben, por lo tanto, a que los programas estadísticos no tienen en cuenta si el número de datos es pequeño (<30) o grande ($\geq$ 30), ya que siempre trabajan con la distribución *t* de Student y nunca utilizan la aproximación Normal.

Si el tamaño de la muestra es grande se usa la misma expresión descrita anteriormente para el caso de desviación típica conocida, pero con la diferencia de que se utiliza la cuasidesviación típica muestral (*s*) en vez de la desviación típica poblacional (σ).

$$I_{\mu}^{1-\alpha}=\left(\bar{x}-z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I en Guisande et al., 2011).

EJEMPLO

Datos de altura y peso en hombres y mujeres de distintos países y ciudades. Se pretende calcular los intervalos de confianza al 95% de la altura y el peso, tanto para hombres como para mujeres. Además los cálculos se hacen para cada país y ciudad.

Los resultados obtenidos se muestran en la siguiente tabla, que indica el valor medio estimado en cada grupo, y los límites de confianza al 95% para la correspondiente media poblacional. Se supone que la variable tiene distribución Normal, y que la desviación típica poblacional es desconocida, y debe ser estimada con los datos muestrales.

Country	City	Gender	variable	LimInf	Media	LimSup
1	1	Female	Height	145,470696	158	170,529304
1	2	Female	Height	160,807156	168,75	176,692845
2	3	Female	Height	155,126813	160,666667	166,206521
2	4	Female	Height	160,329683	167	173,670317
1	1	Male	Height	165,052317	173,333333	181,614349
1	2	Male	Height	168,693723	173,166667	177,63961
2	3	Male	Height	169,789786	179,428571	189,067357

Value

Se obtiene un archivo CSV con los límites inferior y superior del intervalo de confianza y el valor de la media.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Viedma, J.A. (1989) Bioestadística. *Métodos en Medicina y Biología*. Librería Universitaria, Alicante.

Examples

```
## Not run:

data(ZIV1)

IV1(data=ZIV1, varInteres=c("Height","Weight"),
varAgrupacion=c("Country","City","Gender"), alfa=0.05, total=FALSE)


## End(Not run)
```

IV2 *INTERVALO DE CONFIANZA DE LA DIFERENCIA DE MEDIAS DE POBLACIONES NORMALES*

Description

Calcula el intervalo de confianza de la diferencia de medias de poblaciones Normales.

Usage

```
IV2(data, varInteres, varSel, varAgrupacion=NULL, alfa=0.05,  file="Output.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere calcular el intervalo de confianza.
varSel	Se especifica la variable o variables que tienen los dos grupos a estudiar. Es importante mencionar que si la variable fue definida anteriormente como variable agrupación, no es posible usarla en esta sección. Si la variable tiene más de 2 grupos, es necesario especificar cuales son los grupos a estudiar. En el ejemplo se calcula el intervalo de confianza de la diferencia de las medias entre la ciudad 1 y la ciudad 2. Si la variable solo tiene 2 grupos, como por ejemplo «Sexo», no es necesario especificar los grupos a analizar.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
alfa	Nivel de confianza del intervalo.
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

IV. INTERVALOS DE CONFIANZA

IV.2. INTERVALO DE CONFIANZA DE LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES

Cuando se comparan medias de distintas poblaciones, es importante estimar el intervalo de confianza de la diferencia entre las medias. Al igual que con el intervalo que se analizó en la función IV1, el cálculo es diferente dependiendo de que se conozca la desviación típica poblacional (varianza conocida) o deba ser estimada con datos muestrales (varianza desconocida).

IV.2.1. Varianzas conocidas

El intervalo de confianza $\mu_1 - \mu_2$ al nivel de confianza $1 - \alpha$ $\left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I en Guisande et al., 2011).

IV.2.2. Varianzas desconocidas

En el supuesto de que se quiera comparar dos medias, siendo las variables independientes y con distribución Normal, pero se desconocen las varianzas de la población, tenemos que aplicar otras expresiones que son diferentes dependiendo del tamaño de la muestra.

IV.2.2.1. Tamaño de muestra grande (≥30)

Para muestras grandes el intervalo de confianza $\mu_1 - \mu_2$ al nivel de confianza $1-\alpha$ $\left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

Como vemos, la única diferencia con respecto a la situación en la que se conocían las varianzas, es que se usa la cuasivarianza en vez de la varianza.

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I en Guisande et al., 2011).

IV.2.2.2. Tamaño de muestra pequeño (<30)

Se pueden dar dos situaciones: varianzas iguales o diferentes.

IV.2.2.2.1. Varianzas iguales

Suponiendo que desconocemos las varianzas, pero sabemos que no existen diferencias significativas entre ellas, en este caso se utiliza la distribución *t* de Student y el intervalo de confianza $\mu_1 - \mu_2$ al nivel de confianza $1-\alpha$ $\left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\begin{array}{l}(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}, \\ (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\end{array}\right]$$

Se utilizan los valores críticos de la distribución *t* de Student (Tabla 2 del Apéndice I en Guisande et al., 2011).

IV.2.2.2.2. Varianzas diferentes

En aquellos casos en los que se desconocen las varianzas y, además, existen diferencias significativas entre ellas, o simplemente no se sabe si las varianzas son iguales o no, el intervalo $\mu_1 - \mu_2$ al nivel de confianza $1-\alpha$ $\left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ se calcula por medio de la siguiente forma:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

donde f son los grados de libertad y se denomina aproximación de Welch, que se calcula del siguiente modo, tomando f el valor del entero más próximo:

$$f = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Se utilizan los valores críticos de la distribución *t* de Student (Tabla 2 del Apéndice I en Guisande et al., 2011).

EJEMPLO

Datos de altura y peso en hombres y mujeres de distintos países y ciudades. El objetivo es estimar el intervalo de confianza de la diferencia de medias en el peso y en la altura, en hombres y mujeres, entre ciudades. Los cálculos se realizan para cada país y por sexos.

Los resultados obtenidos se muestran en la siguiente tabla, que presenta, para la selección efectuada, los tamaños de muestra, las medias estimadas y su diferencia, el intervalo de confianza al 95% para la diferencia de medias poblacional, y los valores P del contraste de nulidad de la diferencia de medias (un valor menor que 0,05 evidencia que las medias son distintas), y del contraste de igualdad de varianzas (un valor P mayor que 0,05 permite aceptar que las varianzas son iguales en ambos grupos). En todos los casos se supone que la variable tiene distribución Normal y que la varianza poblacional es desconocida, por lo que se estima con los datos de la muestra.

Tipo.de.Grupo	Country	Gender	variable	Freq	n1	n2	media1	media2	limInf	diferencia	limSup	valorPMedia	valorPVar
City.1_2	1	Female	Height	8	4	4	158	168,75	-22,1561705	-10,75	0,65617052	0,06058788	0,47373066
City.1_2	1	Male	Height	12	6	6	173,333333	173,166667	-7,99135135	0,16666667	8,32468468	0,96458869	0,20255664
City.1_2	1	Female	Weight	8	4	4	58,4	62,475	-9,89413761	-4,075	1,74413761	0,13745077	0,28337707
City.1_2	1	Male	Weight	12	6	6	72,7333333	73,0833333	-8,54041391	-0,35	7,84041391	0,91992711	0,02103689

Value

Se obtiene un archivo CSV con los límites inferior y superior del intervalo de confianza de la diferencia de medias de poblaciones Normales.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Examples

```
## Not run:

data(ZIV1)

IV2(data=ZIV1, varInteres=c("Height","Weight"),
varAgrupacion=c("Country","Gender"), alfa=0.05, varSel=list(c("City","1","2")))


## End(Not run)
```

`IV3` *INTERVALO DE CONFIANZA DE LA VARIANZA Y DESVIACION TIPICA DE UNA POBLACION NORMAL*

Description

Calcula el intervalo de confianza de la varianza y desviación típica de una población Normal.

Usage

```
IV3(data, varInteres, varAgrupacion=NULL, alfa=0.05, tipoInterv="desv.est",
file="Output.csv", na="NA", dec= ",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables de cuya varianza se quiere calcular el intervalo de confianza.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
alfa	Nivel de confianza del intervalo.
tipoInterv	Se especifica si se realiza el intervalo de confianza de la varianza "Varianza" o de la desviación típica "desv.est".
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

IV. INTERVALOS DE CONFIANZA

IV.3. INTERVALO DE CONFIANZA DE LA VARIANZA Y DESVIACIÓN TÍPICA DE UNA POBLACIÓN NORMAL

A veces es importante estudiar la mayor o menor concentración de valores alrededor de la media. Por ejemplo, cuando se estudia el tiempo que permanece un individuo inmune después de suministrarle una vacuna, no solo es importante conocer la duración media, sino también que la variabilidad de la duración del efecto no sea muy grande de unos individuos a otros. Para ello se puede estimar la varianza o la desviación típica mediante un intervalo de confianza.

El intervalo de la varianza al nivel de confianza $1-\alpha$ $\left(I_{\sigma^2}^{1-\alpha}\right)$ se calcula por medio del intervalo estimador siguiente, donde s^2 es la cuasivarianza de la muestra:

$$I_{\sigma^2}^{1-\alpha} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right)$$

El intervalo de confianza de la desviación típica al nivel de confianza $1-\alpha$ $\left(I_{\sigma}^{1-\alpha}\right)$ se estima por medio del intervalo estimador siguiente:

$$I_{\sigma}^{1-\alpha} = \left(\frac{\sqrt{n-1}s}{\sqrt{\chi^2_{\frac{\alpha}{2};n-1}}}, \frac{\sqrt{n-1}s}{\sqrt{\chi^2_{1-\frac{\alpha}{2};n-1}}} \right)$$

Se utilizan los valores críticos de la distribución χ^2 (Tabla 3 del Apéndice I en Guisande et al., 2011).

EJEMPLO

Datos del tiempo (en meses) que permanecen inmunes un grupo de mujeres y hombres después de que se les suministra una vacuna. Se quiere saber el intervalo de confianza al 95%, de la desviación típica de las mujeres y los hombres.

Los resultados obtenidos se muestran en la siguiente tabla, que muestra la desviación típica estimada junto con los límites de confianza al 95%, suponiendo que la variable tiene distribución Normal.

Gender	variable	LimInf	DesviacionEstandar	LimSup
F	Time	4,52279729	6,30718963	10,4114996
M	Time	3,93079785	5,625736315	9,87279339

Value

Se obtiene un archivo CSV con los límites inferior y superior del intervalo de confianza de la varianza y desviación típica de una población Normal.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Examples

```
## Not run:

data(ZIV3)

IV3(data=ZIV3, varInteres=c("Time"), varAgrupacion=c("Gender"), alfa=0.05,
tipoInterv="desv.est")


## End(Not run)
```

IV4 — *INTERVALO DE CONFIANZA DE LA RAZON DE VARIANZAS Y PRUEBA DE HOMOGENEIDAD DE VARIANZAS ENTRE POBLACIONES NORMALES*

Description

Calcula el intervalo de confianza de la razón de varianzas de poblaciones Normales y la prueba de homogeneidad de varianzas.

Usage

```
IV4(data, varInteres, varSel, varAgrupacion=NULL, alfa=0.05, file="Output.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables de las que se quiere calcular el intervalo de confianza para el cociente de varianzas y el contraste de homogeneidad.
varSel	Variable o variables que especifican los dos grupos a estudiar. Es importante mencionar que si la variable fue definida anteriormente como variable agrupación, no es posible usarla en esta sección. Si la variable tiene más de 2 grupos, es necesario especificar cuales son los grupos a estudiar.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
alfa	Nivel de confianza del intervalo.
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

IV. INTERVALOS DE CONFIANZA

IV.4. INTERVALO DE CONFIANZA DE LA RAZÓN DE VARIANZAS DE DOS POBLACIONES NORMALES

Determinar el intervalo de confianza de la razón de varianzas es importante, por ejemplo cuando queremos comparar la variabilidad de dos métodos a la hora de realizar un análisis bioquímico o la variabilidad de dos aparatos que nos dan el mismo tipo de medida. La comparación de varianzas entre las variables también es importante porque uno de los requisitos de los tests paramétricos de contrastes de homogeneidad, *t* test (IX1,IX2), ANOVA (IX3), ANCOVA (IX4), etc., es que exista homogeneidad de varianzas entre las variables que se comparan.

El intervalo de la razón de las varianzas al nivel de confianza $1-\alpha$ $\left(I^{1-\alpha}_{\sigma_1^2/\sigma_2^2}\right)$ se calcula de la siguiente forma:

$$I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} = \left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2};n_1-1,n_2-1}}, \frac{s_1^2}{s_2^2}F_{\frac{\alpha}{2};n_2-1,n_1-1}\right)$$

El intervalo de confianza de las desviaciones típicas al nivel de confianza $1-\alpha$ $\left(I^{1-\alpha}_{\sigma_1/\sigma_2}\right)$ se calcula de la siguiente forma:

$$I^{1-\alpha}_{\sigma_1/\sigma_2} = \left(\frac{s_1}{s_2}\frac{1}{\sqrt{F_{\frac{\alpha}{2};n_1-1,n_2-1}}}, \frac{s_1}{s_2}\sqrt{F_{\frac{\alpha}{2};n_2-1,n_1-1}}\right)$$

Se utilizan los valores críticos de la distribución *F* (Tabla 4 del Apéndice I en Guisande et al., 2011).

EJEMPLO

Datos de altura y peso en hombres y mujeres de distintos países y ciudades. El objetivo es estimar el intervalo de confianza de la razón de varianzas en el peso y en la altura, en hombres y mujeres, entre diferentes ciudades, y determinar si las varianzas son homogéneas.

Por ejemplo, en el caso de la comparación de la altura de los hombres entre las ciudades 1 y 2, el intervalo de confianza de la razón de varianzas es entre 0,479 y 24,494. Tanto la probabilidad de la razón de varianzas (p = 0,202), como la del test de Levene (p = 0,083) y la del test Brown & Forsythe (p = 0,217) son mayores de 0,05, por lo tanto se acepta la hipótesis nula de que no existen diferencias significativas en las varianzas de la altura de los hombres entre las ciudades 1 y 2, es decir, que existe homogeneidad de varianzas.

Sin embargo, al comparar el peso de los hombres entre la ciudades 1 y 2, la probabilidad de la razón de varianzas (p = 0,021) y la del test de Levene (p = 0,007) son menores de 0,05, aunque no la del test Brown & Forsythe (p = 0,112), por lo tanto, es razonable pensar que se rechaza la hipótesis nula y, por tanto, existen diferencias significativas entre las varianzas de la altura de los hombres entre las ciudades 1 y 2, es decir, que no existe homogeneidad de varianzas.

Tipo.de.Grupo	Country	Gender	variable	Freq	n1	n2	desv.Est1	desv.Est2	limInf	razon	limSup	valorPVar
City.1_2	1	Female	Height	8	4	4	7,8740	4,9917	0,1612	2,4883	38,4172	0,4737
City.1_2	1	Male	Height	12	6	6	7,8909	4,2622	0,4796	3,4275	24,4944	0,2026
City.1_2	1	Female	Weight	8	4	4	4,2560	2,1235	0,2602	4,0170	62,0192	0,2834
City.1_2	1	Male	Weight	12	6	6	7,8173	2,3878	1,4998	10,7180	76,5952	0,0210

Value

Se obtiene un archivo CSV con el intervalo de confianza de la razón de varianzas y los test de Levene y Brown & Forsythe, los cuales permiten determinar si hay homogeneidad de varianzas en los grupos analizados.

Examples

```
## Not run:

data(ZIV1)

IV4(data=ZIV1, varInteres=c("Height","Weight"), varAgrupacion=c("Country","Gender"),
alfa=0.05, varSel=list(c("City","1","2"),c("City","1","3")))


## End(Not run)
```

IX1 t-*TEST PARA MUESTRAS INDEPENDIENTES*

Description

Se aplica el test de la *t* para muestras independientes y se puede representar los datos con un boxplot o un beanplot.

Usage

```
IX1(data, varInteres, varPoblacion, varComparar1, varComparar2,
trans=NULL, grafico="Boxplot", PAR=NULL, ResetPAR=TRUE, YLAB=NULL,
XLAB=NULL, OrderCat=NULL, LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL,
BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable dependiente.
varPoblacion	Define cual es la variable que actúa como factor.
varComparar1	Primer grupo a comparar de la variable población.
varComparar2	Segundo grupo a comparar de la variable población.
trans	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3 9. EXP (exponencial) 10. ASN (arcoseno)
grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
OrderCat	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

Para el contraste de homogeneidad en el caso de variables cuantitativas contamos con dos tipos de análisis, los **paramétricos**, que proceden de un modelo que obliga a cumplir ciertos supuestos acerca de los parámetros y la distribución de probabilidad de la población de la que se extrajo la muestra, y los **no paramétricos**, que son menos restrictivos para poder ser aplicados. Se debe tener en cuenta que la estadística paramétrica es más precisa, pero las posibilidades de ser aplicada, como veremos más adelante, están limitadas.

IX.1 PRUEBAS PARAMÉTRICAS

IX.1.1. Requisitos

Antes de utilizar las pruebas paramétricas, es indispensable comprobar ciertos requisitos (supuestos) para su aplicación. Los más importantes a tener en cuenta son:

1. Distribución Normal de las poblaciones.

2. Igualdad de varianzas.

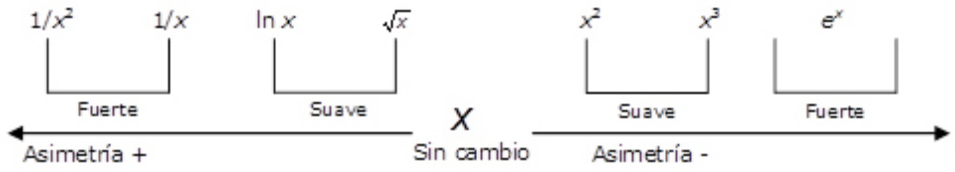
Para la comprobación de la hipótesis de Normalidad hay que realizar un contraste de bondad de ajuste y podemos emplear la prueba de Kolmogorov-Smirnov o la de Shapiro-Wilk explicadas en la función VI3. La prueba de Normalidad se debe hacer sobre los residuos o diferencias entre los valores observados de la variable dependiente y los valores que predice el modelo una vez estimados los parámetros. Si los residuos son Normales, también la variable dependiente tendrá distribución Normal en cada nivel de cada factor. La hipótesis de igualdad de varianzas se verifica comúnmente como se detalló en la función IV4. o como se explica en los ejemplos que veremos a continuación.

Si nuestros datos no cumplen el supuesto de distribución Normal se puede intentar transformar la variable para que adopte una distribución Normal, como veremos en el siguiente apartado. En general, la transformación que "normaliza" los datos también consigue una igualdad de varianzas. En el caso de que no se consiguiera la distribución Normal y la homogeneidad de varianzas, ni siquiera después de transformar los datos, la mejor alternativa sería utilizar un método no paramétrico, aunque a veces algunos tests paramétricos se aplican no suponiendo la igualdad de varianzas, como por ejemplo el t-test modificado por Welch-Satterthwaite.

IX.1.2. Transformaciones

Si decidimos transformar la variable, tendremos varias posibilidades según su tipo de distribución (asimétrica positiva o negativa, ver función II1). La bibliografía nos habla de la llamada escalera de las transformaciones de Tukey, la cual muestra el tipo de transformación recomendada según sea la intensidad de la asimetría o la dirección en la que van los casos extremos (Sánchez, 1999). La Figura IX.1 es una modificación del gráfico realizado por Erickson & Nosanchuk (1977) y adaptada por Guisande et al. (2011).

Figura IX.1. Escalera de transformaciones.



Para elegir la transformación adecuada hay que tener en cuenta la asimetría y la curtosis. Como la gráfica solo muestra la asimetría, la influencia de la curtosis puede dar lugar a que valores inferiores

a la media requieran una transformación de raíz cuadrada y valores superiores a la media requieran elevar al cuadrado. Se debe elegir la transformación que haga más próximos a cero simultáneamente ambos coeficientes.

IX.1.3. *t*-test

IX.1.3.1. Muestras independientes

El *t*-test es el método más común para evaluar las diferencias entre las medias de dos grupos independientes, por ejemplo dos grupos de peces sometidos a dietas diferentes. Para esta prueba, idealmente los sujetos deben asignarse aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al tratamiento (o falta de tratamiento) y no a otros factores. Al aplicar este análisis debemos asegurarnos de que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias. La hipótesis nula (H_0) con la que comúnmente se trabaja es que las medias son iguales. En general no se exige que el número de observaciones en ambos grupos sea igual. Si la H_0 es cierta y suponemos una igualdad de varianzas, el estadístico (*t*-test) seguirá una distribución *t* de Student con n_x + n_y - 2 grados de libertad. Existe también la opción de utilizar el estadístico *t* de Student asumiendo una no igualdad de varianzas, al trabajar con los grados de libertad de Welch-Satterthwaite.

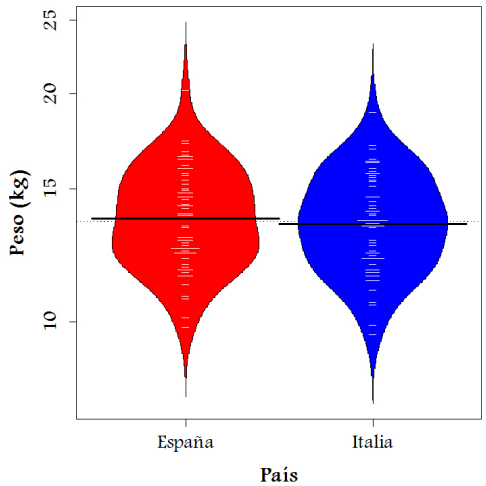
FUNCIONES

Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). La función shapiro.test del paquete base stats para realizar el test de Shapiro-Wilk. El test de Levene se realiza con la función levene.test del paquete lawstat (Gastwirth et al., 2013). La asimetría y la curtosis se realizan con las funciones skewness y kurtosis, respectivamente, del paquete e1071 (Meyer et al., 2014).

EJEMPLO

Datos de peso y altura de niños y niñas entre los 2 y 4 años en Italia y España. El objetivo del trabajo es determinar si existen diferencias significativas en el peso, considerando conjuntamente a niños y niñas, entre ambos países. La Figura IX.2 muestra la mediana y la distribución de los datos del peso de niños y niñas en Italia y España.

Figura IX.2. Beanplot comparando el peso de niños y niñas entre España e Italia.



El test de Shapiro-Wilk muestra que las dos poblaciones de niños en España (p = 0,215) e Italia (p = 0,476) tienen distribución Normal. Además se cumple el requisito de homogeneidad de varianzas, tanto considerando la media (p = 0,644) como la mediana (p = 0,643). Por lo tanto, no es necesario realizar de nuevo el análisis transformando la variable dependiente. En el caso de que no se cumpliera alguno de los dos requisitos, también sale información sobre el asimetría y la curtosis de cada una de las poblaciones, lo cual ayudaría a elegir mejor la transformación a realizar.

```
datos3[, "poblaciones"]: España

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9735, p-value = 0.2152

------------------------------------------------------------
datos3[, "poblaciones"]: Italia

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9811, p-value = 0.4757


[[2]]

     classical Levene's test based on the absolute deviations from the mean
     ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.2151, p-value = 0.6437


[[3]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$valores
Test Statistic = 0.2153, p-value = 0.6435


[[4]]
                                España               Italia
"Sesgo" "0.431370485546375" "0.228137661533395"

[[5]]
                                   España                Italia
"Curtosis"  "0.137026398644089" "-0.411525978179505"
```

Por último, el test de la *t* muestra que no existen diferencias significativas en el peso de los niños y niñas entre España e Italia (t =-0,54, df = 117,8, p = 0,588), con valores medios de 13,8 y 13,6 kgs, respectivamente.

```
        Welch Two Sample t-test

data:  datos3$valores by datos3$populations
t = -0.5427, df = 117.81, p-value = 0.5884
alternative hypothesis: true difference in means between group Italy and group Spain is not equal to 0
95 percent confidence interval:
 -0.9375502  0.5342169
sample estimates:
mean in group Italy mean in group Spain
          13.60167            13.80333
```

Value

Se obtiene un archivo TXT con los resultados del *t*-test y se puede representar un boxplot o un beanplot.

References

Erikson, F.H. & Nosanchuk, T.A. (1977) *Understandig data*. McGraw Hill. Toronto.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Sánchez, J.J. (1999) *Manual de análisis estadístico de los datos*. Alianza Editorial. Madrid.

Examples

```
## Not run:

data(ZIX1)

#Comparación del peso de niños y niñas entre Italia y España

IX1(data = ZIX1, varInteres = "Weight", varPoblacion = "Country", varComparar1 = "Spain",
varComparar2 = "Italy", grafico = "Beanplot", PAR = c("cex.lab=1.6","font.lab=2",
"cex.axis=1.4","family = 'Andalus'", "omi=c(0,0.7,0,0)"), LabelCat = c("España","Italia"),
BEANPLOT = c("col = list(c('red','white'), c('blue','white'))", "ll = 0.04",
"ylab = 'Peso (kg)'", "xlab = 'País'"))


## End(Not run)
```

IX10 *CONTRASTES DE HOMOGENEIDAD NO PARAMETRICOS PARA K MUESTRAS INDEPENDIENTES*

Description

Se aplica el test de suma de rangos de Kruskal-Wallis. Además es posible representar los datos con un boxplot o un beanplot.

Usage

```
IX10(data, varInteres, Factor, ResetPAR=TRUE, grafico="Boxplot",
PAR=NULL,YLAB=NULL, XLAB=NULL, OrderCat=NULL, LabelCat=NULL,
COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL,
MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable dependiente.
Factor	Define cual es la variable que actúa como factor.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
OrderCat	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.2. PRUEBAS NO PARAMÉTRICAS

IX.2.2. Contrastes para *K*-muestras independientes

Los contrastes más utilizados en este caso son los de la mediana y la prueba de Kruskal-Wallis, los cuales se explican brevemente a continuación.

IX.2.2.1. Contraste de la mediana

Consiste en determinar el valor mediano de nuestras observaciones incluyendo todas las k muestras a analizar, y tipificar todos los valores como positivos si están por encima de la mediana y negativos si están por debajo.

Si las muestras fuesen homogéneas, en cada una de ellas aproximadamente la mitad de los valores deberían estar por encima de la mediana y la otra mitad por debajo.

Ésta será la hipótesis nula (que todas las muestras proceden de la misma población).

Con el número de datos positivos y negativos de cada muestra, construimos una tabla de contingencia 2x*k*, a la que se aplica un contraste χ^2 de homogeneidad con la hipótesis nula indicada, es decir, en cada muestra la frecuencia esperada de positivos es igual a las frecuencia esperada de negativos.

Este contraste es poco sensible, por lo que es más preciso utilizar la prueba de Kruskal-Wallis. Sin embargo, en casos en los que los órdenes tienen límites artificiales este contraste, en teoría, es el más adecuado.

IX.2.2.2. ANOVA de Kruskal-Wallis

Es el contraste no paramétrico análogo al análisis de varianza, por lo que se denomina ANOVA, aunque el estadístico de contraste tiene distribución Chi-cuadrado.

Es la prueba más utilizada para más de dos muestras independientes, siendo mucho más sensible que el test de la mediana.

A nivel teórico es una extensión del contraste *U* de Mann-Whitney para más de dos muestras. Esto quiere decir que es un test que mide la tendencia central de las muestras, teniendo como hipótesis nula que las poblaciones testadas tienen la misma mediana.

El desarrollo completo del cálculo de este estadístico se encuentra en Sokal & Rohlf (1981), en el que los pasos a seguir son:

1. Ordenar los datos en rangos (todas las muestras en conjunto).

2. Sumar los rangos de cada muestra y calcular el estadístico de contraste *H* que se compara con un valor de contraste tabulado.

3. Si *H* es mayor que el valor tabulado se rechaza la hipótesis nula, indicando que las muestras son diferentes.

FUNCIONES

Para el gráfico beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El ANOVA de Kruskal-Wallis se realiza con la función kruskal.test del paquete stats.

EJEMPLO

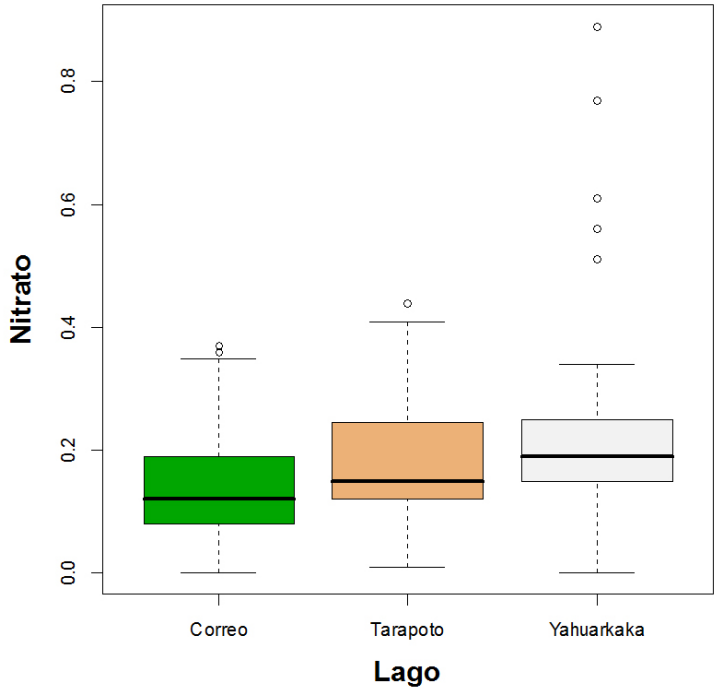
Datos de la concentración de nitrito, nitrato y amonio (μML) en lagos de la amazonia colombiana en dos meses diferentes.

El objetivo es determinar si existen diferencias en la concentración de los nitrato entre todos los lagos.

El resultado del test tiene una p = 0,131, por lo que se acepta la hipótesis nula de homogeneidad (Figura IX.13).

```
        Kruskal-Wallis rank sum test

data:  datos2$valores and datos2$factores
Kruskal-Wallis chi-squared = 4.0706, df = 2, p-value = 0.1306
```

Figura IX.13. Concentración de nitratos (μML) en varios lagos del Amazonas.



Value

Se obtiene un archivo TXT con los resultados del test y se puede representar un boxplot o un beanplot.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*. WH Freeman and Company, New York.

Examples

```
## Not run:

data(ZIX7)

IX10(data=ZIX7, varInteres="Nitrato", Factor="Lago")


## End(Not run)
```

IX11 *CONTRASTES DE HOMOGENEIDAD NO PARAMETRICOS PARA DOS MUESTRAS DEPENDIENTES*

Description

Se aplica la prueba de Wilcoxon para pares relacionados. Además se pueden representar los datos con un boxplot o un beanplot.

Usage

```
IX11(data, varTiempo, grafico="Boxplot", PAR=NULL, ResetPAR=TRUE,
YLAB="Variable problema", XLAB="Medidas en el tiempo", LabelCat=NULL,
COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varTiempo	Columnas en las que se ha medido en el tiempo la variable dependiente.
grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
YLAB	Leyenda del eje y en el boxplot y beanplot.
XLAB	Leyenda del eje x en el boxplot y beanplot.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.2. PRUEBAS NO PARAMÉTRICAS

IX.2.3. Contrastes para dos muestras dependientes

Las pruebas más utilizadas son la del contraste de los signos y la prueba de Wilcoxon para pares relacionados.

IX.2.3.1. Contraste de los signos

Este contraste mide la dirección de las diferencias entre dos muestras pareadas, por lo que es muy útil cuando se trabaja con muestras pareadas a las que se les puede identificar un incremento o descenso en la medición, pero es difícil de cuantificar este cambio. En otras condiciones, el test de Wilcoxon es más sensible.

Se estudia el número de veces que las diferencias de los valores de las dos muestras son positivas, negativas o si no hay diferencias. Con estas premisas podemos contrastar la hipótesis nula de que las muestras son homogéneas, ya que en ese caso habrá aproximadamente tantas diferencias negativas como positivas.

El desarrollo completo de la prueba aparece en Siegel & Castellan (1988), pero los pasos de forma resumida son:

1. Determinación de las diferencias (+ ó -).

2. Ajuste de las diferencias observadas a una distribución binomial. Cuando el número de datos es grande (mayor de 30 datos), se utiliza una aproximación a la distribución Normal (cálculo del estadístico *Z*) de la distribución binomial.

3. Determinación de la probabilidad asociada al estadístico de contraste para aceptar o rechazar la hipótesis nula de que las muestras son homogéneas.

IX.2.3.2. Prueba de Wilcoxon para pares relacionados

Es la prueba análoga a la *t* de Student para muestras relacionadas, siendo casi tan potente como ésta. En lo que se refiere a la sensibilidad, es un contraste mucho más potente que el anterior ya que, aunque también utiliza las diferencias entre los valores de cada caso, el valor absoluto de las diferencias se ordenan en rangos (no simplemente positivo y negativo), habiendo más información sobre las diferencias que en el caso anterior.

Si las muestras son homogéneas, la hipótesis nula, la suma de los rangos de las diferencias positivas tiene que ser similar a la suma de los rangos con valor negativo (aleatoriedad de las diferencias).

Una descripción detallada de esta prueba se encuentra en Siegel & Castellan (1988), los pasos principales son:

1. Ordenar los valores de las muestras de forma ascendente y ordenar en rangos las diferencias en valor absoluto entre ambas muestras.

2. Calcular el estadístico de contraste T y T'.

3. Si T o T' es menor o igual que las cantidades límite que aparecen en la denominada tabla de Wilcoxon, se rechaza la hipótesis nula de que las variables son homogéneas.

En caso de que el tamaño de muestra sea alto (mayor de 100 datos), se puede hacer una aproximación a la Normal del estadístico *T*, pudiéndose calcular el valor *Z* de la distribución Normal que nos devuelve la probabilidad del contraste.

FUNCIONES

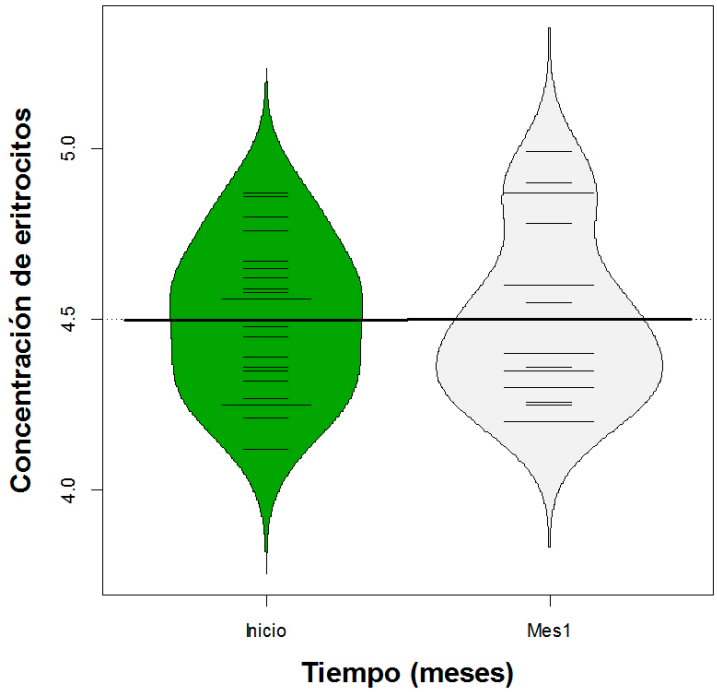
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). La prueba de Wilcoxon se realiza con la función wilcox.test del paquete stats.

EJEMPLO

Datos de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos. El objetivo es determinar si la concentración es diferente entre el inicio y al cabo de un mes.

La Figura IX.14 muestra que la concentración de eritrocitos es muy parecida al inicio y al cabo de un mes, como así lo demuestra el test de Wilcoxon (p = 0,57).

Figura IX.14. Concentración de eritrocitos (en millones por milímetro cúbico) en hombres y mujeres, al inicio y al cabo de un mes de tratamiento.




```
        Wilcoxon signed rank test with continuity correction

data:  datos2$var1 and datos2$var2
V = 108.5, p-value = 0.5697
alternative hypothesis: true location shift is not equal to 0
```

Value

Se obtiene un archivo TXT con los resultados de la prueba de Wilcoxon para dos muestras dependientes y se puede representar un boxplot o un beanplot.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZIX9)

#Variación de la concentración de eritrocitos en el tiempo
IX11(data=ZIX9, varTiempo=c("Inicio", "Mes1"),grafico = "Beanplot" , XLAB="Tiempo (meses)",
YLAB="Concentración de eritrocitos")


## End(Not run)
```

IX12 *CONTRASTES DE HOMOGENEIDAD NO PARAMETRICOS PARA K MUESTRAS INDEPENDIENTES*

Description

Se aplica el ANOVA de Friedman para k-muestras relacionadas. Además se pueden representar los datos con un boxplot o un beanplot.

Usage

```
IX12(data, varTiempo, varBloque, grafico="Boxplot", PAR=NULL, ResetPAR=TRUE,
YLAB="Variable problema", XLAB="Medidas en el tiempo", LabelCat=NULL,
COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varTiempo	Columnas en las que se ha medido en el tiempo la variable dependiente.
varBloque	Variable que identifica cada uno de los casos.
grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.

ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.2. PRUEBAS NO PARAMÉTRICAS

IX.2.4. Contrastes para *K*-muestras dependientes

La prueba más utilizada es el ANOVA de Friedman. Esta prueba contrasta la hipótesis de que las medidas proceden de poblaciones iguales, cuando los datos de cada una de las muestras han sido transformados en rangos de forma ascendente.

Es la prueba análoga al ANOVA para muestras repetidas por lo que, a pesar de que se trata de un contraste χ^2, recibe el nombre de ANOVA.

Los pasos de esta prueba están descritos por Siegel & Castellan (1988). De forma resumida son:

1. Asignar rangos dentro de cada caso de la muestra.

2. Sumar los rangos de cada variable para toda la muestra.

3. Calcular el estadístico χ^2 de contraste y compararlo con el valor tabulado χ^2 crítico con a - 1 grados de libertad, siendo a el número de valores medidos para cada muestra.

FUNCIONES

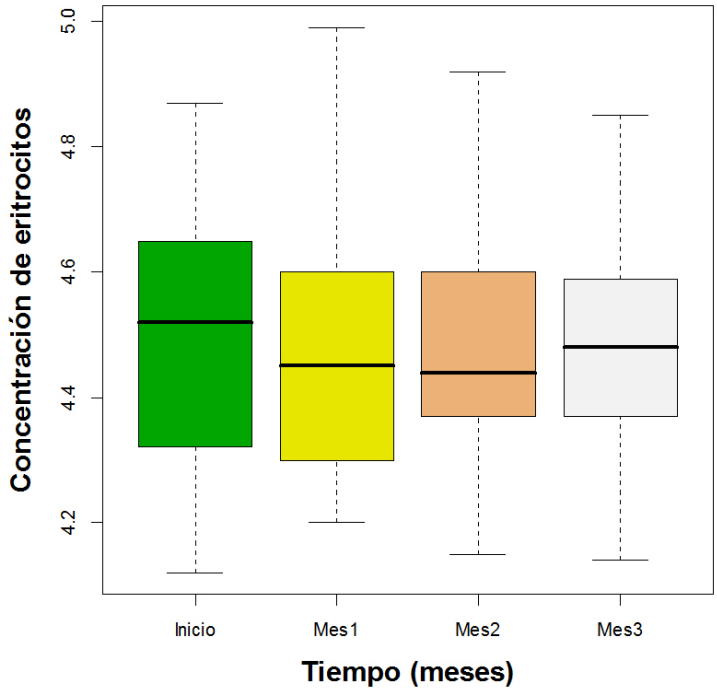
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El ANOVA de Friedman se realiza con la función friedman.test del paquete stats.

EJEMPLO

Datos de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos. El objetivo es determinar si la concentración de glóbulos rojos es diferente considerando todos los tiempos analizados, y considerando el conjunto de hombres y mujeres.

La Figura IX.15 muestra que la concentración de eritrocitos es muy parecida al inicio y en los siguientes meses y, efectivamente los resultados del contraste de Friedman muestran que, en el conjunto de hombres y mujeres, la concentración de glóbulos rojos no cambia en el tiempo (p = 0,967).

Figura IX.15. Concentración de eritrocitos (en millones por milímetro cúbico) en hombres y mujeres, al inicio y en los tres meses siguientes.




```
        Friedman rank sum test

data:  datos1$valores, datos1$variablestot and datos1$bloquetot
Friedman chi-squared = 0.2615, df = 3, p-value = 0.9671
```

Value

Se obtiene un archivo TXT con los resultados del ANOVA de Friedman y se puede representar un boxplot o un beanplot.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZIX10)

#Variación de la concentración de eritrocitos en el tiempo
IX12(data=ZIX10, varTiempo=c("Inicio", "Mes1", "Mes2", "Mes3"),
varBloque="Individuo", XLAB="Tiempo (meses)",
YLAB="Concentración de eritrocitos")


## End(Not run)
```

IX13 *ANALISIS DE COVARIANZA (ANCOVA) NO PARAMETRICO*

Description

Se aplica un análisis de covarianza usando un estadístico tipo ANOVA (Wang & Ye, 2010; Wang, 2013). Además se puede graficar la relación entre la variable dependiente y la covariable, diferenciando entre los grupos del factor.

Usage

```
IX13(data, varInteres, Factor, Covariable, T.AOV=NULL, ResetPAR=TRUE,
grafico=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL, MAIN=NULL, CEXPCH=1.3,
COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable dependiente.
Factor	Factor.
Covariable	Covariable.
T.AOV	Se accede a la función T.aov.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
grafico	Si es NULL no se realiza el gráfico.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
MAIN	Título del gráfico.

CEXPCH	Tamaño de los símbolos del gráfico.
COLOR	Permite modificar los colores del gráfico, pero deben ser tantos como grupos diferentes tenga el factor.
PCH	Vector con los símbolos del gráfico, pero deben ser tantos como grupos diferentes tenga el factor. Si es NULL se calculan automáticamente empezando por el símbolo 15.
LEGEND	Permite incorporar una leyenda al gráfico.
AXIS	Permite añadir ejes al gráfico.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.2. PRUEBAS NO PARAMÉTRICAS

IX.2.5. ANCOVA no paramétrico

Existen también métodos no paramétricos para realizar un análisis de la covarianza. El paquete fANCOVA de R (Wang & Ye, 2010; Wang, 2013) contiene algunas funciones que realizan este análisis basado en la idea de comparar dos curvas para comprobar su igualdad o paralelismo, lo que es válido para curvas de regresión o para funciones de densidad, y por lo tanto se puede considerar equivalente a un análisis paramétrico de la covarianza.

FUNCIONES

Se utiliza la función T.aov del paquete fANCOVA (Wang & Ye, 2010; Wang, 2013), la cual usa un estadístico tipo ANOVA para realizar el ANCOVA no paramétrico.

EJEMPLO 1

Se compara la relación entre temperatura mínima y máxima entre los años 1990 y 2000, considerando los registros diarios de tres ciudades de España: Huelva, Palma de Mallorca y Vigo. El objetivo es determinar si hay diferencias en la temperatura máxima entre los dos años, eliminado el posible efecto de la temperatura mínima que se considera como covariable.

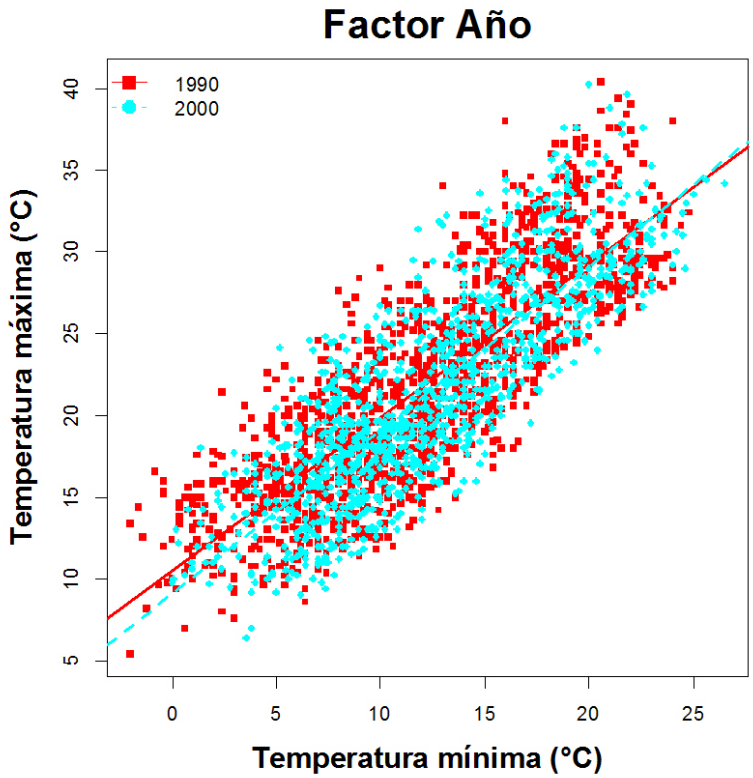
Es importante mencionar que las variables deben ser numéricas, es decir, se ha observado que si se usan variables con texto la función puede dar error. Si es necesario realizar el gráfico con variables de texto, para que por ejemplo quede mejor la leyenda, se puede usar el ANCOVA paramétrico de la función IX4.

Los resultados muestran que no existen diferencias significativas en la relación entre temperatura mínima y máxima entre los años 1990 y 2000 (p = 0,383). Es importante tener en cuenta que si se repite el análisis el valor de probabilidad puede cambiar, es decir, no es siempre da el mismo resultado.

```
        Test the equality of curves based on an ANOVA-type statistic

Comparing 2 nonparametric regression curves
Local polynomial regression with automatic smoothing parameter selection via AICC is used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 curves.
T =  41.35     p-value =  0.4527
```

La Figura IX.16 muestra que existe una clara relación entre temperatura máxima y mínima considerando las tres ciudades.

Figura IX.16. Relación entre la temperatura mínima y máxima, en los años 1990 y 2000, en tres ciudades de España: Huelva, Palma de Mallorca y Vigo.



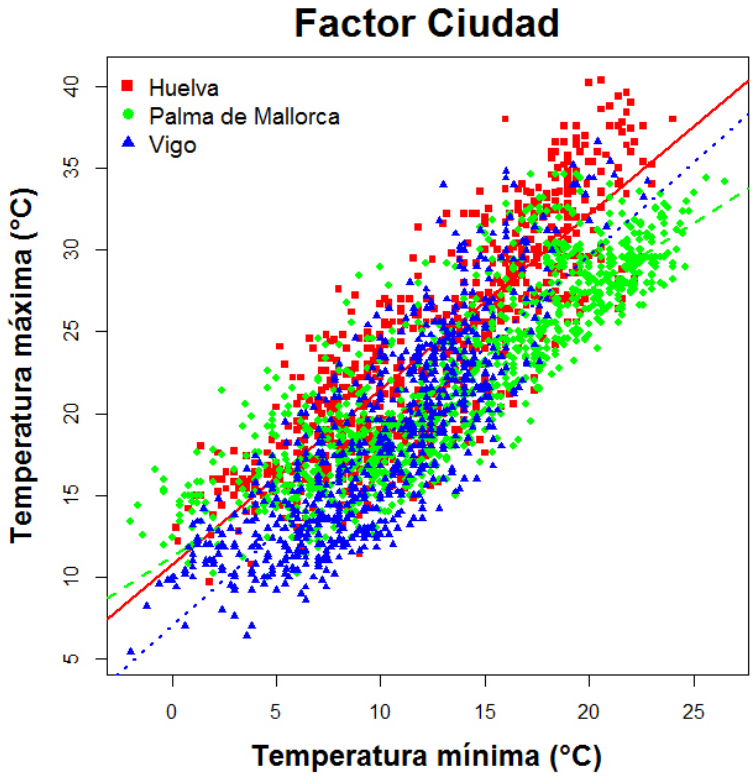
EJEMPLO 2

Se compara la relación entre temperatura mínima y máxima, con datos de registros diarios en los años 1990 y 2000, entre tres ciudades de España: Huelva, Palma de Mallorca y Vigo. El objetivo es determinar si hay diferencias en la temperatura máxima entre las tres ciudades, eliminado el posible efecto de la temperatura mínima que se considera como covariable.

En la Figura IX.17 se observa que existen claras diferencias en la relación entre temperatura mínima y máxima entre las tres ciudades, lo cual es confirmado por el contraste, ya que la p = 0,005.

```
        Test the equality of curves based on an ANOVA-type statistic

Comparing 3 nonparametric regression curves
Local polynomial regression with automatic smoothing parameter selection via AICC is
used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 3 curves.
T =  2.706     p-value =  0.004975
```

Figura IX.17. Relación entre la temperatura mínima y máxima, considerando los años 1990 y 2000, en tres ciudades de España: Huelva, Palma de Mallorca y Vigo.



Value

Se obtiene un archivo TXT con los resultados del ANCOVA no paramétrico y se puede representar un gráfico relacionando la variable dependiente con la covariable, diferenciando además entre los grupos del factor.

References

Wang. X.F. & Ye, D. (2010). On nonparametric comparison of images and regression surfaces. *Journal of Statistical Planning and Inference*, 140: 2875-2884.

Wang, X.F. (2013). Nonparametric Analysis of Covariance. R package version 0.5-1. Available at: http://CRAN.R-project.org/package=fANCOVA.

Examples

```
## Not run:

data(ZIX11)

#EJEMPLO 1
IX13(data=ZIX11, varInteres="T.Max", Factor="Year", Covariable="T.Min",
XLAB="Temperatura mínima (degree C)",
YLAB="Temperatura máxima (degree C)", MAIN="Factor Año", CEXPCH=0.8)

#EJEMPLO 2
IX13(data=ZIX11, varInteres="T.Max", Factor="City", Covariable="T.Min",
XLAB="Temperatura mínima (degree C)",
YLAB="Temperatura máxima (degree C)", MAIN="Factor Ciudad", CEXPCH=0.8, LEGEND=c("x = 'topleft'" ,
"legend=c('Huelva','Palma de Mallorca','Vigo')", "bty = 'n'", "pch=c(15,16,17)",
"col=rainbow(3)", "cex=1.2"))


## End(Not run)
```

IX2 t-*TEST PARA MUESTRAS DEPENDIENTES*

Description

Se aplica el test de la *t* para muestras dependientes y se puede representar los datos con un boxplot o un beanplot.

Usage

```
IX2(data, varInteres, varPoblacion, varComparar1, varComparar2,
trans=NULL, grafico="Boxplot", PAR=NULL, ResetPAR=TRUE, YLAB=NULL,
XLAB=NULL, OrderCat=NULL, LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL,
BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`varInteres`	Variable dependiente.
`varPoblacion`	Define cual es la variable que actúa como factor.
`varComparar1`	Primer grupo a comparar de la variable población.
`varComparar2`	Segundo grupo a comparar de la variable población.
`trans`	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2

3. 1/x
4. LN
5. LOG
6. SQR (raíz cuadrada)
7. x2
8. x3
9. EXP (exponencial)
10. ASN (arcoseno)

grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
OrderCat	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.3. *t*-test

IX.1.3.2. Muestras dependientes

El *t*-test para muestras relacionadas compara las medias de dos variables de un solo grupo. Calcula las diferencias entre los valores de cada variable y contrasta si la diferencia media es significativamente distinta de cero (Sánchez, 1999).

Este test asume que las muestras son dependientes, pareadas o relacionadas, por tanto, un requisito fundamental es tener un número igual de observaciones en ambas variables.

Este *t*-test para muestras dependientes no exige ningún supuesto sobre las varianzas, pero sí requiere que la distribución de las diferencias de los valores de cada par sea Normal. Si no lo es, pero el tamaño de la muestra es grande, el teorema central del límite garantiza que la distribución de

probabilidad de la diferencia de las medias sea aproximadamente Normal, lo que permite utilizar la distribución Normal en lugar de la distribución *t*. Dado que ambas distribuciones son prácticamente iguales cuando el número de grados de libertad es muy grande, esto significa que con muestras grandes podemos aplicar la prueba aunque la distribución de la variable se desvíe de la normalidad.

Las muestras dependientes aparecen a menudo cuando se evalúa una misma variable más de una vez en cada sujeto de la muestra (por ejemplo en intervalos de tiempo diferentes). El *t*-test no se centra en la variabilidad que puede darse entre los individuos, sino en las diferencias que se observan en un mismo sujeto entre un momento y otro. También se aplica este test en estudios de casos y controles donde cada caso se aparea individualmente con un control.

FUNCIONES

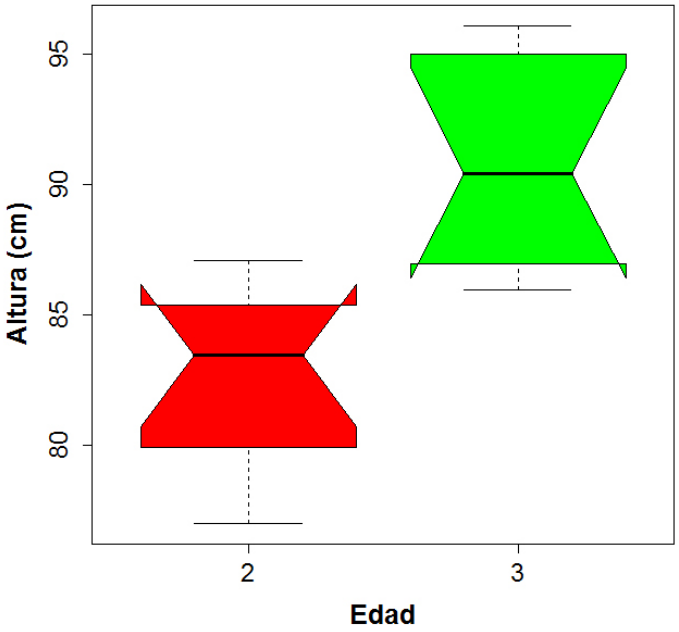
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). La función shapiro.test del paquete base stats para realizar el test de Shapiro-Wilk. El test de Levene se realiza con la función levene.test del paquete lawstat (Gastwirth et al., 2013). La asimetría y la curtosis se realizan con las funciones skewness y kurtosis, respectivamente, del paquete e1071 (Meyer et al., 2014).

EJEMPLO

Datos de peso y altura de niños y niñas entre los 2 y 3 años en Italia y España. El objetivo es determinar si existen diferencias significativas en la altura de las niñas entre los 2 y 3 años en Italia. Los niños y las niñas que se van midiendo y tomando su peso conforme van cumpliendo años son siempre los mismos.

La Figura IX.3 muestra claras disferencias en la altura de las niñas de 2 y 3 años italianas.

Figura IX.3. Boxplot comparando la altura de niñas italianas entre los dos y tres años.



El test de Shapiro-Wilk muestra que las dos poblaciones de niñas de 2 (p = 0,415) y 3 años (p = 0,29) tienen distribución Normal. Además se cumple el requisito de homogeneidad de varianzas, tanto considerando la media (p = 0,831) como la mediana (p = 0,817). Por lo tanto, no es necesario realizar de nuevo el análisis transformando la variable dependiente. En el caso de que no se cumpliera alguno de los dos requisitos, también sale información sobre la asimetría y la curtosis de cada una de las poblaciones, lo cual ayudaría a elegir mejor la transformación a realizar.

```
datos3[, "poblaciones"]: 2

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9266, p-value = 0.4154

------------------------------------------------------------
datos3[, "poblaciones"]: 3

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9113, p-value = 0.2901


[[2]]

     classical Levene's test based on the absolute deviations from the mean
     ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.0468, p-value = 0.8312


[[3]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$valores
Test Statistic = 0.0551, p-value = 0.8171


[[4]]
                                        2                     3
             "Sesgo" "-0.280232427604095" "0.0737002993850063"

[[5]]
                                      2                    3
         "Curtosis" "-1.26371666194686" "-1.69531691801242"
```

Por último, el test de la *t* muestra que existen diferencias significativas en la altura de las niñas italianas entre los 2 y 3 años (t =-25,1, df = 9, p < 0,001), con una diferencia de 8,2 cm.

```
     Paired t-test

data:  datos3$valores by datos3$poblaciones
t = -25.1073, df = 9, p-value = 1.212e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.938817 -7.461183
sample estimates:
mean of the differences
                   -8.2
```

Value

Se obtiene un archivo TXT con los resultados del *t*-test y se puede representar un boxplot o un beanplot.

References

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Kampstra, P (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Sánchez, J.J. (1999) *Manual de análisis estadístico de los datos*. Alianza Editorial. Madrid.

Examples

```
## Not run:

data(ZIX1)

#Comparación de la altura de niñas italianas entre 2 y 3 años.

data1<-subset(ZIX1, Country=="Italy"  &  Gender=="Girl")

IX2(data=data1, varInteres="Height", varPoblacion="Age", varComparar1=2, varComparar2=3,
BOXPLOT= c("xlab='Edad'", "ylab='Altura (cm)'", "cex.lab=1.6","col=c('red','green')",
"notch=TRUE", "cex.axis=1.4"), PAR=c("omi=c(0,0.2,0,0)", "font.lab=2"))


## End(Not run)
```

IX3 *ANALISIS DE VARIANZA (ANOVA)*

Description

Se aplica un análisis de varianza y se puede representar los datos con un boxplot o un beanplot.

Usage

```
IX3(data, varInteres,  Factor1, Factor2=NULL, Factor3=NULL, Factor4=NULL,
SS=3, trans=NULL, ResetPAR=TRUE, mfrow=NULL, grafico1="Boxplot",
PAR1=NULL, OrderCat1=NULL, LabelCat1=NULL, COLOR1=NULL, BOXPLOT1=NULL,
BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL, MTEXT1= NULL, TEXT1=NULL,
grafico2="Boxplot", PAR2=NULL, OrderCat2=NULL, LabelCat2=NULL, COLOR2=NULL,
BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL,
TEXT2=NULL, grafico3="Boxplot", PAR3=NULL, OrderCat3=NULL, LabelCat3=NULL,
```

```
COLOR3=NULL,BOXPLOT3=NULL, BEANPLOT3=NULL, LEGEND3=NULL, AXIS3=NULL,
MTEXT3= NULL, TEXT3=NULL, grafico4="Boxplot", PAR4=NULL, OrderCat4=NULL,
LabelCat4=NULL, COLOR4=NULL,BOXPLOT4=NULL, BEANPLOT4=NULL, LEGEND4=NULL,
AXIS4=NULL, MTEXT4= NULL, TEXT4=NULL, INTPLOT1.2=TRUE, INTPLOT1.3=TRUE,
INTPLOT1.4=TRUE, INTPLOT2.3=TRUE, INTPLOT2.4=TRUE, INTPLOT3.4=TRUE,
INTERPLOT1.2=NULL, INTERPLOT1.3=NULL, INTERPLOT1.4=NULL,
INTERPLOT2.3=NULL,
INTERPLOT2.4=NULL, INTERPLOT3.4=NULL,file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`varInteres`	Variable dependiente.
`Factor1`	Primer factor.
`Factor2`	Segundo factor.
`Factor3`	Tercer factor.
`Factor4`	Cuarto factor.
`SS`	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. El tipo III es adecuado para la mayoría de las aplicaciones, por lo que será el que utilizaremos por defecto. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.
`trans`	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3 9. EXP (exponencial) 10. ASN (arcoseno)
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`mfrow`	Si es NULL y hay varios gráficos, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
`grafico1`	Si es NULL no se realiza el gráfico del factor 1 y las otras opciones son "Boxplot" o "Beanplot".

PAR1	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 1.
OrderCat1	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 1.
LabelCat1	Permite especificar un vector con los nombres de las categorías del gráfico 1.
COLOR1	Vector con el color de las categorías del gráfico 1.
BOXPLOT1	Permite especificar las características del boxplot del factor 1.
BEANPLOT1	Permite especificar las características del beanplot del factor 1.
LEGEND1	Permite incorporar una leyenda al gráfico del factor 1.
AXIS1	Permite añadir ejes al gráfico del factor 1.
MTEXT1	Permite añadir textos en los márgenes del gráfico del factor 1.
TEXT1	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 1.
grafico2	Si es NULL no se realiza el gráfico del factor 2 y las otras opciones son "Boxplot" o "Beanplot".
PAR2	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 2.
OrderCat2	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 2.
LabelCat2	Permite especificar un vector con los nombres de las categorías del gráfico 2.
COLOR2	Vector con el color de las categorías del gráfico 2.
BOXPLOT2	Permite especificar las características del boxplot del factor 2.
BEANPLOT2	Permite especificar las características del beanplot del factor 2.
LEGEND2	Permite incorporar una leyenda al gráfico del factor 2.
AXIS2	Permite añadir ejes al gráfico del factor 2.
MTEXT2	Permite añadir textos en los márgenes del gráfico del factor 2.
TEXT2	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 2.
grafico3	Si es NULL no se realiza el gráfico del factor 3 y las otras opciones son "Boxplot" o "Beanplot".
PAR3	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 3.
OrderCat3	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 3.
LabelCat3	Permite especificar un vector con los nombres de las categorías del gráfico 3.
COLOR3	Vector con el color de las categorías del gráfico 3.
BOXPLOT3	Permite especificar las características del boxplot del factor 3.
BEANPLOT3	Permite especificar las características del beanplot del factor 3.
LEGEND3	Permite incorporar una leyenda al gráfico del factor 3.
AXIS3	Permite añadir ejes al gráfico del factor 3.

MTEXT3	Permite añadir textos en los márgenes del gráfico del factor 3.
TEXT3	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 3.
grafico4	Si es NULL no se realiza el gráfico del factor 4 y las otras opciones son "Boxplot" o "Beanplot".
PAR4	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 4.
OrderCat4	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 4.
LabelCat4	Permite especificar un vector con los nombres de las categorías del gráfico 4.
COLOR4	Vector con el color de las categorías del gráfico 4.
BOXPLOT4	Permite especificar las características del boxplot del factor 4.
BEANPLOT4	Permite especificar las características del beanplot del factor 4.
LEGEND4	Permite incorporar una leyenda al gráfico del factor 4.
AXIS4	Permite añadir ejes al gráfico del factor 4.
MTEXT4	Permite añadir textos en los márgenes del gráfico del factor 4.
TEXT4	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 4.
INTPLOT1.2	Si es TRUE se muestra el gráfico de interacción del factor 1 con el factor 2.
INTPLOT1.3	Si es TRUE se muestra el gráfico de interacción del factor 1 con el factor 3.
INTPLOT1.4	Si es TRUE se muestra el gráfico de interacción del factor 1 con el factor 4.
INTPLOT2.3	Si es TRUE se muestra el gráfico de interacción del factor 2 con el factor 3.
INTPLOT2.4	Si es TRUE se muestra el gráfico de interacción del factor 2 con el factor 4.
INTPLOT3.4	Si es TRUE se muestra el gráfico de interacción del factor 3 con el factor 4.
INTERPLOT1.2	Permite especificar las características del gráfico de interacción del factor 1 con el factor 2.
INTERPLOT1.3	Permite especificar las características del gráfico de interacción del factor 1 con el factor 3.
INTERPLOT1.4	Permite especificar las características del gráfico de interacción del factor 1 con el factor 4.
INTERPLOT2.3	Permite especificar las características del gráfico de interacción del factor 2 con el factor 3.
INTERPLOT2.4	Permite especificar las características del gráfico de interacción del factor 2 con el factor 4.
INTERPLOT3.4	Permite especificar las características del gráfico de interacción del factor 3 con el factor 4.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.4. Análisis de varianza (ANOVA)

En general, al realizar una investigación nos preguntamos si nuestras muestras, que consideramos independientes, pertenecen a una misma población. Los valores de las muestras comúnmente difieren, el problema sería determinar si a pesar de estas diferencias las poblaciones son iguales, y las variaciones son debidas al azar como consecuencia de la aleatoriedad de nuestro muestreo.

La prueba paramétrica más usada para verificar si dos o más medias muestrales proceden de la misma población es el análisis de varianza (ANOVA). El ANOVA al ser un método general, puede ser extendido a más de dos muestras y se puede demostrar que coincide con la *t* de Student si sólo se tienen dos muestras (Salvarrey, 2000; Azzimonti, 2003).

Para la aplicación de un ANOVA es necesario cumplir los siguientes supuestos:

· Las muestras son aleatorias e independientes.

· No existen valores atípicos en la variable dependiente, los cuales se pueden detectar con la función XV7 de StatR.

· La distribución de la población de donde fueron extraídas es Normal.

· Las varianzas son iguales en cada grupo o nivel de los factores considerados.

El ANOVA calcula la variación de las medias y estima la variación "natural" de la población, para luego hacer una comparación entre ambas. La variación natural se mide por medio de la "intravarianza" o "variación del error". Si las muestras son de una misma población, la varianza de las medias es la enésima parte de la varianza de la población. Si se descarta esa situación es porque, aparte del azar, algo más diferencia a las muestras. Si las diferencias son sólo debidas al azar, las dos varianzas corregidas por sus grados de libertad son del mismo orden y su cociente vale más o menos 1. El cociente de varianzas corregidas tiene distribución F, por lo que se busca en las tablas de la distribución F si las relaciones obtenidas son aceptables como cercanas a 1 o no (Salvarrey, 2000).

El análisis de varianza se puede realizar con una variable dependiente (ANOVA univariante) o con varias variables dependientes (ANOVA multivariante o MANOVA). También es posible realizar el análisis de varianza de un factor (ANOVA unifactorial) o con varios factores (ANOVA multifactorial).

El llamado modelo lineal general (MGL) univariante realiza un análisis de regresión y uno de varianza para una variable dependiente mediante uno o más factores o variables independientes (los factores dividen la población en grupos). Con este procedimiento podemos contrastar hipótesis acerca de los efectos de una o más variables (factores o tratamientos) sobre la media de una única variable dependiente. Este análisis exige que la variable dependiente sea cuantitativa y los factores categóricos, es decir, que pueden tener valores numéricos o de cadena (letras).

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors.

La función shapiro.test se usa para realizar el test de Shapiro-Wilk, el modelo lineal se realiza con la función lm y el gráfico de interacción con la función interaction.plot, las tres del paquete base stats.

El ANOVA se realiza con la función Anova del paquete car (Fox et al., 2014).

Para las pruebas posthoc se utiliza la función glhtdel paquete multcomp (Hothorn et al., 2014).

Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013).

El test de Levene se realiza con la función levene.test del paquete lawstat (Gastwirth et al., 2013).

La asimetría y la curtosis se realizan con las funciones skewness y kurtosis, respectivamente, del paquete e1071 (Meyer et al., 2014).

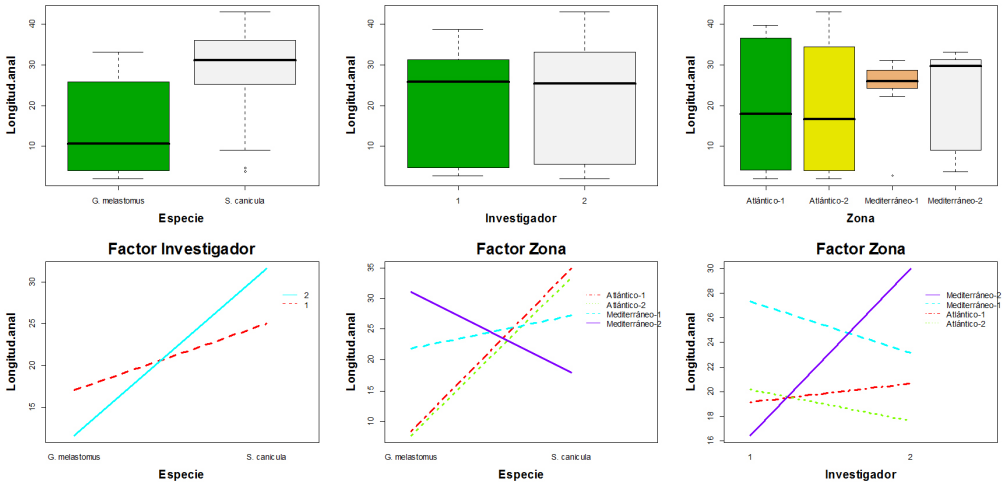
EJEMPLO

Datos biométricos de dos especies de tiburones (*Scyliorhinus canicula* y *Galeus melastomus*), los cuales fueron tomados por dos investigadores en cuatro zonas (dos en el Mediterráneo y dos en el Atlántico). El objetivo es determinar si hay diferencias significativas en la longitud de la aleta anal entre las especies, zonas e investigadores.

En el script se especifica el argumento *mfrow=c(2,3)* para realizar un panel en el que se pongan los 6 gráficos en 2 filas y 3 columnas.

La Figura IX.4 muestra el panel con la mediana y distribución de los valores de la longitud de la aleta anal entre especies, investigadores y zonas, además de los gráficos de interacción entre todos los factores.

Figura IX.4. Longitud de la aleta anal entre especies, investigadores y zonas, además de los gráficos de interacción entre factores.



Los residuos se ajustan a una distribución Normal: test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,345) y el de Shapiro-Wilk (p = 0,108), y existe homogeneidad de varianzas para el factor 1 (p = 0,251), para el factor 2 (p = 0,777) y para el factor 3 (p = 0,084). Por lo tanto, no es necesario transformar los datos.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0865, p-value = 0.3453


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9664, p-value = 0.1078


[[6]]
[[6]][[1]]
[1] "Factor1"

[[6]][[2]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.344, p-value = 0.2513


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 0.0806, p-value = 0.7775


[[6]][[5]]
[1] "Factor3"

[[6]][[6]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 2.3338, p-value = 0.08418
```

A pesar de considerar el efecto «Investigador» y «Zona», hay claras diferencias en la longitud de la aleta anal entre especies (ANOVA, $F_{1,42}$ = 35,6, p < 0,001). Sin embargo, no existen diferencias significativas entre investigadores (ANOVA, $F_{1,42}$= 0,22, p = 0,638), es decir, no existen diferencias de criterio a la hora de tomar la medida. Tampoco existen diferencias entre zonas (ANOVA, $F_{3,42}$= 0,97, p = 0,417). En la tabla de ANOVA se muestra también el efecto conjunto de los factores, considerado como un factor adicional que recoge la posibilidad de que -además del efecto de la especie, investigador y la zona, que se suman- aparezca otro más, que también se suma a los anteriores, que es la interacción entre ellos. Se observa que algunas interacciones son significativas, como «Especie*Investigador» (ANOVA, $F_{1,42}$= 7,73, p = 0,008). Esto se observa en el gráfico de interacción entre los factores especie e investigador (Figura IX.4, figura de la izquierda de la segunda fila), ya que las líneas no son paralelas, sino que se cruzan, lo cual nos está indicando que el investigador 1

obtiene un tamaño mayor de aleta anal para la especie *G. melastomus* que el 2 mientras que, por el contrario, el investigador 2 obtiene un tamaño mayor de aleta anal para la especie *S. canicula* que el 1, es decir, los investigadores miden de forma diferente las especies. Sin embargo, no debe tenerse en cuenta una interacción cuando alguno de los factores que la forman no es significativo, en este caso, el factor investigador.

```
Anova Table (Type III tests)

Response: datos2$valores
                              Sum Sq Df  F value    Pr(>F)
(Intercept)                  27329.4  1 579.0386 < 2.2e-16 ***
datos2$F1                     1682.3  1  35.6438 4.383e-07 ***
datos2$F2                       10.6  1   0.2240  0.638466
datos2$F3                      137.0  3   0.9678  0.416922
datos2$F1:datos2$F2            364.9  1   7.7320  0.008086 **
datos2$F1:datos2$F3           3229.7  3  22.8098 6.407e-09 ***
datos2$F2:datos2$F3            432.1  3   3.0516  0.038815 *
datos2$F1:datos2$F2:datos2$F3  290.1  3   2.0487  0.121602
Residuals                     1982.3 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como en el caso del factor «Zona» hay 4 niveles, se podrían hacer pruebas *post hoc* para averiguar cuales son los niveles distintos. Sin embargo, no tiene sentido hacer comparaciones *post hoc* con un factor que no es significativo, ya que si una sola comparación fuera significativa también lo sería el factor. De todas formas R muestra los resultados *post hoc* del test Tukey, incluso para aquellos factores que solo tienen 2 niveles. Al fijarnos en el Factor 3, que es la «Zona», como era de esperar, se observa que no existen diferencias significativas al comparar cada una de las zonas: zona 2 con la 1 (p = 0,986), zona 3 con la 1 (p = 0,919), zona 4 con la 1 (p = 0,952), zona 3 con la 2 (p = 0,595), zona 4 con la 2 (p = 0,703), y zona 4 con la 3 (p = 1).

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = dv ~ Factor1 * Factor2 * Factor3)

Linear Hypotheses:
                                                 Estimate Std. Error t value
Factor1: S. canicula - G. melastomus == 0        11.44521    1.91704   5.970
Factor2: 2 - 1 == 0                               0.90729    1.91704   0.473
Factor3: Atlántico-2 - Atlántico-1 == 0          -1.46708    2.49878  -0.587
Factor3: Mediterráneo-1 - Atlántico-1 == 0        2.32542    2.59529   0.896
Factor3: Mediterráneo-2 - Atlántico-1 == 0        2.24375    2.84818   0.788
Factor3: Mediterráneo-1 - Atlántico-2 == 0        3.79250    2.56672   1.478
Factor3: Mediterráneo-2 - Atlántico-2 == 0        3.71083    2.82217   1.315
Factor3: Mediterráneo-2 - Mediterráneo-1 == 0    -0.08167    2.90797  -0.028
                                                 Pr(>|t|)
Factor1: S. canicula - G. melastomus == 0          <1e-04 ***
Factor2: 2 - 1 == 0                                 0.995
Factor3: Atlántico-2 - Atlántico-1 == 0             0.986
Factor3: Mediterráneo-1 - Atlántico-1 == 0          0.919
Factor3: Mediterráneo-2 - Atlántico-1 == 0          0.952
Factor3: Mediterráneo-1 - Atlántico-2 == 0          0.595
Factor3: Mediterráneo-2 - Atlántico-2 == 0          0.703
Factor3: Mediterráneo-2 - Mediterráneo-1 == 0       1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Value

Se obtiene un archivo TXT con los resultados del ANOVA, se puede representar un boxplot o un beanplot para cada factor, además de representarse un gráfico de interacción entre todos los factores.

References

Azzimonti, J.C. (2003) *Bioestadística aplicada a Bioquímica y Farmacia*. Universidad Nacional de Misiones. Editorial Universitaria, Argentina.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T., Bretz, F., Westfall, P, Heiberger, R.M. & Schuetzenmeister, A. (2014) Simultaneous Inference in General Parametric Models. R package version 1.3-3. Available at: http://CRAN.R-project.org/package=multcomp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Salvarrey, L. (2000) *Curso de estadística básica*. Universidad de la República del Uruguay.

Examples

```
## Not run:
data(ZIX3)

#Comparación de la longitud de la aleta anal entre dos especies de
#tiburones, medidas por dos investigadores y en varias áreas de muestreo

IX3(data=ZIX3, varInteres="Longitud.anal", Factor1="Especie", Factor2="Investigador",
Factor3="Zona", mfrow=c(2,3))

## End(Not run)
```

IX4 *ANALISIS DE COVARIANZA (ANCOVA)*

Description

Se aplica un análisis de covarianza con gráfica opcional.

Usage

```
IX4(data, varInteres,  Factor1, Factor2=NULL, Factor3=NULL, Covariable1,
Covariable2=NULL, Covariable3=NULL, Covariable4=NULL, SS=3, trans=NULL,
ResetPAR=TRUE, mfrow=NULL,  CEXPCH=1.3, grafico1=TRUE, XLAB=NULL,
YLAB=NULL,
PAR1=NULL, COLOR1=NULL, PCH1=NULL,  MAIN1=NULL, LEGEND1=NULL, AXIS1=NULL,
MTEXT1= NULL, TEXT1=NULL, grafico2=TRUE, PAR2=NULL, COLOR2=NULL, PCH2 =NULL,
MAIN2=NULL, LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, grafico3=TRUE,
PAR3=NULL, COLOR3=NULL,  PCH3=NULL, MAIN3=NULL, LEGEND3=NULL, AXIS3=NULL,
MTEXT3= NULL, TEXT3=NULL, INTPLOT1.2=TRUE, INTPLOT1.3=TRUE,
INTPLOT2.3=TRUE,
INTERPLOT1.2=NULL, INTERPLOT1.3=NULL, INTERPLOT2.3=NULL,
OrderCat1=NULL, LabelCat1=NULL, OrderCat2=NULL, LabelCat2=NULL,  file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable dependiente.
Factor1	Primer factor.
Factor2	Segundo factor.
Factor3	Tercer factor.

Covariable1	Primera covariable.
Covariable2	Segunda covariable.
Covariable3	Tercera covariable.
Covariable4	Cuarta covariable.
SS	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. El tipo III es adecuado para la mayoría de las aplicaciones, por lo que será el que utilizaremos por defecto. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.
trans	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3 9. EXP (exponencial) 10. ASN (arcoseno)
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
mfrow	Si es NULL y hay varios gráficos, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
CEXPCH	Tamaño de los símbolos de las gráficas.
grafico1	Si es NULL no se realiza el gráfico del factor 1 con la covariable 1.
XLAB	Leyenda del eje X.
YLAB	Leyenda del eje Y.
PAR1	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 1.
COLOR1	Permite modificar los colores del gráfico del factor 1, pero deben ser tantos como grupos diferentes tenga el factor 1.
PCH1	Vector con los símbolos del gráfico 1. Si es NULL se calculan automáticamente empezando por el símbolo 15.
MAIN1	Título del gráfico del factor 1.
LEGEND1	Permite modificar una leyenda al gráfico del factor 1.
AXIS1	Permite añadir ejes al gráfico del factor 1.

MTEXT1	Permite añadir textos en los márgenes del gráfico del factor 1.
TEXT1	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 1.
grafico2	Si es NULL no se realiza el gráfico del factor 2 con la covariable 1.
PAR2	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 2.
COLOR2	Permite modificar los colores del gráfico del factor 2, pero deben ser tantos como grupos diferentes tenga el factor 2.
PCH2	Vector con los símbolos del gráfico 2. Si es NULL se calculan automáticamente empezando por el símbolo 15.
MAIN2	Título del gráfico del factor 2.
LEGEND2	Permite modificar una leyenda al gráfico del factor 2.
AXIS2	Permite añadir ejes al gráfico del factor 2.
MTEXT2	Permite añadir textos en los márgenes del gráfico del factor 2.
TEXT2	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 2.
grafico3	Si es NULL no se realiza el gráfico del factor 3 con la covariable 1.
PAR3	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 3.
COLOR3	Permite modificar los colores del gráfico del factor 3, pero deben ser tantos como grupos diferentes tenga el factor 3.
PCH3	Vector con los símbolos del gráfico 3. Si es NULL se calculan automáticamente empezando por el símbolo 15.
MAIN3	Título del gráfico del factor 3.
LEGEND3	Permite modificar una leyenda al gráfico del factor 3.
AXIS3	Permite añadir ejes al gráfico del factor 3.
MTEXT3	Permite añadir textos en los márgenes del gráfico del factor 3.
TEXT3	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 3.
INTPLOT1.2	Si es TRUE se muestra el gráfico de interacción del factor 1 con el factor 2.
INTPLOT1.3	Si es TRUE se muestra el gráfico de interacción del factor 1 con el factor 3.
INTPLOT2.3	Si es TRUE se muestra el gráfico de interacción del factor 2 con el factor 3.
INTERPLOT1.2	Permite especificar las características del gráfico de interacción del factor 1 con el factor 2.
INTERPLOT1.3	Permite especificar las características del gráfico de interacción del factor 1 con el factor 3.
INTERPLOT2.3	Permite especificar las características del gráfico de interacción del factor 2 con el factor 3.
OrderCat1	Permite especificar un vector con el orden en que se muestran las categorías del factor 1 en el gráfico de interacción.

`LabelCat1`	Permite especificar un vector con los nombres de las categorías del factor 1 en el gráfico de interacción.
`OrderCat2`	Permite especificar un vector con el orden en que se muestran las categorías del factor 2 en el gráfico de interacción.
`LabelCat2`	Permite especificar un vector con los nombres de las categorías del factor 2 en el gráfico de interacción.
`file`	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.5. Análisis de covarianza (ANCOVA)

El análisis de la covarianza (ANCOVA) es una técnica estadística cuyo objetivo es eliminar el efecto de posibles variables que puedan estar relacionadas con la variable dependiente. Para controlar estas variables se realiza un análisis de varianza, en el que la variable dependiente es el error en los pronósticos, o residuo, al realizar un análisis de regresión lineal con las covariables como variables independientes y la variable dependiente será igual a la de nuestro ANOVA inicial.

En el ANCOVA la interpretación de sus resultados, al igual que en el ANOVA, se basa en los efectos de los factores sobre nuestras variables dependientes y de las interacciones entre los factores estudiados. Según Steel & Torrie (1985) este análisis se puede usar para:

1. Controlar errores y aumentar la precisión de nuestros análisis.

2. Ajustar las medias de las variables dependientes a las diferencias con los valores de las variables independientes usadas.

3. Ayudar en la interpretación de nuestros datos, especialmente en lo referente a la naturaleza de los efectos de los tratamientos.

4. Particionar una covarianza total o suma de productos cruzados en componentes.

5. Estimar datos que faltan.

Para poder interpretar de forma correcta un ANCOVA es necesario primero realizar el ANOVA, es decir, realizar el análisis sólo teniendo en cuenta la variable dependiente y el o los factores, sin considerar la(s) covariable(s). Esto se debe a que la interpretación del ANCOVA difiere en función de si los resultados obtenidos en el ANCOVA son iguales o diferentes a los obtenidos en el ANOVA.

En el ANCOVA toda variable que no tenga un efecto significativo puede ser eliminada del análisis. Si todas las covariables utilizadas poseen efectos no significativos, las conclusiones del ANCOVA y del ANOVA deben ser muy similares.

Si una o varias covariables tienen efectos significativos pueden ocurrir dos situaciones:

1. Que el resultado del ANOVA y del ANCOVA sea el mismo. Eso significa que el efecto de los factores sobre la variable dependiente no se modifica y, por tanto, el posible efecto de la(s) covariable(s) sobre la variable dependiente no afecta a la relación de esta última con los factores.

2. Si el resultado del ANOVA y del ANCOVA es distinto, puede deberse a dos motivos:

2a. Puede ser debido a que un efecto significativo del ANOVA pasa a no ser significativo y, entonces, el efecto detectado en el ANOVA se debe al efecto de la(s) covariable(s) y no a la(s) variable(s) independientes (los factores).

2b. También puede ser debido a que un efecto no significativo del ANOVA pase a significativo, lo cual indica que el factor, aun no estando relacionado con la variable dependiente globalmente considerada, sí correlaciona con la parte de la variable dependiente no explicada o no atribuible a la(s) covariable(s).

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors. La función shapiro.test se usa para realizar el test de Shapiro-Wilk, el modelo lineal se realiza con la función lm y el gráfico de interacción con la función interaction.plot, las tres del paquete base stats. El ANCOVA se realiza con la función Anova del paquete car (Fox et al., 2014). El test de Levene se realiza con la función levene.test del paquete lawstat (Gastwirth et al., 2013). La asimetría y la curtosis se realizan con las funciones skewness y kurtosis, respectivamente, del paquete e1071 (Meyer et al., 2014).

EJEMPLO

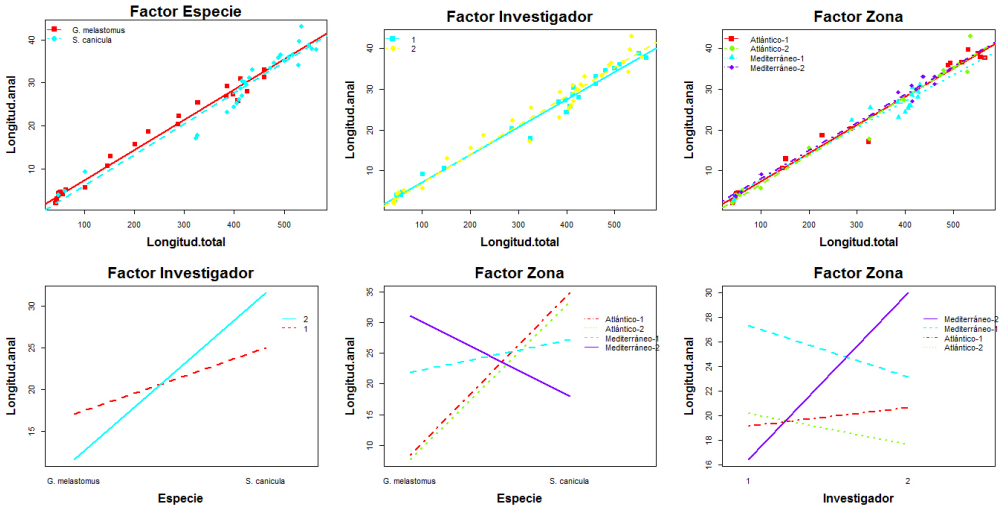
Se compara la longitud de la aleta anal entre dos especies de tiburones, medidas por dos investigadores y en varias áreas de muestreo considerando la longitud total del cuerpo como covariable.

Paso 1.

La Figura IX.5 muestra el panel con la relación entre la longitud total y la longitud de la aleta anal, diferenciando entre las especies, los investigadores y las zonas, además de los gráficos de interacción entre factores.

En primer lugar nos fijamos en la Normalidad de los residuos y la homogeneidad de varianzas. Se observa que, tanto el test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,256) como el de Shapiro-Wilk (p = 0,34) presentan valores de p > 0,05 y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal. El test de Levene considerando la mediana muestra que no se cumple la homogeneidad de varianzas para el factor 1 (p = 0,002), pero si para el factor 2 (p = 0,725) y el factor 3 (p = 0,623). Por lo tanto, es necesario transformar los datos.

Figura IX.5. Relación entre la longitud total y la longitud de la aleta entre especies, investigadores y zonas, junto con los gráficos de interacción entre factores.



```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0921, p-value = 0.2556


[[7]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9771, p-value = 0.3399


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 10.227, p-value = 0.002278


[[8]][[3]]
[1] "Factor2"

[[8]][[4]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 0.1246, p-value = 0.7254
```

Paso 2.

Se repite el análisis pero transformando los datos con el argumento *trans="SQR"*. Ahora se observa que, tanto el test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,576) como el de Shapiro-Wilk (p = 0,361) tienen valores de p > 0,05 y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal. El test de Levene considerando la mediana muestra que se cumple la homogeneidad de varianzas para el factor 1 (p = 0,263), para el factor 2 (p = 0,202) y para el factor 3 (p = 0,159).

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.075, p-value = 0.5761


[[7]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9777, p-value = 0.3613


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.2781, p-value = 0.2631


[[8]][[3]]
[1] "Factor2"

[[8]][[4]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.6696, p-value = 0.2016


[[8]][[5]]
[1] "Factor3"

[[8]][[6]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.7961, p-value = 0.1589
```

La tabla de ANOVA muestra que existe una clara relación entre la longitud total y la longitud de la aleta anal (ANCOVA, $F_{1,41}$ = 576,37, p < 0,001). Al considerar el efecto de la longitud total siguen existiendo claras diferencias en la longitud de la aleta anal entre especies (ANOVA, $F_{1,41}$ = 24,09, p < 0,001), pero no entre investigadores (ANOVA, $F_{1,41}$= 1,44, p = 0,235) ni entre zonas (ANOVA, $F_{3,41}$= 1,06, p = 0,374).

```
Anova Table (Type III tests)

Response: datos2$valores
                             Sum Sq Df  F value    Pr(>F)
(Intercept)                   1.987  1  32.4740 1.160e-06 ***
datos2$C1                    35.265  1 576.3747 < 2.2e-16 ***
datos2$F1                     1.474  1  24.0953 1.504e-05 ***
datos2$F2                     0.089  1   1.4474 0.2358415
datos2$F3                     0.196  3   1.0652 0.3743666
datos2$F1:datos2$F2           0.574  1   9.3866 0.0038534 **
datos2$F1:datos2$F3           1.483  3   8.0818 0.0002405 ***
datos2$F2:datos2$F3           0.301  3   1.6415 0.1946235
datos2$F1:datos2$F2:datos2$F3 1.023  3   5.5707 0.0026676 **
Residuals                     2.509 41
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Value

Se obtiene un archivo TXT con los resultados del ANCOVA, se puede representar un gráfico para cada factor con la covariable 1, además de representarse un gráfico de interacción del factor 1 con cada uno de los otros factores.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Steel, R.G.D. & Torrie, J.H. (1985) *Bioestadística: Principios y procedimientos*. McGraw-Hill, Bogotá.

Examples

```
## Not run:

data(ZIX3)

#Paso 1.
#Sale un gráfico individual para cada factor e interacción entre factores

IX4(data=ZIX3,  varInteres="Longitud.anal", Factor1="Especie",
Factor2="Investigador", Factor3="Zona", Covariable1="Longitud.total")
```

```
#Se combinan los 6 gráficos en uno con mfrow, se especifican distancias
#de los márgenes con PAR(mar=) y se pone color al gráfico del factor 2.

IX4(data=ZIX3,  varInteres="Longitud.anal", Factor1="Especie",
Factor2="Investigador", Factor3="Zona", Covariable1="Longitud.total",
mfrow=c(2,3), PAR1=c("mar=c(5,5,2,1)", "font.lab=2", "cex.lab=1.5"),
COLOR2= c("#00FFFFFF", "#FFFF00FF"))

#Paso 2.

IX4(data = ZIX3 , varInteres = "Longitud.anal" , Factor1 = "Especie" ,
Factor2 = "Investigador" , Factor3 = "Zona" , Covariable1 = "Longitud.total",
trans = "SQR" , mfrow = c(3,1) , PAR1 = c("mar=c(5,5,2,1)", "font.lab=2",
"cex.lab=1.5") , COLOR2 = c("#00FFFFFF", "#FFFF00FF"))


## End(Not run)
```

IX5 *ANOVA PARA FACTORES CON MEDIDAS REPETIDAS*

Description

Se aplica un ANOVA para muestras dependientes y se puede representar los datos con un boxplot o un beanplot.

Usage

```
IX5(data, varTiempo, grafico="Boxplot", SS=2, PAR=NULL, ResetPAR=TRUE,
YLAB="Variable problema", XLAB="Medidas en el tiempo",
LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL,
BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

data	Archivo de datos.
varTiempo	Columnas en las que se ha medido en el tiempo la variable dependiente.
grafico	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
SS	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. Como en un modelo de medidas repetidas todas las casillas del modelo tienen igual número de datos (modelo equilibrado), usaremos por defecto el tipo II. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.

PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
LabelCat	Permite especificar un vector con los nombres de las categorías del gráfico.
COLOR	Vector con el color de las categorías del gráfico.
BOXPLOT	Permite especificar las características del boxplot.
BEANPLOT	Permite especificar las características del beanplot.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.6. ANOVA para factores con medidas repetidas

En la función IX2 se mostró que había una modalidad del *t*-test que se usaba en el caso de querer realizar un contraste de homogeneidad de datos relacionados o pareados. Sin embargo, se puede dar el caso de que el factor tenga más de dos niveles o que se tenga más de un factor. Por ejemplo, seleccionamos varias personas y les pedimos que den una puntuación de 0 a 10 en la calidad de cinco perfumes diferentes. En estos casos el *t*-test solo nos permite comparar dos perfumes entre sí; para comparar los cinco conjuntamente se utiliza el análisis de varianza para medidas repetidas.

El análisis con medidas repetidas está pensado para determinar con mayor precisión el error residual, que será menor si conseguimos controlar o eliminar la variabilidad entre sujetos y dejar únicamente la variabilidad intra sujetos. Cuanto más pequeño es el error residual, más precisión tendremos en la determinación de efectos significativos, ya que cada efecto se compara con la variabilidad residual para saber si es significativo.

Este tipo de análisis de varianza puede ser unifactorial o multifactorial, es decir, trabajar con uno o más factores con medidas repetidas en todos los factores o sólo en alguno de ellos. También permite incluir covariables, es decir, realizar un ANCOVA (un análisis de covarianza).

FUNCIONES

Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El modelo lineal se realiza con la función lm del paquete stats. El ANOVA se realiza con la función Anova del paquete car (Fox et al., 2014).

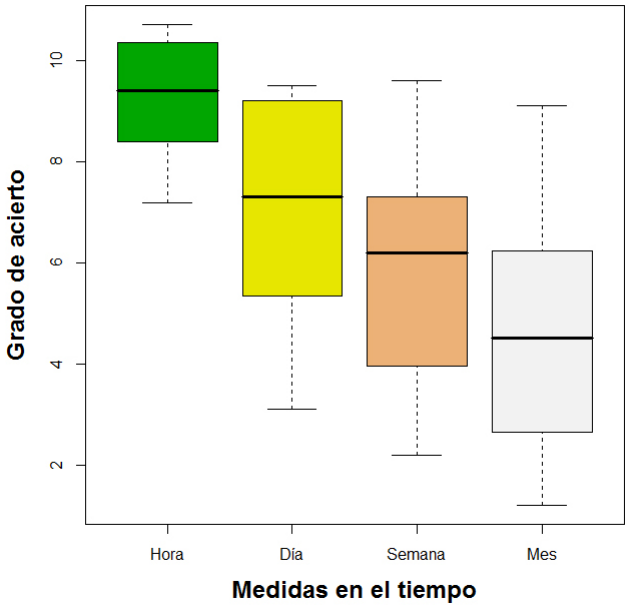
EJEMPLO

En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vinos de diferentes regiones. El objetivo era determinar si la capacidad de acertar la procedencia del vino variaba en el tiempo (después de una hora,

un día, una semana y un mes). Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12.

La Figura IX.6 muestra claramente como el grado de acierto de los catadores va disminuyendo con el tiempo.

Figura IX.6. Grado de acierto de los catadores en el tiempo.



Los resultados indican que el tiempo es un factor significativo ($p < 0{,}001$). Las diferencias observadas en el grado de acierto a lo largo del tiempo no son debidas al azar, sino que la capacidad aprendida de identificar correctamente los vinos se pierde con el tiempo.

```
[1] "ANOVA para medidas repetidas"

[[2]]

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                SS num Df Error SS den Df       F    Pr(>F)
(Intercept) 3374.2      1  238.452     18 254.710 4.547e-12 ***
Factor       245.2      3   56.862     54  77.633 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Un requisito del estadístico *F* en el ANOVA con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad (varianzas iguales y covarianzas nulas), o que la dispersión conjunta

es esférica. La tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como el valor *p* es mayor que 0,05 (p = 0,073) se acepta la hipótesis de esfericidad.

```
Mauchly Tests for Sphericity

       Test statistic  p-value
Factor        0.54716 0.073354


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

        GG eps Pr(>F[GG])
Factor 0.70026  2.282e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          HF eps    Pr(>F[HF])
Factor 0.7948988 5.135716e-16
```

Si no se cumpliera el requisito de esfericidad (si el valor *p* fuese menor que 0,05), se debería utilizar alguna de las dos modificaciones que también se muestran: Greenhouse-Geisser, y Huynh-Feldt, para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor *p* del contraste ANOVA. En este caso ambos valores *p* corregidos $2,282 * 10^{-14}$ y $5,136 * 10^{-16}$ son prácticamente cero, por lo que muestran también una relación significativa entre el tiempo y el acierto.

Una vez probada la relación significativa, debemos averiguar cuáles de los niveles son distintos (al menos dos lo son). Para ello podemos utilizar una prueba *t* con los dos niveles que queremos comparar. La función realiza de forma automática todas las comparaciones entre cada uno de los tiempos. Por ejemplo, se muestra los resultados de la comparación del grado de acierto al cabo de una hora y al cabo de un día. Como el valor p < 0,001, podemos afirmar que el grado de acierto es significativamente distinto entre hora y día.

```
[1] "Comparación t-test entre tiempos"


[[3]][[1]][[1]][[1]][[1]][[1]][[2]]
[1] "t-test entre  Hour  y  Day"

[[3]][[1]][[1]][[1]][[1]][[1]][[3]]

     Paired t-test

data:  datos1[, i] and datos1[, x]
t = 7.2223, df = 18, p-value = 1.019e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.716780 3.125326
sample estimates:
mean of the differences
               2.421053
```

Al hacer múltiples comparaciones como ésta, debemos recordar que es necesario utilizar algún tipo de protección, especialmente cuando se hacen muchas comparaciones, para evitar el aumento de la

probabilidad de error de tipo I. Por ejemplo, podemos dividir el nivel de significación por el número de comparaciones (método Bonferroni).

Value

Se obtiene un archivo TXT con los resultados del ANOVA para muestras dependientes y se puede representar un boxplot o un beanplot.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Examples

```
## Not run:

data(ZIX5)

#Variación de la capacidad de acertar la procedencia del
#vino en función del tiempo

IX5(data = ZIX5 , varTiempo = c("Hour","Day","Week","Month") ,
YLAB = "Grado de acierto" , LabelCat = c("Hora","Día","Semana","Mes"))


## End(Not run)
```

IX6 *ANOVA PARA FACTORES CON MEDIDAS REPETIDAS Y NO REPETIDAS*

Description

Se aplica un ANOVA en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos), además de poder representar los datos con un boxplot o un beanplot, y se representa un gráfico de interacción de cada uno de los factores con el factor tiempo.

Usage

```
IX6(data, varTiempo, Factor1, Error, Factor2=NULL, Factor3=NULL, SS=2, ResetPAR=TRUE,
mfrow=NULL, YLAB="Variable problema", XLAB="Medidas en el tiempo",
graficoT="Boxplot", PART=NULL, LabelCatT=NULL, COLORT=NULL, BOXPLOTT=NULL,
BEANPLOTT=NULL, LEGENDT=NULL, AXIST=NULL, MTEXTT= NULL, TEXTT=NULL,
grafico1="Boxplot", PAR1=NULL, OrderCat1=NULL, LabelCat1=NULL, COLOR1=NULL,
BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL, MTEXT1= NULL,
TEXT1=NULL, grafico2="Boxplot",  PAR2=NULL, OrderCat2=NULL, LabelCat2=NULL,
COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL,
AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, grafico3="Boxplot", PAR3=NULL,
OrderCat3=NULL, LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL,
BEANPLOT3=NULL, LEGEND3=NULL, AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL,
INTERPLOT1=NULL, INTERPLOT2=NULL, INTERPLOT3=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varTiempo	Columnas en las que se ha medido en el tiempo la variable dependiente.
Factor1	Primer factor.
Error	Es la variable que codifica los sujetos estudiados.
Factor2	Segundo factor.
Factor3	Tercer factor.
SS	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. Como en un modelo de medidas repetidas todas las casillas del modelo tienen igual número de datos (modelo equilibrado), usaremos por defecto el tipo II. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
mfrow	Si es NULL y hay varios gráficos, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
graficoT	Si es NULL no se realiza el gráfico del factor tiempo y las otras opciones son "Boxplot" o "Beanplot".
PART	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor tiempo.
LabelCatT	Permite especificar un vector con los nombres de las categorías del gráfico del factor tiempo.

COLORT	Vector con el color de las categorías del gráfico del factor tiempo.
BOXPLOTT	Permite especificar las características del boxplot del gráfico del factor tiempo.
BEANPLOTT	Permite especificar las características del beanplot gráfico del factor tiempo.
LEGENDT	Permite incorporar una leyenda al gráfico del factor tiempo.
AXIST	Permite añadir ejes al gráfico del factor tiempo.
MTEXTT	Permite añadir textos en los márgenes del gráfico del factor tiempo.
TEXTT	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor tiempo.
grafico1	Si es NULL no se realiza el gráfico del factor 1 y las otras opciones son "Boxplot" o "Beanplot".
PAR1	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 1.
OrderCat1	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 1.
LabelCat1	Permite especificar un vector con los nombres de las categorías del gráfico 1.
COLOR1	Vector con el color de las categorías del gráfico 1.
BOXPLOT1	Permite especificar las características del boxplot del factor 1.
BEANPLOT1	Permite especificar las características del beanplot del factor 1.
LEGEND1	Permite incorporar una leyenda al gráfico del factor 1.
AXIS1	Permite añadir ejes al gráfico del factor 1.
MTEXT1	Permite añadir textos en los márgenes del gráfico del factor 1.
TEXT1	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 1.
grafico2	Si es NULL no se realiza el gráfico del factor 2 y las otras opciones son "Boxplot" o "Beanplot".
PAR2	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 2.
OrderCat2	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 2.
LabelCat2	Permite especificar un vector con los nombres de las categorías del gráfico 2.
COLOR2	Vector con el color de las categorías del gráfico 2.
BOXPLOT2	Permite especificar las características del boxplot del factor 2.
BEANPLOT2	Permite especificar las características del beanplot del factor 2.
LEGEND2	Permite incorporar una leyenda al gráfico del factor 2.
AXIS2	Permite añadir ejes al gráfico del factor 2.
MTEXT2	Permite añadir textos en los márgenes del gráfico del factor 2.
TEXT2	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 2.
grafico3	Si es NULL no se realiza el gráfico del factor 3 y las otras opciones son "Boxplot" o "Beanplot".

`PAR3`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 3.
`OrderCat3`	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 3.
`LabelCat3`	Permite especificar un vector con los nombres de las categorías del gráfico 3.
`COLOR3`	Vector con el color de las categorías del gráfico 3.
`BOXPLOT3`	Permite especificar las características del boxplot del factor 3.
`BEANPLOT3`	Permite especificar las características del beanplot del factor 3.
`LEGEND3`	Permite incorporar una leyenda al gráfico del factor 3.
`AXIS3`	Permite añadir ejes al gráfico del factor 3.
`MTEXT3`	Permite añadir textos en los márgenes del gráfico del factor 3.
`TEXT3`	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 3.
`INTERPLOT1`	Permite especificar las características del gráfico de interacción del factor 1.
`INTERPLOT2`	Permite especificar las características del gráfico de interacción del factor 2.
`INTERPLOT3`	Permite especificar las características del gráfico de interacción del factor 3.
`file`	ARCHIVO TXT. Nombre del archivo de salida.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.7. ANOVA para factores con medidas repetidas y no repetidas

Véase la sección *details* de la función IX5 o Guisande et al. (2011) para la explicación sobre este tipo de ANOVA.

FUNCIONES

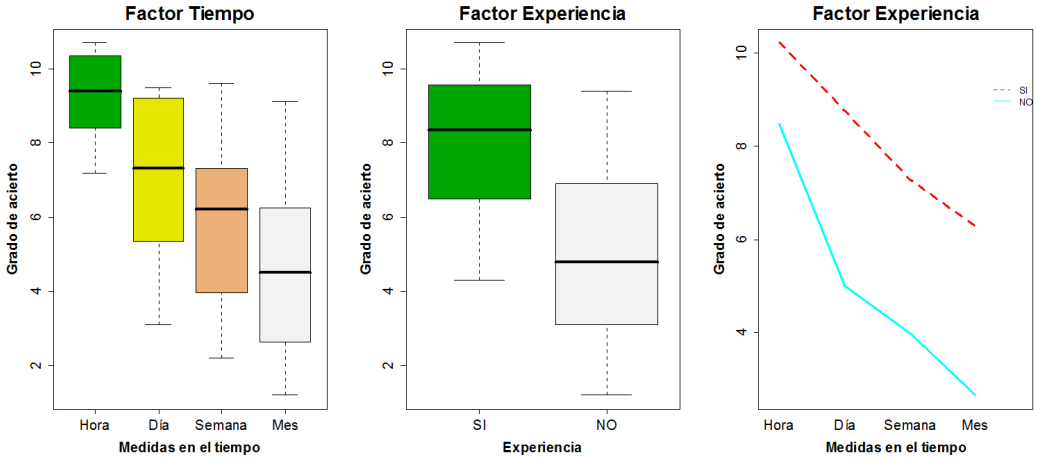
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El modelo lineal se realiza con la función lm y el gráfico de interacción con la función interaction.plot, ambas del paquete stats. El ANOVA se realiza con la función Anova del paquete car (Fox et al., 2014).

EJEMPLO

En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vinos de diferentes regiones. El objetivo era determinar si la capacidad de acertar la procedencia del vino variaba en el tiempo (después de una hora, un día, una semana y un mes), y en función de la experiencia. Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12.

La Figura IX.7 muestra claramente como el grado de acierto de los catadores va disminuyendo con el tiempo (panel izquierdo), pero los catadores con experiencia tienen un grado de acierto mayor que los que no la tienen (panel central) y, por último, el gráfico de interacción parece mostrar que el grado de acierto, o intensidad del recuerdo, disminuye con el tiempo desde una hora hasta un mes, de forma similar en ambos grupos (ambas líneas son aproximadamente paralelas, especialmente entre un día y un mes). La intensidad es menor cuando el tiempo pasa, y los expertos mantienen siempre un grado de acierto dos o tres puntos más alto que los no expertos. Sin embargo, es necesario comprobar si esta diferencia es estadísticamente significativa.

Figura IX.7. Panel mostrando el grado de acierto de los catadores en el tiempo (panel izquierdo), el valor medio de los catadores con y sin experiencia (panel central) y el gráfico de interacción (panel derecho).



El factor «Experiencia» es significativo, con $p < 0{,}001$. El factor «Tiempo» también lo es ($p < 0{,}001$), como habíamos comprobado ya anteriormente, y también la interacción entre ambos, con $p = 0{,}006$, lo que significa que la evolución del recuerdo es distinta en ambos grupos, y las curvas no son realmente paralelas como nos había parecido antes. Por tanto, la capacidad de identificar el origen del vino disminuye con diferente intensidad a lo largo del tiempo entre los expertos y no expertos.

```
[1] "ANOVA para medidas repetidas y no repetidas"

[[2]]

Univariate Type II Repeated-Measures ANOVA Assuming Sphericity

                     SS num Df Error SS den Df         F    Pr(>F)
(Intercept)       3374.2      1    54.896     17 1044.9128 < 2.2e-16 ***
Experience         183.6      1    54.896     17   56.8426 8.105e-07 ***
Factor             245.2      3    44.612     51   93.4530 < 2.2e-16 ***
Experience:Factor   12.2      3    44.612     51    4.6679  0.005876 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Un requisito del estadístico *F* en el ANOVA con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad (varianzas iguales y covarianzas nulas), o que la dispersión conjunta es esférica. La tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como el valor *p* es mayor que 0,05 ($p = 0{,}075$) se acepta la hipótesis de esfericidad.

```
Mauchly Tests for Sphericity

                   Test statistic  p-value
Factor                    0.52822 0.074782
Experience:Factor         0.52822 0.074782


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                    GG eps Pr(>F[GG])
Factor             0.69471  4.598e-15 ***
Experience:Factor 0.69471    0.01484 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                      HF eps   Pr(>F[HF])
Factor            0.7936535 6.695820e-17
Experience:Factor 0.7936535 1.096768e-02
```

Si no se cumpliera el requisito de esfericidad (si el valor *p* fuese menor que 0,05), se debería utilizar alguna de las dos modificaciones que también se muestran: Greenhouse-Geisser, y Huynh-Feldt, para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor *p* del contraste ANOVA. En este caso ambos valores *p* corregidos $4,598 * 10^{-15}$ y $6,69 * 10^{-17}$ son prácticamente cero, por lo que muestran también una relación significativa entre el tiempo y el acierto.

Value

Se obtiene un archivo TXT con los resultados del ANOVA y se puede representar un boxplot o un beanplot para cada factor, además de representarse un gráfico de interacción de cada uno de los factores con el factor tiempo.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Examples

```
## Not run:

data(ZIX5)

#Variación de la capacidad de acertar la procedencia del
#vino en función del tiempo y la experiencia del catador
```

```
IX6(data = ZIX5, varTiempo = c("Hour","Day","Week","Month"), Factor1 = "Experience",
Error = "Sommelier", mfrow = c(1,3), YLAB = "Grado de acierto",
PART = c("cex.axis = 1.5", "cex.lab = 1.5", "font.lab = 2", "mar = c(5,5,3,2)"),
LabelCatT = c("Hora","Día","Semana","Mes"), LabelCat1 = c("SI","NO"),
BOXPLOT1 = c("ylab=YLAB", "xlab='Experiencia'", "col=color", "main='Factor Experiencia'",
"cex.main=2"), INTERPLOT1 = c("ylab=YLAB", "xlab=XLAB", "legend=TRUE",
"main='Factor Experiencia'","cex.main=2", "col=rainbow(length(unique(datos4[,Factor1])))",
"trace.label=''"))


## End(Not run)
```

IX7 *ANCOVA PARA FACTORES CON MEDIDAS REPETIDAS Y NO REPETIDAS*

Description

Se aplica un ANCOVA en el cual hay factores con medidas repetidas (factores intra-sujetos), factores sin medidas repetidas(factores inter-sujetos) y una covariable, además de poder representar los datos con un boxplot o un beanplot, y se representa un gráfico de interacción de cada uno de los factores con el factor tiempo.

Usage

```
IX7(data, varTiempo, Factor1, Covariable, Error, Factor2=NULL, Factor3=NULL, SS=2,
ResetPAR=TRUE, mfrow=NULL, YLAB="Variable problema",
XLAB="Medidas en el tiempo", graficoT="Boxplot", PART=NULL, LabelCatT=NULL,
COLORT=NULL, BOXPLOTT=NULL, BEANPLOTT=NULL, LEGENDT=NULL, AXIST=NULL,
MTEXTT= NULL, TEXTT=NULL, grafico1="Boxplot", PAR1=NULL, OrderCat1=NULL,
LabelCat1=NULL, COLOR1=NULL, BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL,
AXIS1=NULL, MTEXT1= NULL, TEXT1=NULL, grafico2="Boxplot",  PAR2=NULL,
OrderCat2=NULL, LabelCat2=NULL, COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL,
LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, grafico3="Boxplot",
PAR3=NULL, OrderCat3=NULL, LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL,
BEANPLOT3=NULL, LEGEND3=NULL, AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL,
INTERPLOT1=NULL, INTERPLOT2=NULL, INTERPLOT3=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varTiempo	Columnas en las que se ha medido en el tiempo la variable dependiente.
Factor1	Primer factor.
Covariable	Covariable.
Error	Es la variable que codifica los sujetos estudiados.
Factor2	Segundo factor.

Factor3	Tercer factor.
SS	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. Como en un modelo de medidas repetidas todas las casillas del modelo tienen igual número de datos (modelo equilibrado), usaremos por defecto el tipo II. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
mfrow	Si es NULL y hay varios gráficos, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
YLAB	Leyenda del eje Y en el boxplot y beanplot.
XLAB	Leyenda del eje X en el boxplot y beanplot.
graficoT	Si es NULL no se realiza el gráfico del factor tiempo y las otras opciones son "Boxplot" o "Beanplot".
PART	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor tiempo.
LabelCatT	Permite especificar un vector con los nombres de las categorías del gráfico del factor tiempo.
COLORT	Vector con el color de las categorías del gráfico del factor tiempo.
BOXPLOTT	Permite especificar las características del boxplot del gráfico del factor tiempo.
BEANPLOTT	Permite especificar las características del beanplot gráfico del factor tiempo.
LEGENDT	Permite incorporar una leyenda al gráfico del factor tiempo.
AXIST	Permite añadir ejes al gráfico del factor tiempo.
MTEXTT	Permite añadir textos en los márgenes del gráfico del factor tiempo.
TEXTT	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor tiempo.
grafico1	Si es NULL no se realiza el gráfico del factor 1 y las otras opciones son "Boxplot" o "Beanplot".
PAR1	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 1.
OrderCat1	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 1.
LabelCat1	Permite especificar un vector con los nombres de las categorías del gráfico 1.
COLOR1	Vector con el color de las categorías del gráfico 1.
BOXPLOT1	Permite especificar las características del boxplot del factor 1.
BEANPLOT1	Permite especificar las características del beanplot del factor 1.

LEGEND1	Permite incorporar una leyenda al gráfico del factor 1.
AXIS1	Permite añadir ejes al gráfico del factor 1.
MTEXT1	Permite añadir textos en los márgenes del gráfico del factor 1.
TEXT1	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 1.
grafico2	Si es NULL no se realiza el gráfico del factor 2 y las otras opciones son "Boxplot" o "Beanplot".
PAR2	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 2.
OrderCat2	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 2.
LabelCat2	Permite especificar un vector con los nombres de las categorías del gráfico 2.
COLOR2	Vector con el color de las categorías del gráfico 2.
BOXPLOT2	Permite especificar las características del boxplot del factor 2.
BEANPLOT2	Permite especificar las características del beanplot del factor 2.
LEGEND2	Permite incorporar una leyenda al gráfico del factor 2.
AXIS2	Permite añadir ejes al gráfico del factor 2.
MTEXT2	Permite añadir textos en los márgenes del gráfico del factor 2.
TEXT2	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 2.
grafico3	Si es NULL no se realiza el gráfico del factor 3 y las otras opciones son "Boxplot" o "Beanplot".
PAR3	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 3.
OrderCat3	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 3.
LabelCat3	Permite especificar un vector con los nombres de las categorías del gráfico 3.
COLOR3	Vector con el color de las categorías del gráfico 3.
BOXPLOT3	Permite especificar las características del boxplot del factor 3.
BEANPLOT3	Permite especificar las características del beanplot del factor 3.
LEGEND3	Permite incorporar una leyenda al gráfico del factor 3.
AXIS3	Permite añadir ejes al gráfico del factor 3.
MTEXT3	Permite añadir textos en los márgenes del gráfico del factor 3.
TEXT3	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 3.
INTERPLOT1	Permite especificar las características del gráfico de interacción del factor 1.
INTERPLOT2	Permite especificar las características del gráfico de interacción del factor 2.
INTERPLOT3	Permite especificar las características del gráfico de interacción del factor 3.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.7. ANOVA para factores con medidas repetidas y no repetidas

Véase la sección *details* de la función IX5 o Guisande et al. (2011) para la explicación sobre este tipo de ANOVA.

FUNCIONES

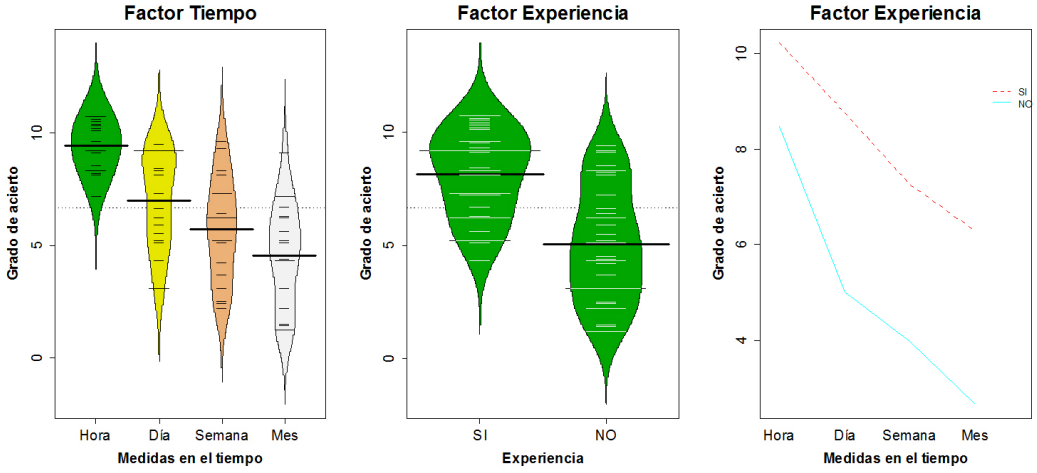
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El modelo lineal se realiza con la función lm, el ANOVA con la función aov y el gráfico de interacción se realiza con la función interaction.plot, todas ellas del paquete stats.

EJEMPLO

En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vinos de diferentes regiones. El objetivo era determinar si la capacidad de acertar la procedencia del vino variaba en el tiempo (después de una hora, un día, una semana y un mes), y en función de la experiencia. Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12. Además, se tuvo en cuenta la posible influencia de la edad de los catadores, la cual se consideró como una covariable.

La Figura IX.8 muestra claramente como el grado de acierto de los catadores va disminuyendo con el tiempo (panel izquierdo), pero los catadores con experiencia tienen un grado de acierto mayor que los que no la tienen (panel central) y, por último, el gráfico de interación parece mostrar que el grado de acierto, o intensidad del recuerdo, disminuye con el tiempo desde una hora hasta un mes, de forma similar en ambos grupos (ambas líneas son aproximadamente paralelas, especialmente entre un día y un mes). La intensidad es menor cuando el tiempo pasa, y los expertos mantienen siempre un grado de acierto dos o tres puntos más alto que los no expertos. Sin embargo, es necesario comprobar si esta diferencia es estadísticamente significativa.

Figura IX.8. Panel mostrando el grado de acierto de los catadores en el tiempo (panel izquierdo), el valor medio de los catadores con y sin experiencia (panel central) y el gráfico de interacción (panel derecho).



El factor «Experiencia» es significativo, con p < 0,001. El factor «Tiempo» también lo es (p < 0,001), como habíamos comprobado ya anteriormente, y también la interacción entre ambos, con p = 0,044, lo que significa que la evolución del recuerdo es distinta en ambos grupos, y las curvas no son realmente paralelas como nos había parecido antes. Por tanto, la capacidad de identificar el origen del vino disminuye con diferente intensidad a lo largo del tiempo entre los expertos y no expertos. Por último, la edad también tiene un efecto significativo sobre el grado de acierto, con p < 0,001.

```
[1] "ANCOVA para medidas repetidas y no repetidas"

[[2]]

Univariate Type II Repeated-Measures ANOVA Assuming Sphericity

                     SS num Df Error SS den Df         F    Pr(>F)
(Intercept)       3374.2      1   22.340     16 2416.6845 < 2.2e-16 ***
Age                 32.6      1   22.340     16   23.3178 0.0001851 ***
Experience         109.0      1   22.340     16   78.0653 1.494e-07 ***
Factor             245.2      3   29.968     48  130.9367 < 2.2e-16 ***
Age:Factor          14.6      3   29.968     48    7.8186 0.0002383 ***
Experience:Factor    5.4      3   29.968     48    2.8989 0.0445054 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Un requisito del estadístico *F* en el ANOVA con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad (varianzas iguales y covarianzas nulas), o que la dispersión conjunta es esférica. La tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como el valor *p* es mayor que 0,05 (p = 0,57) se acepta la hipótesis de esfericidad.

```
Mauchly Tests for Sphericity

                   Test statistic p-value
Factor                     0.7706 0.57395
Age:Factor                 0.7706 0.57395
Experience:Factor          0.7706 0.57395


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                   GG eps Pr(>F[GG])
Factor             0.8426  < 2.2e-16 ***
Age:Factor         0.8426  0.0006031 ***
Experience:Factor 0.8426  0.0549458 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     HF eps   Pr(>F[HF])
Factor            1.013762 4.102332e-23
Age:Factor        1.013762 2.382577e-04
Experience:Factor 1.013762 4.450543e-02
```

Si no se cumpliera el requisito de esfericidad (si el valor *p* fuese menor que 0,05), se debería utilizar alguna de las dos modificaciones que también se muestran: Greenhouse-Geisser, y Huynh-Feldt, para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor *p* del contraste ANOVA. En este caso ambos valores *p* corregidos $2,2 * 10^{-16}$ y $4,1 * 10^{-23}$ son prácticamente cero, por lo que muestran también una relación significativa entre el tiempo y el acierto.

Value

Se obtiene un archivo TXT con los resultados del ANOVA y se puede representar un boxplot o un beanplot para cada factor, además de representarse un gráfico de interacción de cada uno de los factores con el factor tiempo.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Examples

```
## Not run:

data(ZIX5)

#Variación de la capacidad de acertar la procedencia del
#vino en función del tiempo, y la experiencia y edad del catador (covariable)

IX7(data = ZIX5, varTiempo = c("Hour","Day","Week","Month"), Factor1 = "Experience",
Covariable = "Age" , Error = "Sommelier", mfrow = c(1,3), YLAB = "Grado de acierto",
graficoT = "Beanplot", PART = c("cex.axis = 1.5", "cex.lab = 1.5", "font.lab = 2",
"mar = c(5,5,3,2)"), LabelCatT = c("Hora","Día","Semana","Mes"), grafico1 = "Beanplot",
LabelCat1 = c("SI","NO"), BEANPLOT1 = c("ylab=YLAB", "xlab='Experiencia'",
"col=color", "main='Factor Experiencia'","cex.main=2"), INTERPLOT1 = c("ylab=YLAB",
"xlab=XLAB", "legend=TRUE", "main='Factor Experiencia'", "cex.main=2",
"col=rainbow(length(unique(datos4[,Factor1])))", "trace.label=''"))


## End(Not run)
```

IX8 *ANOVA ANIDADO*

Description

Se aplica un análisis de varianza anidado y se puede representar los datos con un boxplot o un beanplot.

Usage

```
IX8(data, varInteres,  Factor1, CF1=NULL, Factor2, CF2=1,
Factor3=NULL,
CF3=NULL, Factor4=NULL, CF4=NULL, combinado=FALSE,  trans=NULL, ResetPAR=TRUE,
mfrow=NULL, grafico1="Boxplot", PAR1=NULL,  OrderCat1=NULL, LabelCat1=NULL,
COLOR1=NULL, BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL,
MTEXT1= NULL, TEXT1=NULL, grafico2="Boxplot", PAR2=NULL, OrderCat2=NULL,
LabelCat2=NULL, COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL,
AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, grafico3="Boxplot", PAR3=NULL,
OrderCat3=NULL, LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL, BEANPLOT3=NULL,
LEGEND3=NULL, AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL, grafico4="Boxplot",
PAR4=NULL, OrderCat4=NULL, LabelCat4=NULL, COLOR4=NULL, BOXPLOT4=NULL,
BEANPLOT4=NULL, LEGEND4=NULL, AXIS4=NULL, MTEXT4= NULL, TEXT4=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable dependiente.
Factor1	Primer factor.
CF1	Factores de los que anida el Factor1. Si este factor no anida de ninguno se pone NULL, que es la opción por defecto.
Factor2	Segundo factor.
CF2	Factores de los que anida el Factor2. Por ejemplo, un valor de 1 significaría que anida del factor 1.
Factor3	Tercer factor.
CF3	Factores de los que anida el Factor3. Por ejemplo, un vector c(2,1) significaría que anida de los factores 2 y 1. Un valor de 2 significa que anida solamente del factor 2.
Factor4	Cuarto factor.
CF4	Factores de los que anida el Factor4. Por ejemplo, un vector c(3,2,1) significaría que anida de los factores 3, 2 y 1. Un vector c(3,1) significaría que anida de los factores 3 y 1.
combinado	Si es TRUE se muestra un panel gráfico que combina los factores. Sin embargo, si el número de combinaciónes es mayor de 12, entonces se desactiva automáticamente y pasa a FALSE.
trans	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3

9. EXP (exponencial)
10. ASN (arcoseno)

`ResetPAR` Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.

`mfrow` Si es NULL y hay varios gráficos, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.

`grafico1` Si es NULL no se realiza el gráfico del factor 1 y las otras opciones son "Boxplot" o "Beanplot".

`PAR1` Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 1.

`OrderCat1` Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 1.

`LabelCat1` Permite especificar un vector con los nombres de las categorías del gráfico 1.

`COLOR1` Vector con el color de las categorías del gráfico 1.

`BOXPLOT1` Permite especificar las características del boxplot del factor 1.

`BEANPLOT1` Permite especificar las características del beanplot del factor 1.

`LEGEND1` Permite incorporar una leyenda al gráfico del factor 1.

`AXIS1` Permite añadir ejes al gráfico del factor 1.

`MTEXT1` Permite añadir textos en los márgenes del gráfico del factor 1.

`TEXT1` Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 1.

`grafico2` Si es NULL no se realiza el gráfico del factor 2 y las otras opciones son "Boxplot" o "Beanplot".

`PAR2` Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 2.

`OrderCat2` Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 2.

`LabelCat2` Permite especificar un vector con los nombres de las categorías del gráfico 2.

`COLOR2` Vector con el color de las categorías del gráfico 2.

`BOXPLOT2` Permite especificar las características del boxplot del factor 2.

`BEANPLOT2` Permite especificar las características del beanplot del factor 2.

`LEGEND2` Permite incorporar una leyenda al gráfico del factor 2.

`AXIS2` Permite añadir ejes al gráfico del factor 2.

`MTEXT2` Permite añadir textos en los márgenes del gráfico del factor 2.

`TEXT2` Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 2.

`grafico3` Si es NULL no se realiza el gráfico del factor 3 y las otras opciones son "Boxplot" o "Beanplot".

PAR3	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 3.
OrderCat3	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 3.
LabelCat3	Permite especificar un vector con los nombres de las categorías del gráfico 3.
COLOR3	Vector con el color de las categorías del gráfico 3.
BOXPLOT3	Permite especificar las características del boxplot del factor 3.
BEANPLOT3	Permite especificar las características del beanplot del factor 3.
LEGEND3	Permite incorporar una leyenda al gráfico del factor 3.
AXIS3	Permite añadir ejes al gráfico del factor 3.
MTEXT3	Permite añadir textos en los márgenes del gráfico del factor 3.
TEXT3	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 3.
grafico4	Si es NULL no se realiza el gráfico del factor 4 y las otras opciones son "Boxplot" o "Beanplot".
PAR4	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico del factor 4.
OrderCat4	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico 4.
LabelCat4	Permite especificar un vector con los nombres de las categorías del gráfico 4.
COLOR4	Vector con el color de las categorías del gráfico 4.
BOXPLOT4	Permite especificar las características del boxplot del factor 4.
BEANPLOT4	Permite especificar las características del beanplot del factor 4.
LEGEND4	Permite incorporar una leyenda al gráfico del factor 4.
AXIS4	Permite añadir ejes al gráfico del factor 4.
MTEXT4	Permite añadir textos en los márgenes del gráfico del factor 4.
TEXT4	Permite añadir texto en cualquier zona de la parte interna del gráfico del factor 4.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.1.8. ANOVA anidado

El ANOVA anidado es un diseño experimental en el que los niveles de uno o más factores dependen de los niveles de otros factores. Es una extensión de un ANOVA de un factor, en el cual se divide cada grupo en subgrupos. Se dice que un factor B está anidado en otro factor A (o que sus niveles están anidados en los de A) cuando todos los niveles del factor B aparecen asociados a un único nivel del factor A. Se denota como B incluido en A. Es posible tener varios factores anidados en el diseño experimental, es decir, tener grupos, subgrupos para cada grupo, subsubgrupos para cada subgrupo, etc.

El análisis de varianza anidado tiene una hipótesis nula para cada factor. En un ANOVA anidado de dos vías, una hipótesis nula sería que los subgrupos dentro de cada grupo tienen la misma media, la segunda hipótesis nula sería que los grupos tienen la misma media. Los requisitos del ANOVA anidado son iguales a los del ANOVA, es decir, se requiere que los residuos tengan distribución Normal y homogeneidad de varianzas.

Un ejemplo de un ANOVA anidado es plantar una especie de árbol en tres hábitats diferentes, por ejemplo, en el valle, en la ladera de la montaña y en la alta montaña, y repetir ese diseño experimental en dos zonas diferentes, por ejemplo, una zona que es un parque natural y, por tanto, existe cierto grado de protección, y otra zona que no tiene ningún tipo de control. El objetivo sería determinar si existen variaciones en la tasa de crecimiento de los árboles (variable dependiente) entre las dos zonas (factor primario), y entre los hábitats dentro de cada zona (factor secundario anidado).

En otro ejemplo, supongamos que se está probando la hipótesis nula de si monos sometidos a estrés y no sometidos a estrés tienen la misma cantidad de anticuerpos. Como los monos ocupan mucho espacio, es necesario tenerlos en cuatro habitaciones diferentes y en cada una de ellas hay monos de los dos tipos, a los que se les ha sometido y no ha sometido a estrés. Se puede dar el caso de que las condiciones de las habitaciones (luz, temperatura, etc.) pueden causar algunas diferencias en el nivel de estrés y, por tanto, tal vez en el nivel de anticuerpos. Para eliminar este potencial efecto de las diferentes habitaciones en el diseño experimental, se debe hacer una ANOVA anidado en el que el primer factor sería las habitaciones (con cuatro niveles, las 4 habitaciones) y el factor anidado sería el nivel de estrés de los monos (con dos niveles, estresados y no estresados). Con este diseño se puede ver si existen diferencias en la cantidad de anticuerpos (variable dependiente) entre habitaciones y en la cantidad de anticuerpos entre el grupo de monos estresados y no estresados dentro de cada habitación.

Además de informar si existen diferencias en las medias entre los niveles de cada factor, en las medias entre cada nivel del subfactor dentro de cada factor, etc., el ANOVA anidado es muy útil para futuros diseños experimentales, ya que informa del grado de variabilidad dentro de cada factor. Por ejemplo, imaginemos que se miden los niveles de glucógeno en varios tipos de músculos en ratas sometidas y no sometidas a estrés. Por tanto, el factor músculo estaría anidado dentro del factor estrés y la variable dependiente sería el nivel de glucógeno. Se puede dar el caso de que la variabilidad en el contenido de glucógeno entre músculos sea muy pequeña, mientras que entre ratas sea alta. Por tanto, en el siguiente diseño habría que centrar el esfuerzo en medir el glucógeno en un mayor número de ratas y tan solo en un músculo dentro de cada rata, es decir, ya no sería necesario un ANOVA anidado pues nos quedaríamos con un solo factor, el nivel de estrés al que se somete a las ratas.

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors.

La función shapiro.test del paquete base stats para realizar el test de Shapiro-Wilk. El ANOVA se realiza con la función Anova del paquete car (Fox et al., 2014). Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013).

El test de Levene se realiza con la función levene.test del paquete lawstat (Gastwirth et al., 2013). La asimetría y la curtosis se realizan con las funciones skewness y kurtosis, respectivamente, del paquete e1071 (Meyer et al., 2014).

EJEMPLO 1

Paso 1.

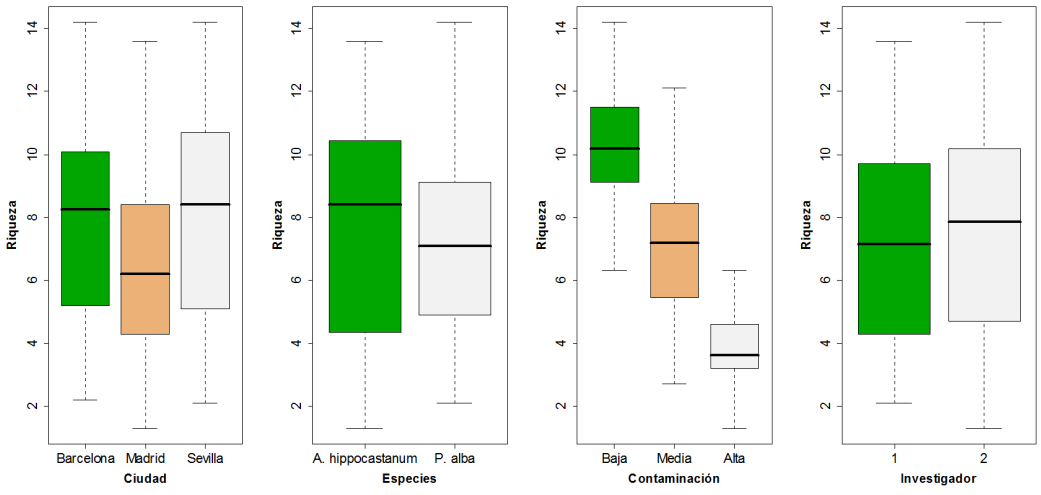
Los líquenes son organismos considerados bioindicadores de contaminación, ya que se ha demostrado que su presencia suele ser menor en zonas más contaminadas. Los datos que se usan en el ejemplo son el promedio de especies de líquenes encontradas en dos especies de árbol, *Populus alba* (Chopo blanco) y *Aesculus hippocastanum* (Castaño de indias), en tres ciudades de España (Madrid, Barcelona y Sevilla), en tres tipos de ambientes (baja, media y alta contaminación) y mostrando también el investigador que realizó en recuento de especies, por si podría verse afectado este dato en función de la persona que realiza el muestreo. El promedio es el resultado de contar las especies de líquenes en 40 árboles de cada una de las especies (*Populus alba* y *Aesculus hippocastanum*), en cada una de las ciudades, en los tres tipos de ambientes y por cada uno de los investigadores.

El objetivo es determinar si el número de especies de líquenes varía en función del investigador, que hace el recuento. Como el número de especies de líquenes puede verse afectado por la especie de árbol y/o por el grado de contaminación de las ciudades, es necesario realizar un ANOVA anidado.

Lo más importante en un ANOVA anidado es anidar de forma correcta los factores. En este estudio el factor 1 (Ciudad) no se anida y, por ello, se dejan las condiciones por defecto *CF1=NULL*. El factor 2 (Especie) está anidado en el 1 *CF2=1*. El factor 3 esta anidado en los factores 2 y 1 *CF3=c(2,1)*. Por último, el factor investigador está anidado en los tres anteriores *CF4=c(3,2,1)*.

La Figura IX.9 muestra los promedios y rango de la riqueza de los líquenes para cada una de las categorías de los factores analizados. Se observa que la riqueza es mayor en Barcelona y Sevilla, y menor en Madrid. Hay más riqueza en la especie *Aesculus hippocastanum* que en *Populus alba*, aunque es necesario comprobar si las diferencias son significativas. Se observa claramente que la riqueza disminuye a medida que aumenta la contaminación y, en principio, los dos investigadores hicieron un recuento de especies de líquenes similar.

Figura IX.9. Panel mostrando el promedio y rango de la riqueza de líquenes entre ciudades, especies, grado de contaminación y entre investigadores.



En primer lugar nos fijamos en la Normalidad de los residuos. Se observa que tanto el test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,775), como el de Shapiro-Wilk (p = 0,615) muestran p > 0,05 y, por tanto, los residuos se ajustan a una distribución Normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0391, p-value = 0.7754


[[4]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9931, p-value = 0.6146
```

El test de Levene considerando la mediana muestra que también se cumple la homogeneidad de varianzas (p > 0,05) para casi todos los factores, con la excepción del factor 3 (contaminación). Por tanto, es necesario realizar una transformación.

```
[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 4.765, p-value = 0.009746
```

Paso 2.

Se transforman los datos aplicando la raíz cuadrada con el argumento *trans = "SQR"*. Se observa de nuevo que tanto el test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,308), como el de Shapiro-Wilk (p = 0,501) muestran p > 0,05 y, por tanto, los residuos se ajustan a una distribución Normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0529, p-value = 0.3085


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9921, p-value = 0.501
```

Además ahora hay homogeneidad de varianzas en todos los factores, ya que p > 0,05.

```
[1] "Factor1"

[[6]][[2]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 2.8225, p-value = 0.06237


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.6406, p-value = 0.4246



[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.0336, p-value = 0.9669


[[6]][[7]]
[1] "Factor4"

[[6]][[8]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.0232, p-value = 0.8791
```

Los resultados del ANOVA se muestran en la siguiente tabla. Existen diferencias significativas entre las ciudades en el número de especies de líquenes (ANOVA, $F_{2,130}$= 15,42, p < 0,001), y también entre las dos especies de árboles (ANOVA, $F_{3,130}$= 3,17, p = 0,026). Existen claras diferencias significativas en el número de especies de líquenes observadas en función del grado de contaminación (ANOVA, $F_{12,130}$= 39,343, p < 0,001). En lo que se refiere al objetivo del ejemplo, los dos investigadores realizan un recuento de líquenes semejante ya que no existen diferencias significativas entre ellos (ANOVA, $F_{18,30}$= 1,45, p = 0,117).

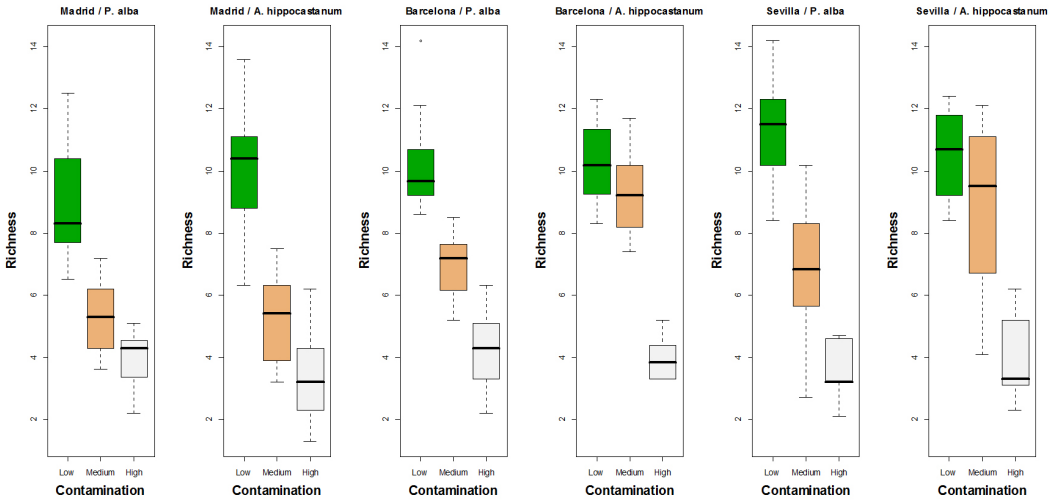
```
[1] "ANOVA ANIDADO"

[[2]]
                                      Df Sum Sq Mean Sq F value    Pr(>F)
City                                   2   2.76   1.381  15.420 9.79e-07 ***
City:Species                           3   0.85   0.284   3.175    0.0264 *
City:Species:Contamination            12  42.28   3.523  39.343  < 2e-16 ***
City:Species:Contamination:Researcher 18   2.34   0.130   1.454    0.1175
Residuals                            130  11.64   0.090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EJEMPLO 2

En el segundo ejemplo solamente se trabaja con tres factores para determinar si la riqueza de líquenes varía en función del grado de contaminación, por tanto no se incluye el factor investigador. Además se hace otro supuesto de anidamiento, de tal forma que el factor 2 (Especie) está anidado en el 1 *CF2=1*, pero el factor 3 solamente anida del 2 y no del 1 *CF3=2*.

El argumento *combinado=TRUE*, lo cual significa que se realiza automáticamente un panel gráfico que combina los factores (Figura IX.10). Como el tercer factor es la contaminación, es el que sale en el eje X de todas las gráficas y se combina con cada una de las categorías de los otros factores. Al ser el tercer factor el que se combina con el resto de factores, todos los cambios se deben realizar a nivel de este tercer factor. Por ejemplo, la modificación del orden de las categorías con el argumento *OrderCat3=c("Low","Medium","High")*, o las modificaciones del gráfico con el argumento *PAR3*. Es posible que de el error *figure margins too large*, en cuyo caso se maximiza la ventana del gráfico y se vuelve a correr el script.

Figura IX.10. Panel mostrando el promedio y rango de la riqueza de líquenes para cada nivel de contaminación combinando con las categorías de los otros factores.



En primer lugar nos fijamos en la Normalidad de los residuos. Se observa que tanto el test Kolmogorov-Smirnov con la corrección de Lilliefors (p = 0,421), como el de Shapiro-Wilk (p = 0,4) muestran p > 0,05 y, por tanto, los residuos se ajustan a una distribución Normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0492, p-value = 0.4211


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9912, p-value = 0.4003
```

El test de Levene considerando la mediana muestra que también se cumple la homogeneidad de varianzas (p > 0,05) para todos los factores. Por lo tanto, no es necesario transformar los datos.

```
[1] "Factor1"

[[6]][[2]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 2.8556, p-value = 0.0604


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.9585, p-value = 0.1636


[[6]][[5]]
[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.9874, p-value = 0.1404
```

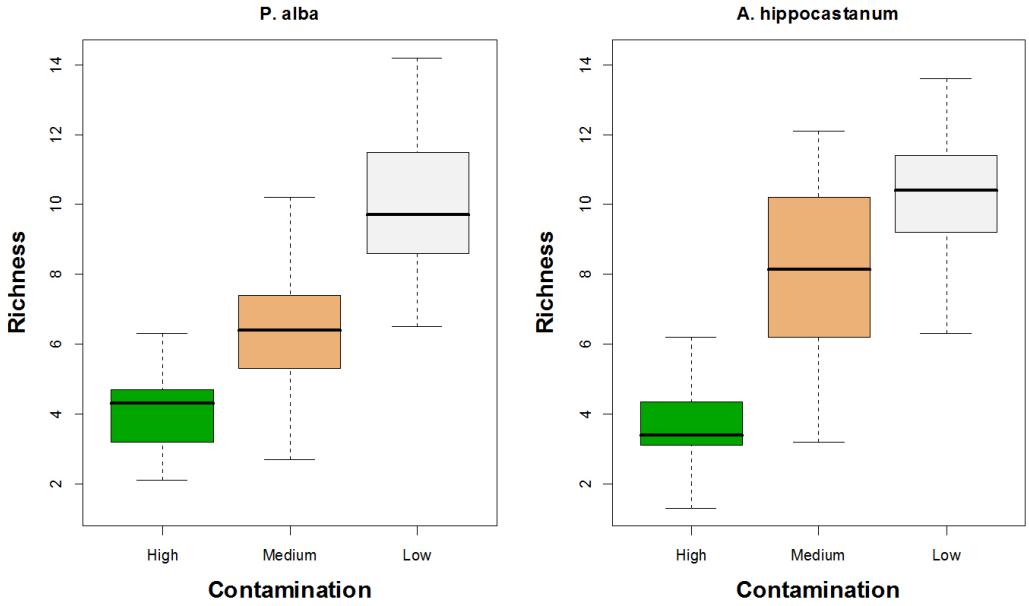
Los resultados del ANOVA se muestran en la siguiente tabla. Existen diferencias significativas entre las ciudades en el número de especies de líquenes (ANOVA, $F_{2,156}$= 13,19, p < 0,001), pero no entre las dos especies de árboles considerando la situación en la que se anida en el factor ciudad(ANOVA, $F_{2,156}$= 2,09, p = 0,127). En lo que se refiere al objetivo del ejemplo, existen diferencias significativas en el número de especies de líquenes observadas en función del grado de contaminación (ANOVA, $F_{4,156}$= 94,347, p < 0,001).

```
[1] "ANOVA ANIDADO"

[[2]]
                      Df Sum Sq Mean Sq F value    Pr(>F)
City                   2   74.0   37.02  13.191 5.09e-06 ***
Species                1   14.9   14.93   5.320    0.0224 *
City:Species           2   11.7    5.86   2.087    0.1275
Species:Contamination  4 1059.2  264.80  94.347  < 2e-16 ***
Residuals            156  437.8    2.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EJEMPLO 3

En el último script se trabaja con 2 factores(Figura IX.11).

Figura IX.11. Panel mostrando el promedio y rango de la riqueza de líquenes para cada nivel de contaminación combinando con las especies.



Value

Se obtiene un archivo TXT con los resultados del ANOVA anidado y se puede representar un boxplot o un beanplot para cada factor o realizar una combinación entre factores.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Examples

```
## Not run:

data(ZIX6)

#Ejemplo 1. Se trabaja con los 4 factores

#Paso 1

IX8(data = ZIX6 , varInteres = "Richness", Factor1 = "City", Factor2 = "Species",
CF2=1, Factor3 = "Contamination", CF3=c(2,1), Factor4 = "Researcher",CF4=c(3,2,1),
mfrow = c(1,4) , PAR1 = c("mar=c(5,5,3,2)","font.lab=2",  "cex.lab=1.5","cex.axis=1.7"),
BOXPLOT1 = c("col = color" , "xlab = 'Ciudad'" , "ylab = 'Riqueza'") ,
BOXPLOT2 = c("col = color", "xlab = 'Especies'", "ylab = 'Riqueza'"),
OrderCat3 = c("Low","Medium","High"), LabelCat3 = c("Baja","Media","Alta"),
BOXPLOT3 = c("col = color", "xlab = 'Contaminación'", "ylab = 'Riqueza'"),
BOXPLOT4 = c("col = color", "xlab = 'Investigador'", "ylab = 'Riqueza'"))

#Paso 2. Se transforman los datos

IX8(data = ZIX6 , varInteres = "Richness", Factor1 = "City", Factor2 = "Species",
CF2=1, Factor3 = "Contamination", CF3=c(2,1), Factor4 = "Researcher",CF4=c(3,2,1),
mfrow = c(1,4) , PAR1 = c("mar=c(5,5,3,2)","font.lab=2",  "cex.lab=1.5","cex.axis=1.7"),
BOXPLOT1 = c("col = color" , "xlab = 'Ciudad'" , "ylab = 'Riqueza'") ,
BOXPLOT2 = c("col = color", "xlab = 'Especies'", "ylab = 'Riqueza'"),
OrderCat3 = c("Low","Medium","High"), LabelCat3 = c("Baja","Media","Alta"),
BOXPLOT3 = c("col = color", "xlab = 'Contaminación'", "ylab = 'Riqueza'"),
BOXPLOT4 = c("col = color", "xlab = 'Investigador'", "ylab = 'Riqueza'"), trans = "SQR")

##Ejemplo 2. Se trabaja con los 3 factores

windows(16,8)

IX8(data=ZIX6, varInteres="Richness",   Factor1 = "City" , Factor2 = "Species" ,
CF2=1, Factor3 = "Contamination" ,CF3=2, combinado=TRUE, OrderCat3=c("Low",
"Medium","High"), PAR3 = c("mar=c(5,5,3,2)","font.lab=2", "cex.lab=1.7") )

##Ejemplo 3. Dos factores

windows(16,8)

IX8(data=ZIX6, varInteres="Richness", Factor1= "Species",
Factor2="Contamination",
CF2=1, combinado=TRUE, OrderCat3=c("Low", "Medium","High"), PAR3 =c("mar=c(5,5,3,2)",
"font.lab=2", "cex.lab=1.7"))

## End(Not run)
```

IX9 *CONTRASTES DE HOMOGENEIDAD NO PARAMETRICOS PARA DOS MUESTRAS INDEPENDIENTES*

Description

Se aplica el test de Kolmogorov-Smirnov para dos muestras y el contraste de la suma de rangos de Wilcoxon, además de poder representar los datos con un boxplot o un beanplot.

Usage

```
IX9(data, varInteres, varPoblacion, varComparar1, varComparar2,
grafico="Boxplot", PAR=NULL, ResetPAR=TRUE, YLAB=NULL, XLAB=NULL,
OrderCat=NULL, LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL,
LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`varInteres`	Variable dependiente.
`varPoblacion`	Define cual es la variable que actúa como factor.
`varComparar1`	Primer grupo a comparar de la variable población.
`varComparar2`	Segundo grupo a comparar de la variable población.
`grafico`	Si es NULL no se realiza ningún gráfico y las otras opciones son "Boxplot" o "Beanplot".
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`YLAB`	Leyenda del eje Y en el boxplot y beanplot.
`XLAB`	Leyenda del eje X en el boxplot y beanplot.
`OrderCat`	Permite especificar un vector con el orden en que se muestran las categorías en el gráfico.
`LabelCat`	Permite especificar un vector con los nombres de las categorías del gráfico.
`COLOR`	Vector con el color de las categorías del gráfico.
`BOXPLOT`	Permite especificar las características del boxplot.
`BEANPLOT`	Permite especificar las características del beanplot.
`LEGEND`	Permite incorporar una leyenda.
`AXIS`	Permite añadir ejes.
`MTEXT`	Permite añadir textos en los márgenes del gráfico.
`TEXT`	Permite añadir texto en cualquier zona de la parte interna del gráfico.
`file`	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

IX.2. PRUEBAS NO PARAMÉTRICAS

En muchas ocasiones no podemos asumir los supuestos sobre normalidad e igualdad de varianzas en nuestros datos, por tanto es necesario utilizar la llamada estadística no paramétrica o métodos de distribución libre. Igualmente, podemos utilizar los tests no paramétricos cuando los datos de la población están en forma de rango, cuando el procedimiento de obtención de la muestra impide asimilarla con una sucesión de valores cuantitativos (Pérez, 2004).

Tienen como ventaja que son de aplicación más general que las paramétricas (a las variables distribuidas normalmente también se les puede aplicar estos contrastes), ya que no exigen ninguna condición sobre el tipo de distribución. Sin embargo, son menos sensibles a la detección de diferencias que las pruebas paramétricas, aunque de forma general se puede decir que la coincidencia entre los resultados obtenidos con las pruebas paramétricas y las no paramétricas es superior al 90%.

Dentro de los contrastes de homogeneidad no paramétricos podemos distinguir dos tipos diferentes en función de la naturaleza de las muestras. Tendríamos muestras independientes cuando los elementos sean seleccionados aleatoriamente de forma separada o no relacionada (por ejemplo, medida del nivel de azúcar en hombres y mujeres), y muestras dependientes, cuando las mediciones si están relacionadas, generalmente porque corresponden al mismo individuo (por ejemplo, medida del nivel de azúcar en diabéticos, antes y después de un tratamiento).

IX.2.1. Contrastes para dos muestras independientes

Son los contrastes análogos al contraste paramétrico *t* de Student. Existen tres pruebas principales que tratan de determinar si las diferencias observadas entre dos muestras son debidas al azar o si pertenecen a dos poblaciones distintas. En todas ellas, los datos se ordenan de forma ascendente sin considerar su grupo de pertenencia y, en función de su posición, se les asigna un número de orden o rango.

IX.2.1.1. Prueba U *de Mann-Whitney*

Es una prueba que compara la tendencia central de dos muestras, que no es necesario que tengan el mismo tamaño, partiendo de la hipótesis nula de que en ambas muestras la medida central es la misma.

Es un contraste muy utilizado, aunque en el caso de que se quiera medir no sólo diferencias en la tendencia central de los datos sino en otras características como, por ejemplo, asimetría o dispersión de los datos, es más adecuado utilizar el test de Wald-Wolfowitz.

Una premisa de esta prueba es que los valores de cada una de las muestras son diferentes, es decir, que no hay solapamiento de datos (valores repetidos). En caso de que existan solapamientos, el contraste se vuelve más conservador (es más difícil encontrar diferencias entre muestras). Si el número de solapamientos es muy grande, sobre todo en muestras pequeñas y con variables continuas, se aconseja la utilización del test de Kolmogorov-Smirnov para dos muestras, aunque existe una corrección para solapamientos del propio estadígrafo.

A las muestras ordenadas de menor a mayor se les asigna un rango de forma ascendente (1,2,3,...), habiendo tantos rangos como datos. En caso de solapamiento o empate se suman los rangos implicados en el mismo y se divide por el número de datos solapados.

El desarrollo completo de esta prueba aparece en Sokal & Rohlf (1981). Consiste en calcular el estadístico *U* con el que es posible conocer si la media de los rangos es significativamente diferente entre las dos muestras. Si el número total de datos es inferior a 20, se utiliza una tabla para la

determinación de valores críticos de *U*. En cambio, si el número total de datos es superior a 20, se puede utilizar una aproximación a la Normal a partir del estadístico *U*, que nos indica el valor *Z* de dicha Normal y su probabilidad.

IX.2.1.2. Test de rachas de Wald-Wolfowitz

Es una prueba que compara la distribución de dos muestras, que no es necesario que tengan el mismo tamaño, partiendo de la hipótesis nula de que la distribución de rangos es aleatoria, es decir, las muestras son homogéneas.

Los datos de las dos muestras se ordenan conjuntamente de forma ascendente y a cada grupo de valores se le asigna un valor de racha. Las rachas son secuencias de valores del mismo grupo, cuando las muestras han sido ordenadas.

Una descripción completa del estadístico aparece en el apartado 25.6 del libro *Biostatistical Analysis* (Zar, 1999).

Si dos muestras presentan una misma distribución cabe esperar que, en el ordenamiento de datos de menor a mayor, ambas muestras estén muy mezcladas (aleatorizadas), es decir, que el número de rachas sea alto. Con una tabla de contraste se determina si el número de rachas encontradas es significativamente grande y se puede suponer la aleatoriedad de los rangos y, por tanto, que no hay diferencias entre las muestras. Si el número de datos es grande, al igual que sucedía con el contraste de Mann-Whitney, este estadístico se ajusta a una Normal y se puede analizar la homogeneidad de las muestras calculando el estadístico *Z*.

Este estadístico es menos potente que el contraste de Mann-Whitney, pero tiene la ventaja de que detecta diferencias no sólo de tendencia central sino de dispersión y asimetría.

IX.2.1.3. Prueba de Kolmogorov-Smirnov para dos muestras

Tiene como ventaja que mide las diferencias existentes entre las frecuencias relativas acumuladas de las dos muestras, por lo que se detectan diferencias no sólo en la tendencia central, sino también en la dispersión y forma (simetría, apuntamiento) de las muestras.

Cuando el tamaño de la muestra es grande, es mejor el estadístico *U* de Mann-Whitney. Esta prueba estadística se usa para contrastar la H_0 de que las dos muestras proceden de una misma población, por tanto exige comparar dos funciones de distribución muestral, observando la diferencia máxima entre ellas. El desarrollo completo del cálculo de este estadístico se encuentra en Sokal & Rohlf (1981), en el que los pasos a seguir son:

1. Ordenar las observaciones de cada muestra y distribuirlas en clases.
2. Determinar las frecuencias acumuladas de las clases en cada muestra.
3. Buscar la máxima diferencia entre la frecuencia acumulada de ambas muestras.
4. Determinar los valores críticos de D_{max} en una tabla.
5. Si D_{max} es mayor que el valor crítico, se rechaza la hipótesis nula de que las muestras proceden de la misma población.

FUNCIONES

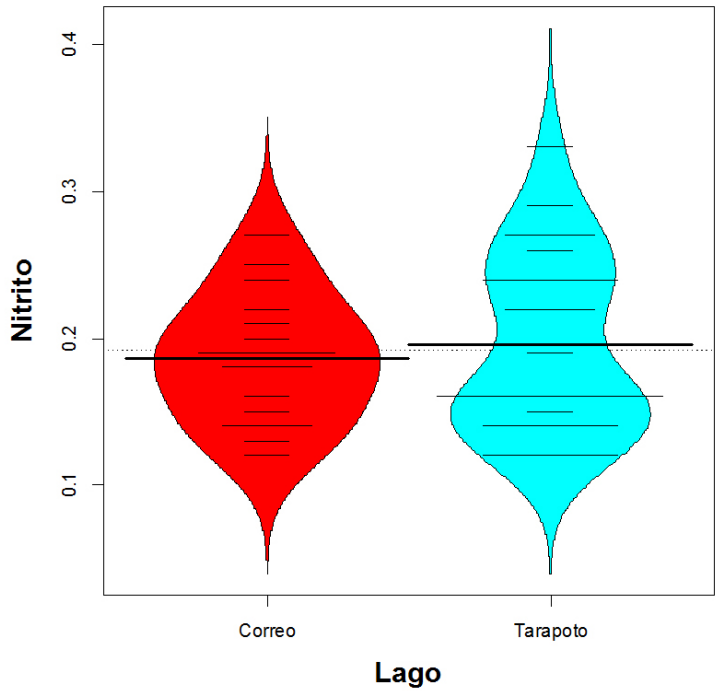
Para el beanplot se utiliza la función beanplot del paquete beanplot (Kampstra, 2008; Kampstra, 2013). El test de la U de Mann-Whitney se realiza con la función wilcox.test y el test de Kolmogorov-Smirnov para dos muestras con la función ks.test, ambas del paquete stats.

EJEMPLO

Datos de la concentración de nitrito, nitrato y amonio (μML) en lagos de la amazonia colombiana en dos meses diferentes. El objetivo es determinar si existen diferencias en la concentración de nitritos entre dos lagos del Amazonas.

La Figura IX.12 muestra que la concentración es muy parecida en ambos lagos.

Figura IX.12. Beanplot mostrando la mediana y distribución de las concentraciones de nitrito (μML) en dos lagos del Amazonas.



El test de Wilcoxon de suma de rangos (también conocido como prueba U de Mann-Whitney) con una p = 0,794 acepta la hipótesis de igualdad, al igual que el de Kolmogorov-Smirnov con una p = 0,831. Es importante mencionar que cuando existen empates, es decir, existen valores iguales en los grupos a comparar, R no calcula el valor exacto de probabilidad, ni en el test de Wilcoxon de suma de rangos ni en el test de Kolmogorov-Smirnov y, por ello, se pueden observar pequeñas diferencias con otros programas estadísticos.

```
        Wilcoxon rank sum test with continuity correction

data:  datos3a$valores and datos3b$valores
W = 185.5, p-value = 0.794
alternative hypothesis: true location shift is not equal to 0


[[2]]

        Exact two-sample Kolmogorov-Smirnov test

data:  datos3a$valores and datos3b$valores
D = 0.19949, p-value = 0.6093
alternative hypothesis: two-sided
```

Value

Se obtiene un archivo TXT con los resultados de los dos test y se puede representar un boxplot o un beanplot.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Pérez, C (2004) *Técnicas de análisis multivariante de datos*. Aplicaciones con SPSS. Pearson Educación. Madrid.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*. WH Freeman and Company, New York.

Zar, J.H. (1999) *Biostatistical Analysis*. Prentice Hall, New Jersey.

Examples

```
## Not run:

data(ZIX7)

IX9(data=ZIX7, varInteres="Nitrito", varPoblacion="Lago", varComparar1="Correo",
varComparar2="Tarapoto", grafico="Beanplot")


## End(Not run)
```

PAR *FUNCION PAR*

Description

Esta función permite realizar modificaciones en los gráficos.

Usage

```
PAR(adj=0.5, ann=TRUE, bg=NULL, bty="o", cex.axis=1, cex.lab=1, cex.main=1,
cex.sub=1, col.axis="black", col.lab="black", col.main="black", col.sub="black",
family="Arial", fg="black", font.axis=1, font.lab=1, font.main=2, font.sub=1,
lab=c(5,5,7), las=0, lend=0, lty= 1, lwd=1, mai=NULL, mar=NULL, mex=NULL,
mgp=c(3,3,0), oma=NULL, omd=NULL, omi=NULL, pch=NULL, ps=NULL, pty="m",
tck=NA, tcl=NA, xaxp=NULL, xaxs="r", xaxt="s", yaxp=NULL, yaxs="r", yaxt="s")
```

Arguments

adj	Especifica la alineación de los textos en los gráficos. Puede tomar cualquier valor entre 0 (texto a la izquierda) y 1 (texto a la derecha).
ann	Si se selecciona FALSE elimina las leyendas de los ejes y los títulos de los gráficos cuando se usa con la función plot.default.
bg	Color de fondo del gráfico.
bty	Permite definir el tipo de marco de los gráficos: "o" (marco completo), "c" (no se representa el borde derecho), "u" (no se representa el borde superior), "l" (no se representan los bordes superior y derecho), "7" (no se representan los bordes inferior e izquierdo), "]" (no se representa el borde izquierdo).
cex.axis	Tamaño de los rótulos de los ejes.
cex.lab	Tamaño del texto de las leyendas.
cex.main	Tamaño del texto del título el gráfico.
cex.sub	Tamaño del texto del subtítulo.
col.axis	Define el color de los rótulos de los ejes.
col.lab	Define el color del texto de las leyendas.
col.main	Define el color del texto del título el gráfico.
col.sub	Define el color del texto del subtítulo.
family	Especifica el tipo de letra de los textos.
fg	Define el color del marco y líneas del gráfico.
font.axis	Valor numérico que define la fuente de los rotulos de los ejes. El valor 1 es la letra normal, 2 es negrita, 3 es itálica y 4 es itálica y negrita.
font.lab	Valor numérico que define la fuente de las leyendas. El valor 1 es la letra normal, 2 es negrita, 3 es itálica y 4 es itálica y negrita.
font.main	Valor numérico que define la fuente del título del gráfico. El valor 1 es la letra normal, 2 es negrita, 3 es itálica y 4 es itálica y negrita.
font.sub	Valor numérico que define la fuente del subtítulo del gráfico. El valor 1 es la letra normal, 2 es negrita, 3 es itálica y 4 es itálica y negrita.
lab	Valor numérico con el formato c(x,y,len) que define el número de divisiones de los ejes. El valor de «len» hay que especificarlo pero R no lo utiliza.
las	Define la posición de las etiquetas de los ejes: 0 es paralelo al eje, 1 es horizontal, 2 es perpendicular al , y 3 es vertical.
lend	Define el estilo del final de la línea y se puede especificar con un número o letras: 0 ó "round" es un extremo redondeado, 1 ó "butt" es un extremo grueso y 2 ó "square" es un extremo cuadrado.
lty	Define el tipo de línea: 0 Sin línea. 1 Línea sólida. 2 Línea discontinua. 3 Línea punteada. 4 Línea de puntos y rayas. 5 Línea de guión largo. 6 Línea de doble raya.
lwd	Define el grosor de línea
mai	Un vector numérico con el formato c(abajo, izquierda, arriba, derecha) que define los márgenes de la figura en pulgadas.

`mar`	Un vector numérico con el formato c(abajo, izquierda, arriba, derecha) que define el número de líneas de los márgenes de la figura.
`mex`	Factor de expansión del tamaño de caracteres en los márgenes.
`mgp`	Vector numérico con el formato c(t,e,l) que especifica la posición de la leyenda de los ejes (t), el texto de las divisiones de los ejes (e) y la línea de los ejes (l). El valor por defecto es (3,1,0).
`oma`	Un vector numérico con el formato c(abajo, izquierda, arriba, derecha) que define los márgenes del texto en el panel, en número de líneas.
`omd`	Vector numérico con el formato c(x1,x2,y1,y2) que define los márgenes internos donde se ubicará el gráfico, en una escala de 0 a 1.
`omi`	Un vector numérico con el formato c(abajo, izquierda, arriba, derecha) que define los márgenes del panel en pulgadas.
`pch`	Valor numérico entre 1 y 25 que define el tipo de símbolo (el símbolo que corresponde a cada número se muestra en el siguiente cuadro). También se puede poner entre "" cualquier carácter, como se muestra a modo de ejemplo en el cuadro con a, *, !, ¡ y $.
`ps`	Tamaño del punto de texto (unidad de medida de los caracteres) en 1/72 pulgadas.
`pty`	Carácter que indica si el gráfico es cuadrado o se alarga lateralmente cuando se amplía la ventana. Con "s" genera una figura cuadrada y con "m" el gráfico se amplia al máximo posible.
`tck`	Valor alfanumérico que define el tamaño de las marcas de los ejes, en proporción al área de la región de la figura. El valor 1 pone líneas interiores en el gráfico. Con tck=NA deja la marca por defecto.
`tcl`	Valor alfanumérico que define el tamaño de las marcas de los ejes. Se diferencia de «tck» en que las medidas son reales y no en proporción. Con tcl=NA deja la marca por defecto.
`xaxp`	Vector numérico con el formato c(x1, x2, n), donde x1 y x2 son los extremos de las marcas de eje x, y n el número de intervalos o marcas en el eje.
`xaxs`	"r" (regular) extiende un 4% más de los límites dados en «xlim», y calcula el número de intervalos que mejor se ajustan a esos límites, y "i" (interno) busca el número de intervalos más adecuado a los datos originales.
`xaxt`	Valor lógico que especifica si se representan los intervalos del eje x, con las opciones "s" o "n".
`yaxp`	Vector numérico con el formato c(x1, x2, n), donde x1 y x2 son los extremos de las marcas de eje y, y n el número de intervalos o marcas en el eje.
`yaxs`	"r" (regular) extiende un 4% más de los límites dados en «ylim», y calcula el número de intervalos que mejor se ajustan a esos límites, y "i" (interno) busca el número de intervalos más adecuado a los datos originales.
`yaxt`	Valor lógico que especifica si se representan los intervalos del eje y, con las opciones "s" o "n".

VI1 *CONTRASTES DE BONDAD DE AJUSTE-POISSON*

Description

Se determina si una muestra se ajusta a una distribución de Poisson.

Usage

```
VI1(data, varInteres, varAgrupacion=NULL, distribucion="Normal", combinar=TRUE,
grafica=TRUE, ni=12, tituloEjeX="", col.bar="cyan", col.line="red", lty=2,
lwd=2, legend.pos="topleft", file="Output.csv", na="NA", dec=",",
row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de bondad de ajuste.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
distribucion	Tipo de distribución teórica a la que se contrasta el ajuste de la muestra: "Binomial", "Poisson" o "Normal".
combinar	Es aconsejable que si el número de casos esperados dentro de una clase es menor que 5, se combinen varias clases en una hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. Por ello es conveniente dejar el valor por defecto TRUE.
grafica	GRÁFICO. Valor lógico que define si se representa la gráfica con las distribuciones esperada y real.
ni	GRÁFICO. Número de barras (intervalos) en el gráfico de la distribución Normal.
tituloEjeX	GRÁFICO. Leyenda del eje X.
col.bar	GRÁFICO. Color de las barras.
col.line	GRÁFICO. Color de la línea de la distribución teórica.
lty	GRÁFICO. Tipo de línea de la distribución teórica.
lwd	GRÁFICO. Grosor de la línea de la distribución teórica.
legend.pos	GRÁFICO. Posición de la leyenda en el gráfico: "topleft", "topright", "bottomleft" o "bottomright"
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

VI. CONTRASTES DE BONDAD DE AJUSTE

Para determinar si las variables se ajustan a una distribución Normal, Poisson, Binomial, etc., es necesario utilizar contrastes estadísticos, los cuales se denominan pruebas de bondad de ajuste.

Existen distintas pruebas de bondad de ajuste que se utilizan en función del tipo de datos y la distribución teórica esperada. Una clasificación de los ajustes más empleados es:

1. Muestras categorizadas (los datos son códigos asignados a los valores de una variable cualitativa o a clases en las que se agrupan los valores de una variable cuantitativa): Chi-cuadrado

2. Muestras no categorizadas (variables cuantitativas, continuas o discretas, no agrupadas en intervalos o clases).

a) Para todas las distribuciones: Prueba Kolmogorov-Smirnov (test K-S).

b) Distribución Normal: Contraste de normalidad Shapiro-Wilk.

VI.1. CHI-CUADRADO

Se puede aplicar tanto a distribuciones continuas (con los datos previamente agrupados en clases) como a distribuciones discretas o variables cualitativas. Se basa en cuantificar las diferencias entre las frecuencias observadas en cada clase y las esperadas, partiendo de la hipótesis nula de que los datos se ajustan a una distribución *F(x)* (que puede ser Normal, Poisson, etc.). Para su aplicación, en las *n* clases existentes, se contabiliza el número de casos observados (O_i) y, a través de la función teórica que se desea contrastar, se calcula el número de casos esperados (E_i). A partir de estas frecuencias se calcula el valor del estadístico χ^2, con la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Se aconseja que si el número de casos esperados dentro de una clase es menor de 5 se combinen varias clases en una, hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. No se debe utilizar cuando hay pocos datos.

A continuación se deben calcular los grados de libertad (υ) de la muestra. Si los valores esperados se pueden calcular de forma previa al muestreo, el número de υ es *n* - 1. Sin embargo, si para calcular los valores esperados es necesario estimar algunos parámetros mediante estadísticos obtenidos en el muestreo (como, por ejemplo, μ o σ para una distribución Normal) el número de υ es *n* - *r* -1 donde *r* es el número de estadísticos necesarios para calcular el valor esperado. Una vez determinados χ^2 y υ, se busca en las tabla χ^2 (Tabla 3 del Apéndice I en Guisande et al., 2011) el valor crítico para esos υ y para el nivel de significación elegido (generalmente α = 0,05). Si el χ^2 crítico es mayor que χ^2 calculado se acepta la hipótesis nula de que los datos observados se ajustan a la distribución esperada, mientras que si χ^2 crítico es menor que χ^2 calculado la hipótesis se rechaza. La mayoría de los programas estadísticos calculan directamente el valor *p* del contraste; si es menor que el nivel de significación elegido (0,05 o 0,01 habitualmente) la hipótesis se rechaza.

VI.2. TEST DE KOLMOGOROV-SMIRNOV

Es la prueba adecuada para contrastar la normalidad de una muestra si el número de datos es grande (*n* > 30), aunque se puede usar tanto para muestras grandes como pequeñas. También se puede usar para contrastar otras distribuciones como la Binomial o de Poisson. Es un test muy conservador que se aplica a variables continuas. Se basa en la determinación de la máxima diferencia (*D*) entre las frecuencias acumuladas observadas (AO_i) y las frecuencias acumuladas esperadas (AE_i), partiendo

de la hipótesis nula de que los datos se ajustan a una distribución determinada. La fórmula del test es $D = max|AO_i - AE_i|$.

Este test fue recalculado para una distribución Normal estudiando las frecuencias esperadas a partir de la media y varianza de la muestra (Lilliefors, 1967) y se conoce como test *K-S-L*. Sin embargo, su aplicación es limitada cuando existen pocos datos (son necesarias 100 observaciones para distinguir entre una Normal con μ = 0 y σ^2 = 1 y una distribución uniforme entre $-\sqrt{3}$ y $\sqrt{3}$.

Una vez calculado el estadístico *D* se contrasta con un valor *D* crítico para el nivel de significación elegido que aparece en la Tabla 5 del Apéndice I (Guisande et al., 2011) y Tabla 6 del Apéndice I (Guisande et al., 2011, para el test *K-S-L*). La hipótesis nula se acepta cuando *D* observado es inferior al valor de *D* tabulado o cuando el valor *p* del contraste es mayor que el nivel de significación (por ejemplo mayor que 0,05).

VI.3. TEST DE SHAPIRO-WILK

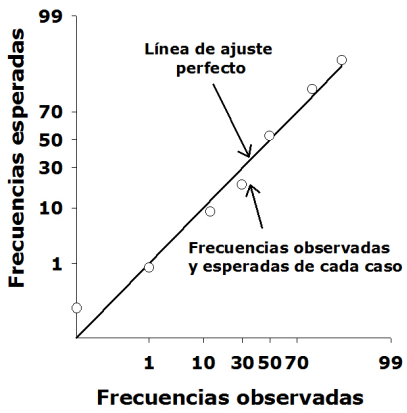
Es la prueba más recomendable para contrastar la normalidad de una muestra, sobre todo si se trabaja con un número pequeño de datos (*n* < 30). Solo se utiliza para contrastar la distribución Normal (Shapiro & Wilk, 1965). Se basa en medir el ajuste de los datos a una recta probabilística Normal (Figura VI.1). Si el ajuste fuera perfecto los puntos formarían una recta de 45º (frecuencia observada igual a frecuencia esperada). El estadístico de contraste se expresa por medio de la siguiente ecuación:

$$W = \frac{1}{\sum_{j=1}^{n}(x_j - \mu)^2}\left[\sum_{j=1}^{h} a_{j,n}(x_{(n-j+1)} - x_j)\right]^2$$

donde *n* es el número de datos, x_j es el dato en orden ascendente de muestra que ocupa el lugar *j*, μ es la media, *h* es *n*/2 si *n* es par o (*n*-1)/2 si *n* es impar y $a_{j,n}$ es un valor tabulado.

Una vez calculado el estadístico *W* se contrasta con un valor *W* crítico para el nivel de significación elegido. Como este estadístico mide el ajuste a una recta y no la distancia a la distribución Normal (como era en los casos anteriores), y se puede interpretar de forma aproximada como un coeficiente de correlación entre los valores observados y esperados (un valor próximo a 1 indica buen ajuste, y próximo a 0 malo), la hipótesis nula se acepta cuando el valor *W* es superior al valor de contraste tabulado (valor de ajuste muy alto).

Figura VI.1. Representación de las probabilidades esperadas y observadas de una muestra junto con la línea de ajuste perfecto a una distribución normal.

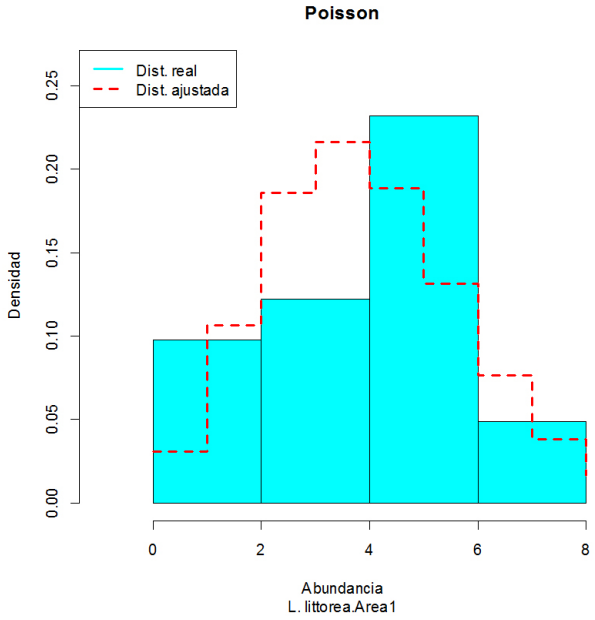


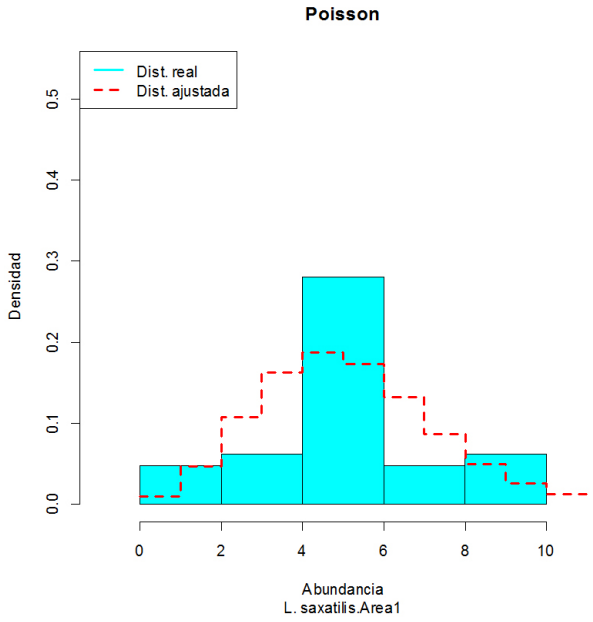
EJEMPLO

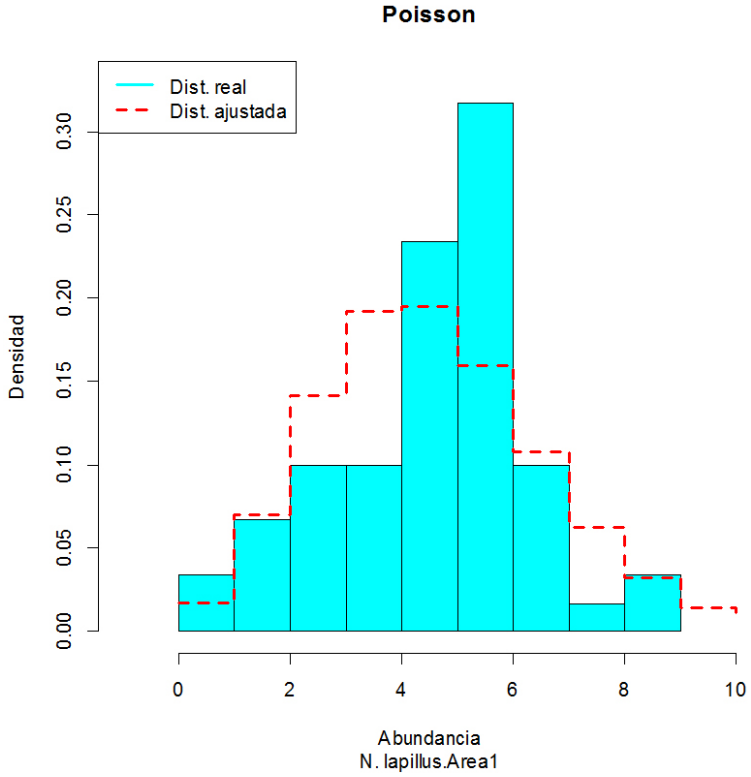
Datos de la abundancia de dos especies de caracoles de mar herbívoras (*Littorina littorea* y *Littorina saxatilis*) y una especie carnívora (*Nucella lapillus*), en dos zonas de muestreo. Se desea conocer si la distribución es aleatoria en las distintas zonas, es decir, si siguen una distribución de Poisson.

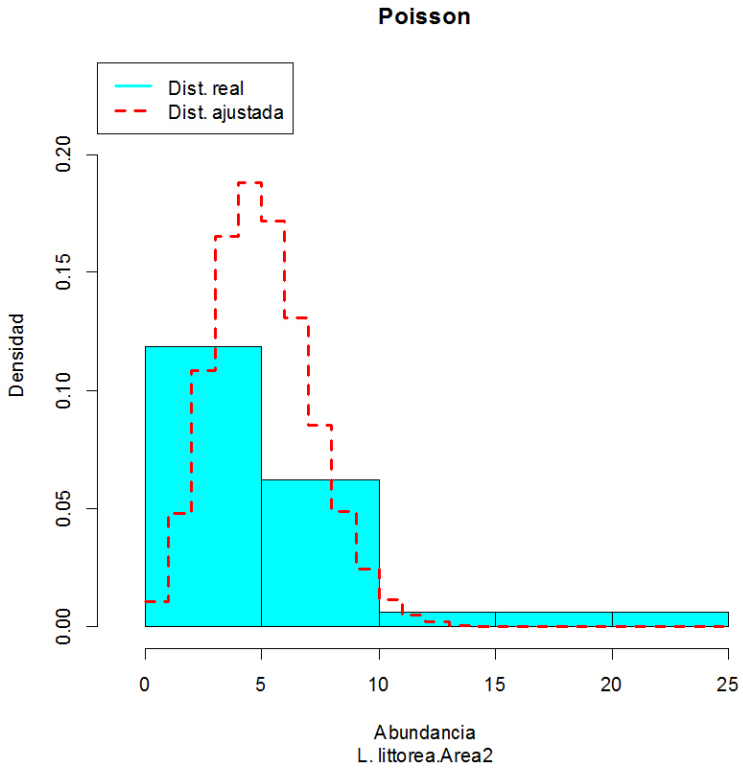
Para poder ver todas las gráficas, las cuales se muestran a continuación (Figura VI.2), solo hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.

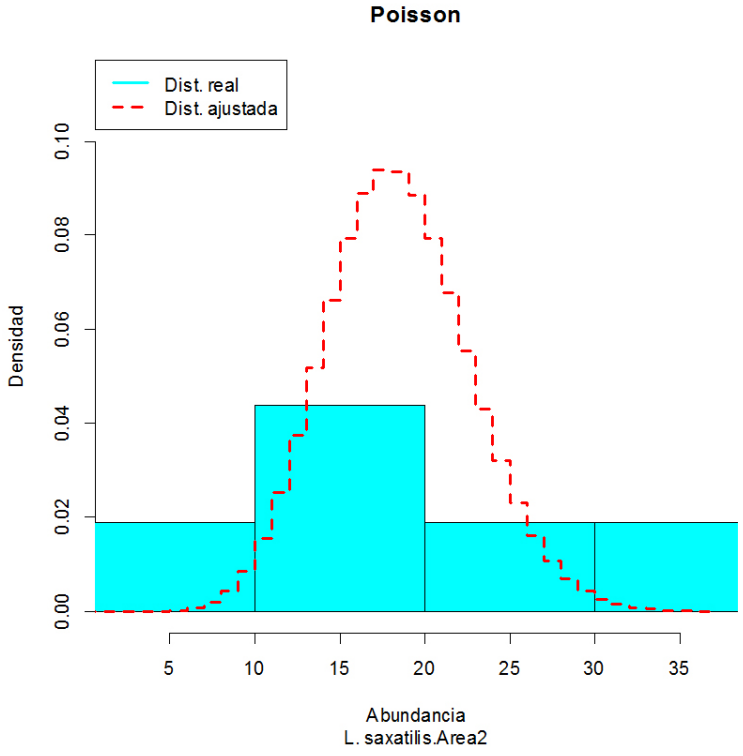
Figura VI.2. Histogramas de frecuencias para cada una de las especies en las dos zonas.

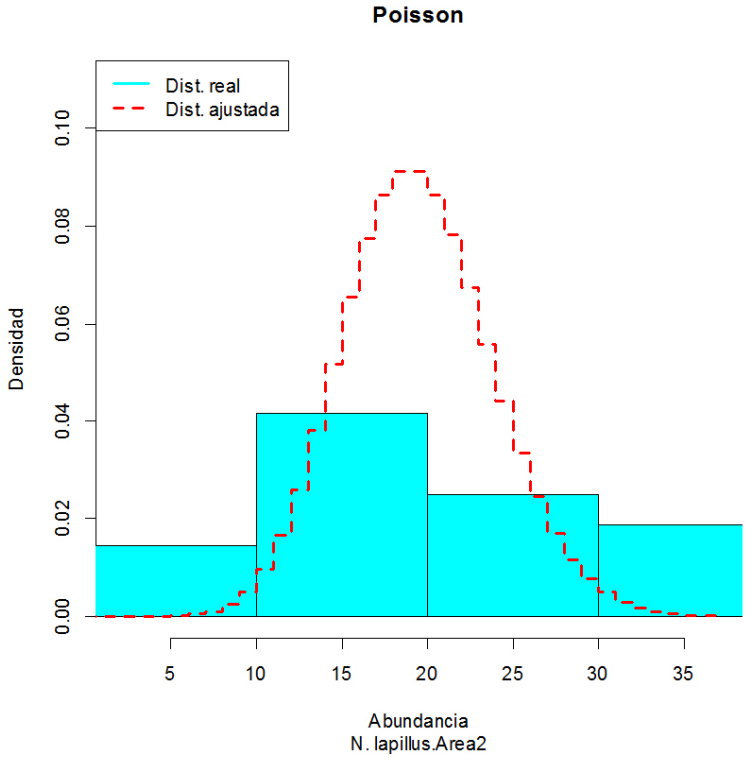



Poisson
Dist. real
Dist. ajustada
Densidad
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Abundancia
L. saxatilis.Area1


Poisson
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
Abundancia
N. lapillus.Area1


Poisson
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
Abundancia
L. littorea.Area2


Poisson
Dist. real
Dist. ajustada
Densidad
0.00
0.02
0.04
0.06
0.08
0.10
5
10
15
20
25
30
35
Abundancia
L. saxatilis.Area2



Los resultados se muestran a continuación:

Nombres	Species	Variable	D de K-S	Valor-p K-S	X Chisq2	Grados de libertad ajustado	Valor-p Chisq2 ajustado
L. littorea.Area1	L. littorea	Area1	0,28872113	0,00214984	12,9149482	4	0,011698936
L. littorea.Area2	L. littorea	Area2	0,15153169	0,45445897	5,12697276	3	0,162732211
L. saxatilis.Area1	L. saxatilis	Area1	0,29565653	0,00743713	12,0614729	3	0,00717552
L. saxatilis.Area2	L. saxatilis	Area2	0,27091062	0,19084638	NA	NA	NA
N. lapillus.Area1	N. lapillus	Area1	0,31581487	1,27E-05	15,7339841	5	0,007646062
N. lapillus.Area2	N. lapillus	Area2	0,2213481	0,01812666	28,0103039	4	1,24E-05

En el caso del test χ^2, la prueba ajustada agrupa las categorías para conseguir frecuencias esperadas no menores que cinco. Cuando los resultados de las pruebas no ajustada y ajustada son distintos, eso significa que ha sido necesario agrupar y, por lo tanto, que la prueba no ajustada no es válida. Los resultados muestran que en todos los casos los valores de probabilidad para el test χ^2 de la prueba ajustada son diferentes de la no ajustada y, por tanto, en todos los casos se tiene que considerar la prueba ajustada como válida.

Para *Littorina littorea* en la zona 1, el valor *p* del contraste χ^2 es menor que 0,05 en la prueba ajustada (p = 0,011), por tanto se rechaza la hipótesis de que la distribución de esta especie es aleatoria en la zona 1, es decir, no sigue una distribución de Poisson. El test de Kolmogorov-Smirnov (*K-S*) corrobora lo antes mencionado ya que p = 0,002 (menor que 0,05). Por el contrario, para la misma especie en la zona 2, el valor *p* del contraste χ^2 es mayor que 0,05 en la prueba ajustada (p = 0,162), lo cual indica que la distribución puede ser de Poisson en esa zona. El test *K-S* corrobora lo antes mencionado ya que p = 0,454 (mayor que 0,05).

En el caso de *Littorina saxatilis*, el valor *p* del contraste χ^2 es menor que 0,05 en la zona 1 (p = 0,007 ajustado), indicando que no sigue una distribución de Poisson, y no se puede estimar en la zona 2.

Con el test de *K-S* la probabilidad también es menor que 0,05 en la zona 1 (p = 0,007), indicando que la distribución no es de Poisson. El test *K-S* es más apropiado que el test χ^2 para variables cuantitativas, por lo que daremos validez al test *K-S*. En cualquier caso, si se decide aplicar varias pruebas válidas con el mismo fin, como en este ejemplo que usamos los test χ^2 y *K-S*, y alguna de ellas rechaza la hipótesis nula y otras no, se debe rechazar la hipótesis nula.

Para *Nucella lapillus*, en las dos zonas, tanto el χ^2 de las pruebas ajustadas como el test *K-S*, el valor p < 0,05, rechazando la hipótesis de distribución aleatoria de Poisson.

Value

Se obtiene un archivo CSV con los resultados de los contrastes y la gráfica con las distribuciones esperada y real si se seleccionó que se mostrara.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Lilliefors, H. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62: 399-402.

Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52: 591-611.

Examples

```
## Not run:

data(ZVI1)

#Distribución de Poisson
VI1(data=ZVI1, varInteres=c("Area1","Area2"), varAgrupacion=c("Species"),
distribucion="Poisson", tituloEjeX="Abundancia")


## End(Not run)
```

VI2 *CONTRASTES DE BONDAD DE AJUSTE-BINOMIAL*

Description

Se determina si una muestra se ajusta a una distribución Binomial.

Usage

```
VI2(data, varInteres, varAgrupacion=NULL, distribucion="Normal", combinar=TRUE,
grafica=TRUE, ni=12, tituloEjeX="", col.bar="cyan", col.line="red", lty=2,
lwd=2, legend.pos="topleft", file="Output.csv", na="NA", dec=",",
row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de bondad de ajuste.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
distribucion	Tipo de distribución teórica a la que se contrasta el ajuste de la muestra: "Binomial", "Poisson" o "Normal".
combinar	Es aconsejable que si el número de casos esperados dentro de una clase es menor que 5, se combinen varias clases en una hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. Por ello es conveniente dejar el valor por defecto TRUE.
grafica	GRÁFICO. Valor lógico que define si se representa la gráfica con las distribuciones esperada y real.
ni	GRÁFICO. Número de barras (intervalos) en el gráfico de la distribución Normal.
tituloEjeX	GRÁFICO. Leyenda del eje X.
col.bar	GRÁFICO. Color de las barras.
col.line	GRÁFICO. Color de la línea de la distribución teórica.
lty	GRÁFICO. Tipo de línea de la distribución teórica.
lwd	GRÁFICO. Grosor de la línea de la distribución teórica.
legend.pos	GRÁFICO. Posición de la leyenda en el gráfico: "topleft", "topright", "bottomleft" o "bottomright"
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

VI. CONTRASTES DE BONDAD DE AJUSTE

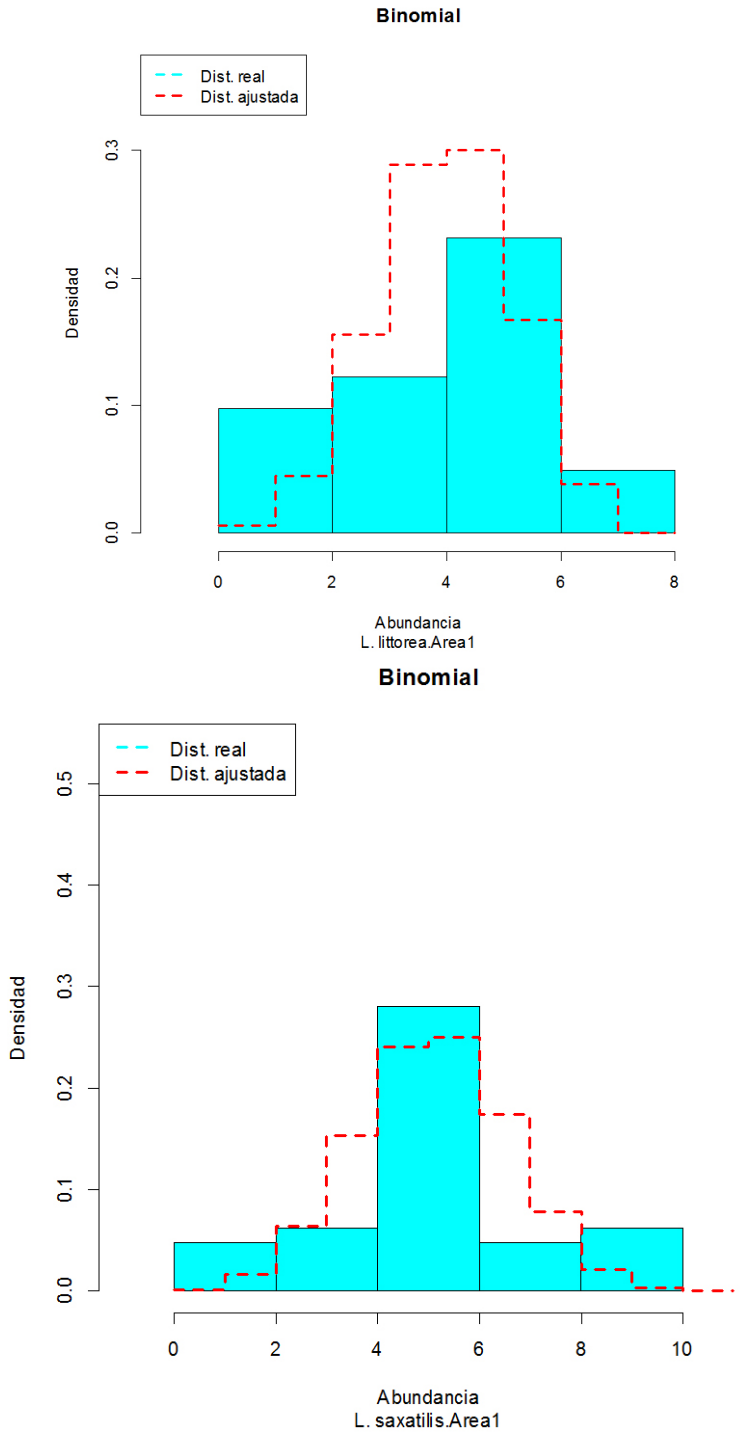
Véase la sección *details* de la función VI1.

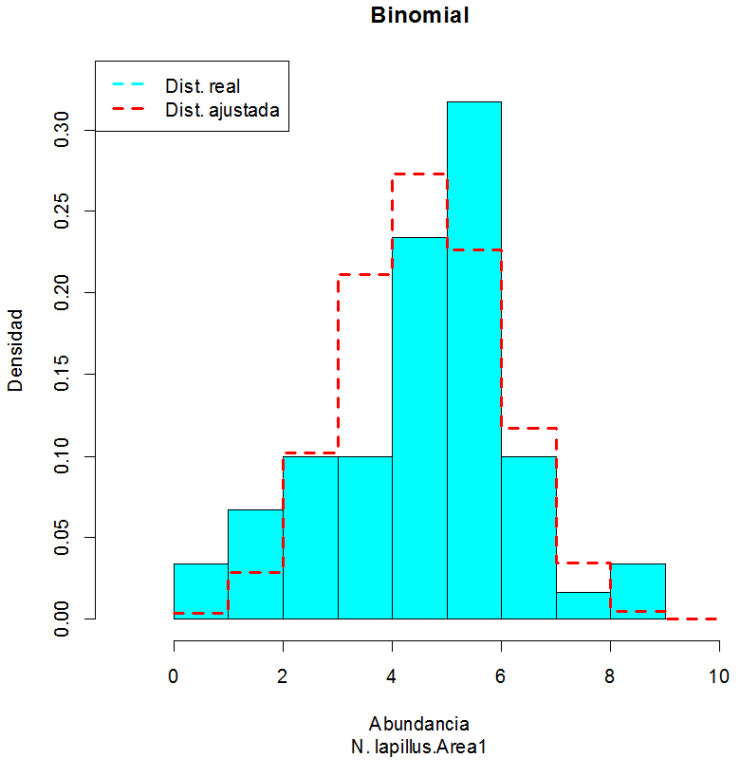
EJEMPLO

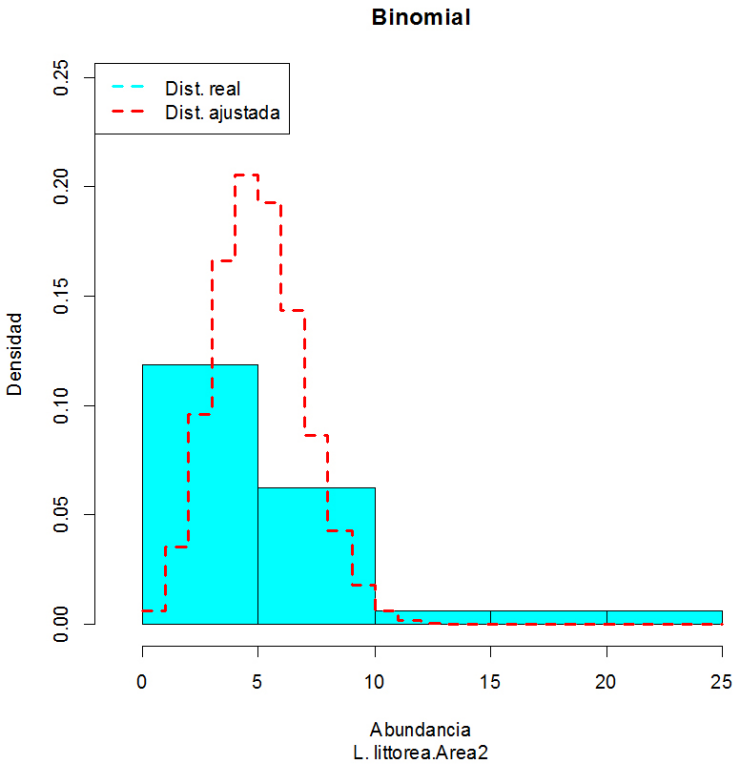
Datos de la abundancia de dos especies de caracoles de mar herbívoras (*Littorina littorea* y *Littorina saxatilis*) y una especie carnívora (*Nucella lapillus*), en dos zonas de muestreo. Se desea conocer si la distribución es Binomial en ambas zonas y para cada especie.

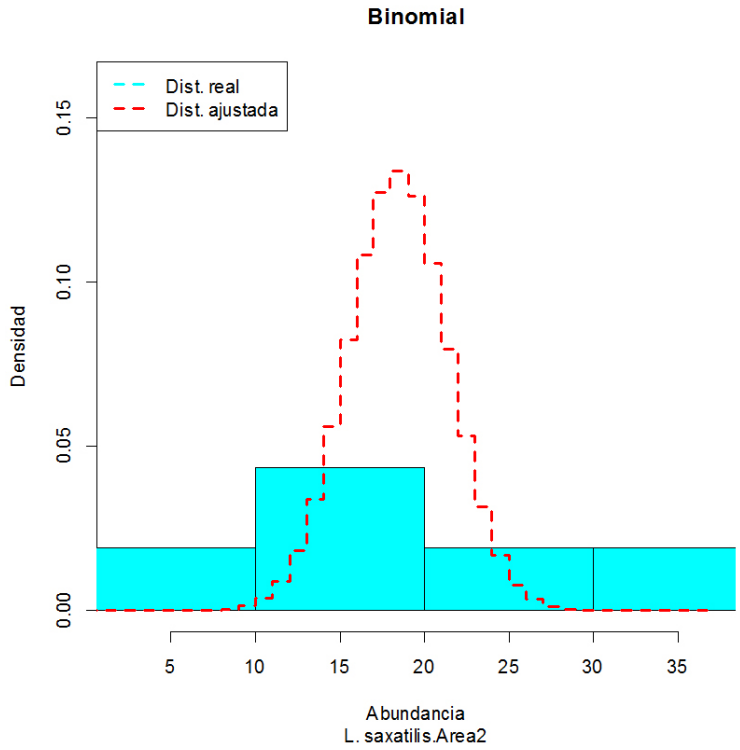
Para poder ver todas las gráficas, las cuales se muestran a continuación (Figura VI.3), solo hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.

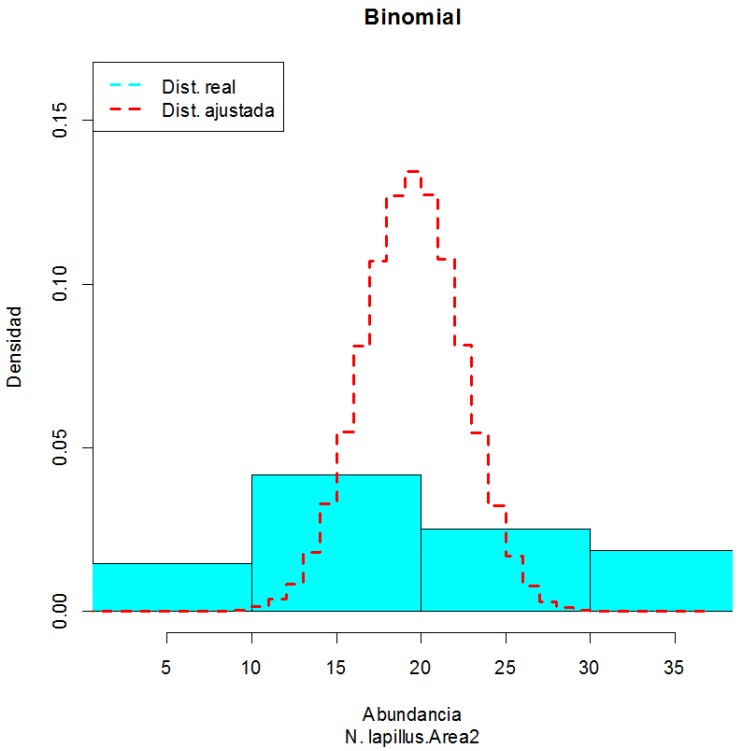
Figura VI.3. Histogramas de frecuencias para cada una de las especies en las dos zonas.




Binomial
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
Abundancia
N. lapillus.Area1


Binomial
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
25
Abundancia
L. littorea.Area2


Binomial
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundancia
L. saxatilis.Area2


Binomial
Dist. real
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundancia
N. lapillus.Area2

Los resultados se muestran a continuación:

Nombres	Species	Variable	D de K-S	Valor-p K-S	X Chisq2	Grados de libertad ajustado	Valor-p Chisq2 ajustado
L. littorea.Area1	L. littorea	Area1	0,35564476	6,26E-05	11,8569317	2	0,002662564
L. littorea.Area2	L. littorea	Area2	0,14610655	0,50168429	4,97694249	2	0,083036812
L. saxatilis.Area1	L. saxatilis	Area1	0,3189149	0,00297925	2,74052173	2	0,25404068
L. saxatilis.Area2	L. saxatilis	Area2	0,35754244	0,03345251	NA	NA	NA
N. lapillus.Area1	N. lapillus	Area1	0,3183428	1,05E-05	8,02796411	3	0,045437293
N. lapillus.Area2	N. lapillus	Area2	0,31841638	0,00011855	70,6281522	5	7,58E-14

En el caso del test χ^2, la prueba ajustada agrupa las categorías para conseguir frecuencias esperadas no menores que cinco. Cuando los resultados de las pruebas no ajustada y ajustada son distintos, eso significa que ha sido necesario agrupar y, por lo tanto, que la prueba no ajustada no es válida. Los resultados muestran que en todos los casos los valores de probabilidad para el test χ^2 de la prueba ajustada son diferentes de la no ajustada y, por tanto, en todos los casos se tiene que considerar la prueba ajustada como válida.

Cuando p < 0,05 se rechaza hipótesis nula y, por tanto, la distribución de las especies no sigue una Binomial. Por el contrario, en el caso de *L. littorea* en la zona 2 (p = 0,083) y *L. saxatilis* en la zona 1 (p = 0,254) las especies muestran una distribución Binomial.

Value

Se obtiene un archivo CSV con los resultados de los contrastes y la gráfica con las distribuciones esperada y real si se seleccionó que se mostrara.

Examples

```
## Not run:

data(ZVI1)

#Distribución Binomial
VI2(data=ZVI1, varInteres=c("Area1","Area2"), varAgrupacion=c("Species"),
distribucion="Binomial", tituloEjeX="Abundancia")


## End(Not run)
```

VI3 *CONTRASTES DE BONDAD DE AJUSTE-NORMAL*

Description

Se determina si una muestra se ajusta a una distribución Normal.

Usage

```
VI3(data, varInteres, varAgrupacion=NULL, distribucion="Normal", combinar=TRUE,
grafica=TRUE, ni=12, tituloEjeX="", col.bar="cyan", col.line="red", lty=2,
lwd=2, legend.pos="topleft", file="Output.csv", na="NA", dec=",",
row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de bondad de ajuste.
varAgrupacion	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
distribucion	Tipo de distribución teórica a la que se contrasta el ajuste de la muestra: "Binomial", "Poisson" o "Normal".
combinar	Es aconsejable que si el número de casos esperados dentro de una clase es menor que 5, se combinen varias clases en una hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. Por ello es conveniente dejar el valor por defecto TRUE.
grafica	GRÁFICO. Valor lógico que define si se representa la gráfica con las distribuciones esperada y real.
ni	GRÁFICO. Número de barras (intervalos) en el gráfico de la distribución Normal.
tituloEjeX	GRÁFICO. Leyenda del eje X.
col.bar	GRÁFICO. Color de las barras.
col.line	GRÁFICO. Color de la línea de la distribución teórica.
lty	GRÁFICO. Tipo de línea de la distribución teórica.
lwd	GRÁFICO. Grosor de la línea de la distribución teórica.
legend.pos	GRÁFICO. Posición de la leyenda en el gráfico: "topleft", "topright", "bottomleft" o "bottomright"
file	ARCHIVO CSV. Nombre del archivo de salida.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

VI. CONTRASTES DE BONDAD DE AJUSTE

Véase la sección *details* de la función VI1.

EJEMPLO

Datos de la abundancia de dos especies de caracoles de mar herbívoras (*Littorina littorea* y *Littorina saxatilis*) y una especie carnívora (*Nucella lapillus*), en dos zonas de muestreo. Se desea conocer si la distribución es Nomal en ambas zonas y para cada especie.

Los resultados se muestran a continuación:

Nombres	Species	Variable	n	Asimetría	D de K-S	Valor-p K-S	D de K-S Lilliefors	Valor-p K-S Lilliefors	W de Shapiro	Valor-p Shapiro
L. littorea.Area1	L. littorea	Area1	41	-0,52822325	0,1839558	0,12468631	0,183955816	0,001266339	0,900981844	0,001772749
L. littorea.Area2	L. littorea	Area2	32	2,50715703	0,3052724	0,005138	0,30527239	3,28E-08	0,690832471	6,36E-07
L. saxatilis.Area1	L. saxatilis	Area1	32	-0,07679185	0,2048303	0,13637883	0,204830284	0,001486642	0,937785769	0,064796923
L. saxatilis.Area2	L. saxatilis	Area2	16	0,32062788	0,1742211	0,71641553	0,174221083	0,21716141	0,919296589	0,16434109
N. lapillus.Area1	N. lapillus	Area1	60	-0,32175659	0,1848496	0,03313438	0,184849635	2,43E-05	0,941172835	0,006112648
N. lapillus.Area2	N. lapillus	Area2	48	0,21846098	0,1278772	0,41240806	0,127877217	0,047854752	0,930699577	0,007236301

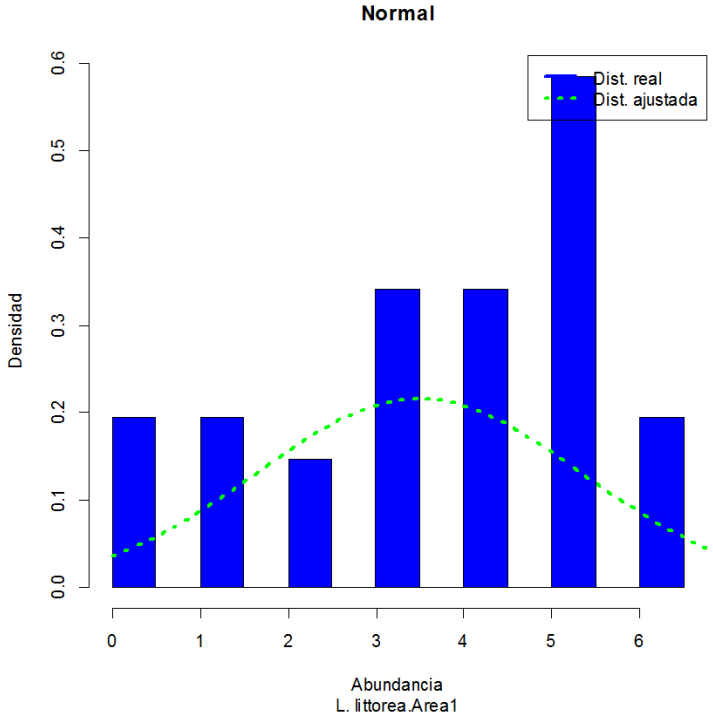
Los resultados también muestran la asimetría de cada una de las series de datos ya que, en el supuesto de que la población no tenga una distribución Normal, es posible transformar los datos y la asimetría nos indica el tipo de transformación que hay que realizar (véase sección *details* de la función IX1).

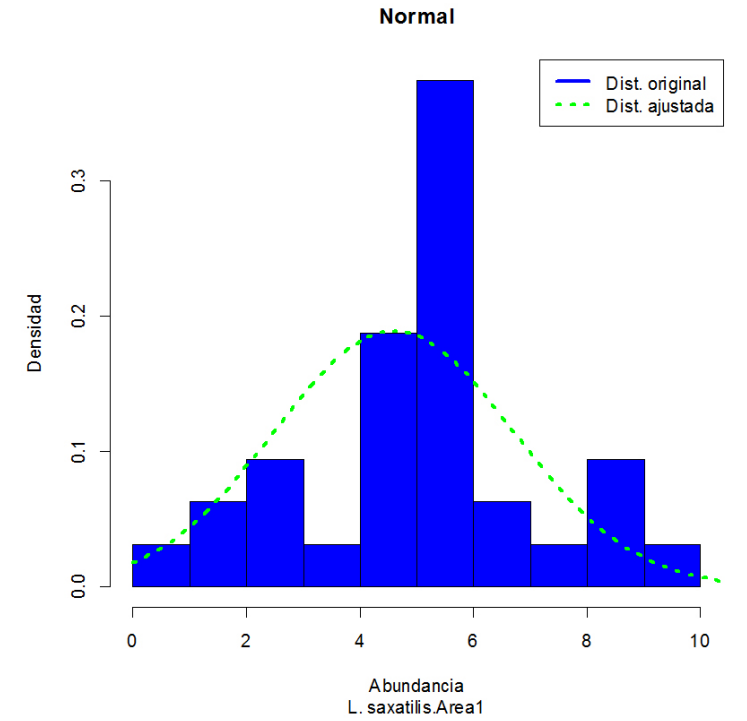
Solamente existe una discrepancia al comparar el test *K-S* con el test Shapiro-Wilk. La especie *L. saxatilis* en la zona 1 no tiene una distribución Normal según el test *K-S* (p = 0,001), mientras que el test Shapiro-Wilk si muestra distribución Normal (p = 0,065).

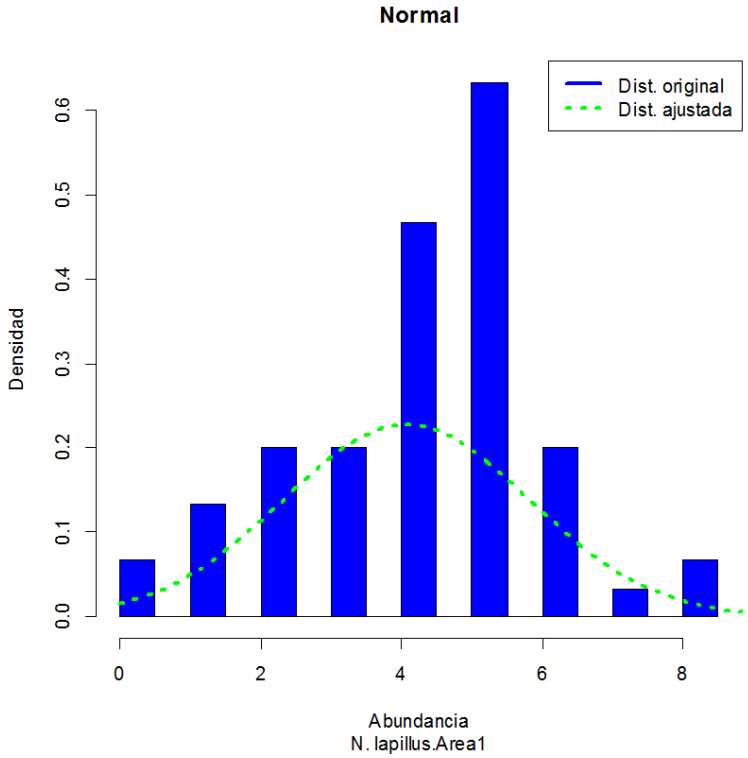
Como ya se ha mencionado en otras ocasiones, si existe discrepancia entre dos pruebas, se debe tomar una decisión conservadora, es decir, rechazar la hipótesis nula y, por tanto, en este ejemplo se rechaza que la distribución sea Normal.

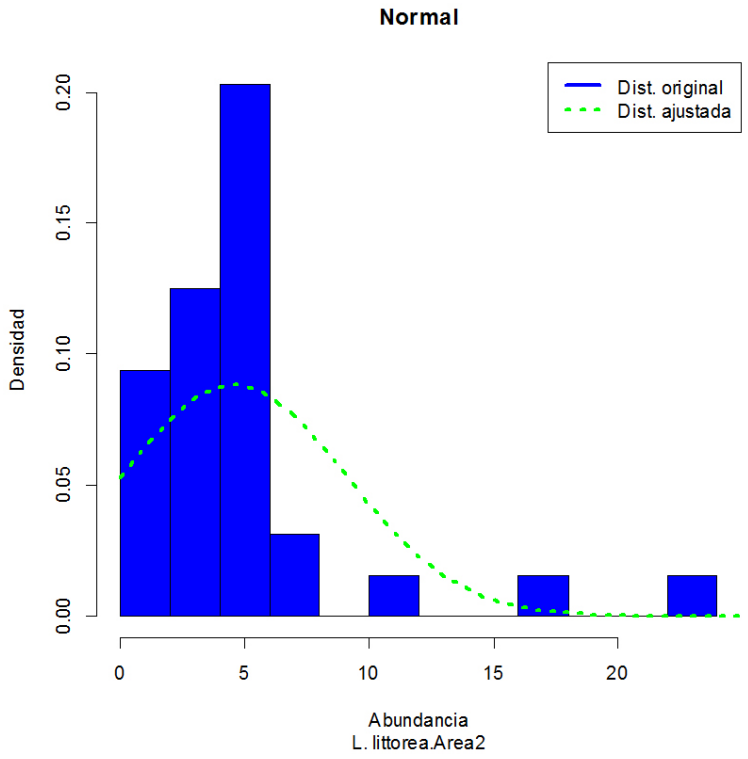
Para poder ver todas las gráficas, las cuales se muestran a continuación (Figura VI.4), solo hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.

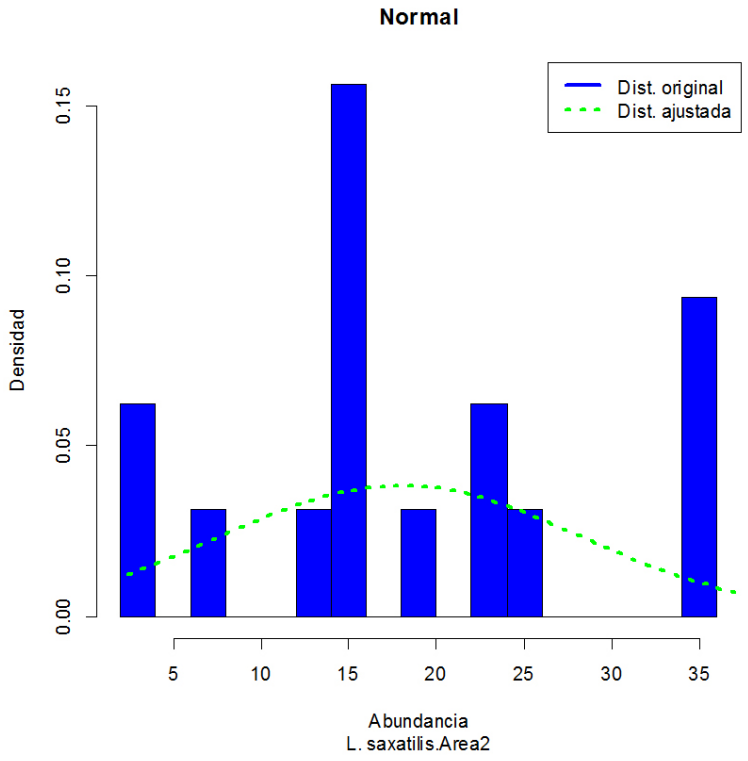
Figura VI.4. Histogramas de frecuencias para cada una de las especies en las dos zonas.

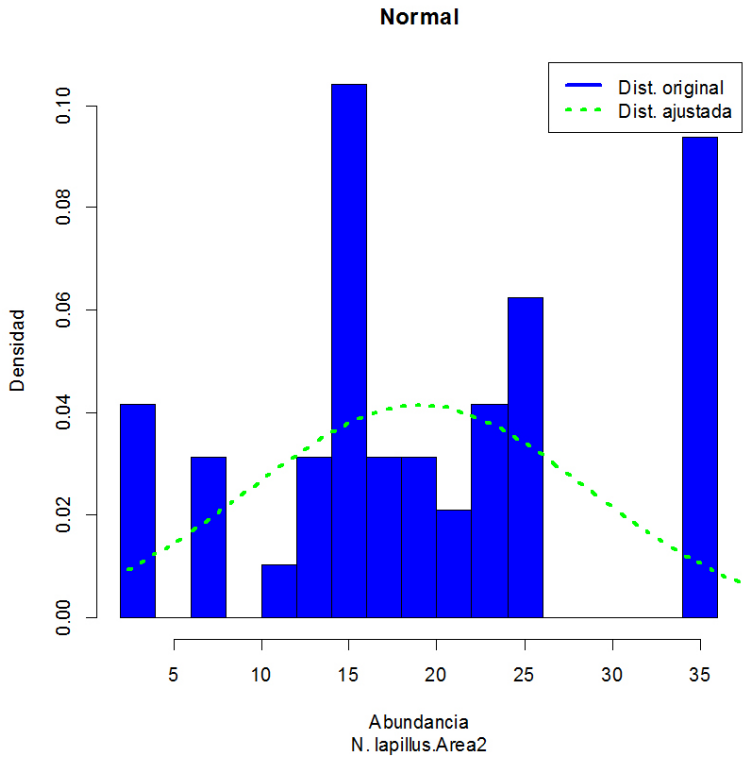



Normal
Dist. original
Dist. ajustada
Densidad
0.0
0.1
0.2
0.3
0
2
4
6
8
10
Abundancia
L. saxatilis.Area1


Normal
Dist. original
Dist. ajustada
Densidad
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
Abundancia
N. lapillus.Area1


Normal
Dist. original
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
Abundancia
L. littorea.Area2


Normal
Dist. original
Dist. ajustada
Densidad
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundancia
L. saxatilis.Area2



Value

Se obtiene un archivo CSV con los resultados de los contrastes y la gráfica con las distribuciones esperada y real si se seleccionó que se mostrara.

Examples

```
## Not run:

data(ZVI1)

#Distribución Normal
VI3(data=ZVI1, varInteres=c("Area1","Area2"), varAgrupacion=c("Species"),
tituloEjeX="Abundancia", col.bar="blue", col.line="green", lty=3, lwd=3,
legend.pos="topright")


## End(Not run)
```

VII1 *CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS-MUESTRAS INDEPENDIENTES Y VARIABLES POLITOMICAS*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, se diferencian en un determinado carácter cualitativo.

Usage

```
VII1(data, varInteres, varAgrupa=NULL, pruebas=c(1,2,3),grafica=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de homogeneidad.
varAgrupa	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
pruebas	TIPO DE CONTRASTE DE HOMOGENEIDAD. 1. Prueba chi-cuadrado. 2. Prueba G sin corrección de Williams. 3. Prueba G con corrección de Williams. 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2). 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral. 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que. 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que. 8. Prueba ANOVA de Friedman. 9. Prueba de McNemar sin corrección de continuidad. 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2). 11. Prueba de Cochran.
grafica	GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.
compress	GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.

xlim	GRÁFICO. Vector con los límites del eje X.
ylim	GRÁFICO. Vector con los límites del eje Y.
spacing	GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.
keep_aspect_ratio	GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.
xscale	GRÁFICO. Ancho de barra de las categorías.
yspace	GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.
main	GRÁFICO. Título principal del gráfico.
sub	GRÁFICO. Subtítulo del gráfico.
residuals_type	GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.
shade	GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.
gp_axis	GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.
grid.edit	GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.
text.grid	GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.
gp.grid	GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.
labeling	GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_left", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.
labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.
legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.

sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

En esta función se realizan los tratamientos estadísticos para los contrastes de homogeneidad que sirven para trabajar con variables cualitativas, que son variables que no se expresan numéricamente y pueden ser dicotómicas (la variable solo puede tomar dos valores, como por ejemplo mujer o hombre) o politómicas (la variable puede tener tres o más valores), existiendo dentro de estas últimas dos tipos: nominales y ordinales.

En un contraste de homogeneidad lo que se pretende es comprobar si dos o más muestras pertenecen a una misma población o que dos muestras pertenecientes a una misma población no hayan sufrido alteraciones. En otras palabras, estamos interesados en conocer si los datos de mediciones de una variable en varios grupos o épocas tienen una misma distribución, o si cada medición es diferente. Podemos distinguir dos tipos de contrastes en función de la naturaleza de las variables, ya que se puede trabajar con mediciones independientes (por ejemplo, dos muestras elegidas en zonas distintas) o pareadas (por ejemplo, podemos medir la misma variable dos veces).

VII.1. MUESTRAS INDEPENDIENTES

VII.1.1. Variable politómicas

Las pruebas son el contraste Chi-cuadrado y la razón de verosimilitud (test G).

VII.1.1.1. Chi-cuadrado

La hipótesis nula de esta prueba supone que las muestras pertenecen a la misma población y, por tanto, la proporción de frecuencias esperadas es la misma en todas las muestras. Para contrastarla se calculan los valores de frecuencia esperados para cada celda, multiplicando las dos frecuencias marginales y dividiendo por el tamaño de muestra, y se comparan con las frecuencias observadas. En caso de que las diferencias sean pequeñas y, en consecuencia, el estadístico de contraste menor que un valor crítico, se acepta la hipótesis de que las muestras son homogéneas (tienen una misma distribución). La fórmula para el cálculo de los valores esperados es:

$$E_{ij} = \frac{n_{mi} n_{cj}}{n}$$

Una vez obtenidos los valores esperados se calcula el estadístico de contraste χ^2 con la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Se busca en la Tabla χ^2 (Tabla 3 del Apéndice I en Guisande et al., 2011) el valor crítico para (*h-1*)*(*k-1*) grados de libertad y para el nivel de significación elegido (generalmente 0,05). Si χ^2 crítico es mayor que χ^2 calculado se acepta la hipótesis nula de que son homogéneas, mientras que

si χ^2 crítico es menor que χ^2 calculado, (o p < nivel de significación), la hipótesis se rechaza. Para la aplicación de esta prueba se requiere que las frecuencias esperadas no tomen valores menores que 5.

VII.1.1.2 Razón de verosimilitud (test G)

Este contraste es muy similar al test χ^2, ya que también cuantifica diferencias entre valores esperados y observados. Además, los grados de libertad se calculan del mismo modo que en el test χ^2 de homogeneidad de muestras y el valor crítico del test se obtiene a partir de la Tabla χ^2 (Tabla 3 del Apéndice I en Guisande et al., 2011).

La tabla de contingencia y el cálculo de los valores esperados son los mismos que los utilizados en el test χ^2 de homogeneidad. La fórmula del estadístico de contraste es:

$$G = 2 \sum_{m=1}^{h} \sum_{c=1}^{k} O_{mc} ln \frac{O_{mc}}{E_{mc}}$$

Del mismo modo que en el contraste anterior, se busca en la Tabla χ^2 (Tabla 3 del Apéndice I en Guisande et al., 2011) el valor crítico con los grados de libertad y el nivel de significación determinados. Si χ^2 crítico es mayor que G se acepta la hipótesis nula de que las muestras son homogéneas.

El estadístico G tiene una distribución aproximadamente χ^2. Existen algunas correcciones que mejoran la aproximación, como la de Williams (1976), aplicable cuando las diferentes categorías tienen frecuencias similares.

FUNCIONES

Se usó la función assoc del paquete vcd (Meyer et al., 2006; 2013). Para más detalles de como usar esta función consultar el ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO 1

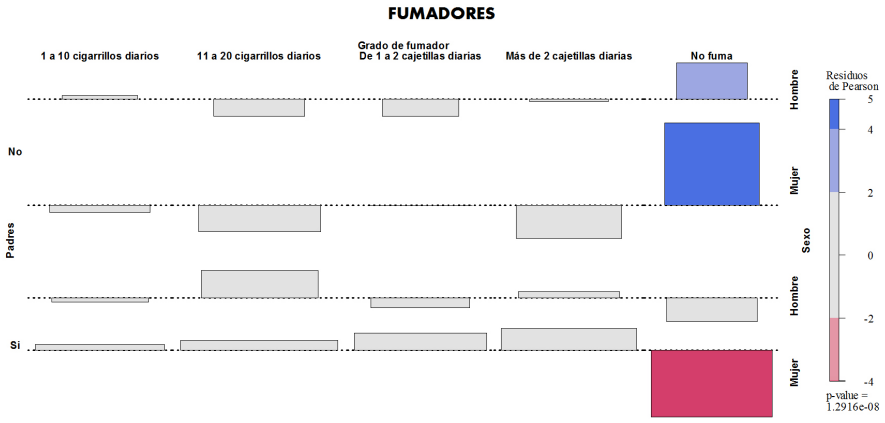
Datos del grado en que fuman hombres y mujeres en distintos centros de trabajo. Las categorías utilizadas fueron: 1 (no fumador), 2 (entre 1 y 10 cigarrillos diarios), 3 (entre 11 y 20 cigarrillos diarios), 4 (entre 1 y 2 cajetillas diarias) y 5 (más de 2 cajetillas diarias). También hay información de si alguno de los padres de estos trabajadores son fumadores y las categorías son: trabajadores en que alguno de sus padres es fumador (categoría con valor 1) y el otro grupo son aquellos en que ninguno de sus padres es fumador (categoría con valor 0). Se quiere determinar, en hombres y en mujeres, si hay diferencias entre el grupo de personas con alguno de los padres fumadores y el grupo en que sus padres no fuman, en la proporción de los distintos tipos de grados de fumador.

Para ello, en el script definimos las variables a comparar con el argumento *varInteres = c("Parents", "Smoker")* y como queremos que el análisis se haga también separando por sexos definimos el argumento *varAgrupa = "Gender"*.

La Figura VII1.1 muestra la representación gráfica de la tabla de contingencia. En este gráfico también se muestra el resultado del test χ^2 con una p < 0,001. Por tanto, se rechaza la hipótesis nula de que las muestras son homogéneas, es decir, existen diferencias significativas entre el grupo con alguno de sus padres fumadores y el grupo sin padres fumadores, en el porcentaje de los distintos tipos de fumadores. El gráfico da también una información adicional muy importante, las categorías donde las diferencias son significativas y que están contribuyendo a que, en su conjunto, el test χ^2 sea significativo. Las categorías de color gris no son significativamente diferentes, mientras que las categorías con color son las significativamente diferentes. En concreto, se observa que en la categoría de padres no fumadores, tanto para hombres como para mujeres, existe una proporción

significativamente mayor de personas que no fuman (las barras están por encima de la línea punteada). Por el contrario, el número de personas que no fuman es significativamente menor en el grupo de padres fumadores(la barra está por debajo de la línea punteada). Por tanto, el hecho de que los padres no fumen parece propiciar que sus hijos no fumen. Sin embargo, una vez que la persona fuma, el grado en que fuman no varía en función de si los padres fuman o no fuman, ya que se observa que en todas las categorías de fumadores el color de las barras es gris.

Figura VII1.1 Representación gráfica de la tabla de contingencia de las variables sexo, si los padres fuman o no fuman y el grado de fumador.



```
$Hombre
$Hombre$tabla
      Grado
Padres 1 a 10 cigarrillos diarios 11 a 20 cigarrillos diarios
    No                         14                          15
    Si                         21                          42
      Grado
Padres De 1 a 2 cajetillas diarias Más de 2 cajetillas diarias No fuma
    No                          10                          14      19
    Si                          20                          26      13

$Hombre$listaPruebas
$Hombre$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 10.0273, df = 4, p-value = 0.03997


$Hombre$listaPruebas[[2]]

     Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 9.8787, X-squared df = 4, p-value
= 0.04252


$Hombre$listaPruebas[[3]]

     Log likelihood ratio (G-test) test of independence with Williams'
     correction

data:  tbl1
Log likelihood ratio statistic (G) = 9.7053, X-squared df = 4, p-value
= 0.0457
```

```
$Mujer
$Mujer$tabla
      Grado
Padres 1 a 10 cigarrillos diarios 11 a 20 cigarrillos diarios
    No                         22                          26
    Si                         42                          63
      Grado
Padres De 1 a 2 cajetillas diarias Más de 2 cajetillas diarias No fuma
    No                          25                          16      44
    Si                          48                          52      11

$Mujer$listaPruebas
$Mujer$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 50.8757, df = 4, p-value = 2.37e-10


$Mujer$listaPruebas[[2]]

     Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 50.9257, X-squared df = 4, p-value
= 2.313e-10


$Mujer$listaPruebas[[3]]

     Log likelihood ratio (G-test) test of independence with Williams'
     correction

data:  tbl1
Log likelihood ratio statistic (G) = 50.4451, X-squared df = 4, p-value
= 2.915e-10
```

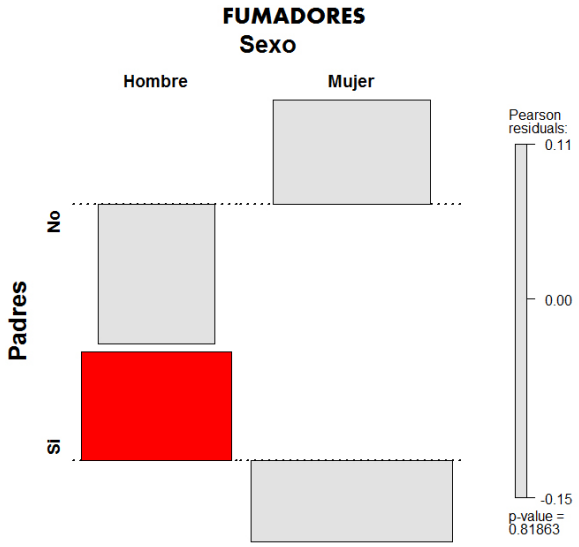
En el script se definió el argumento *varAgrupa="Gender"* y, por ello, se realizaron los test para hombres y mujeres por separado. En el caso de los hombres, tanto el test χ^2 (p = 0,039) como el test de la *G* con la corrección de Williams (p = 0,0457) mostraron diferencias significativas. Por tanto, existen diferencias significativas entre el grupo con alguno de sus padres fumadores y el grupo sin padres fumadores, en el porcentaje de los distintos tipos de fumadores en los hombres, que como se mostró en la Figura VII1.1, se debían a un mayor número de hombres no fumadores en el grupo de padres no fumadores.

En el caso de las mujeres, las diferencias fueron incluso más significativas, tanto en el test χ^2 (p < 0,001) como en el test de la *G* con la corrección de Williams (p < 0,001). Al igual que en los hombres, la Figura VII1.1 mostró que estas diferencias se debían a un mayor número de mujeres no fumadoras, en el grupo de padres no fumadores.

EJEMPLO 2

El objetivo en este segundo ejemplo, usando los mismos datos del ejemplo anterior, es comparar si la proporción de hombres y mujeres muestreada en el estudio fue diferente en el grupo de padres fumadores en comparación con el de no fumadores.

Figura VII1.2 Representación gráfica de la tabla de contingencia de las variables sexo y si los padres fuman o no fuman.



La Figura VII1.2 muestra que no hay diferencias significativas (χ^2, p = 0,818). Por tanto, la proporción hombres y mujeres era igual en el grupo de padres fumadores y no fumadores. En el ejemplo se muestra también como resaltar con un color diferente un categoría con el argumento *text.grid="rect:Parents=Si,Gender=Hombre"* (Figura VII1.2).

Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Meyer, D., Zeileis, A. & Hornik, K. (2013) Visualizing Categorical Data. R package version 1.3-1. Available at: http://CRAN.R-project.org/package=vcd.

Williams, D.A. (1976) Improved llikelihood ratio test for complete contingency tables. *Biometrika*, 63: 33-37.

Examples

```
## Not run:

data(ZVII1)
```

```
#Diferencias en el grado de fumador en función de si
#los padres son fumadores y separando por sexos
VII1(data=ZVII1, varInteres=c("Parents","Smoker"), varAgrupa="Gender",
pruebas=c(1,2,3), labeling_args= list(set_varnames=c(Parents="Padres",
Gender="Sexo", Smoker="Grado de fumador"), offset_varnames = c(left =-0.5,
top=-0.5), offset_labels=c(top=-0.3), rot_labels=c(left=0)),
main="FUMADORES", main_gp=gpar(fontsize=20, fontface=2,
fontfamily="Aharoni"), gp_varnames = gpar(fontsize=12,fontface=2),
gp_labels=gpar(fontsize = 12, fontface = 2),
legend=legend_resbased(fontfamily= "Times New Roman", fontsize=12.9,
text="Residuos\nde Pearson"), legend_width=unit(6, "lines"))

#Se muestra como resaltar una categoría
VII1(data=ZVII1, varInteres=c("Parents","Gender"), pruebas=c(1,2),
grid.edit=TRUE, text.grid="rect:Parents=Si,Gender=Hombre",
gp.grid = gpar(fill = "red"),
labeling_args= list(set_varnames=c(Parents="Padres", Gender="Sexo")),
margins=unit(c(4,2,2,4), "lines"), main="FUMADORES", main_gp=gpar(fontsize=20,
fontface=2, fontfamily="Aharoni"), gp_varnames = gpar(fontsize=20,fontface=2),
gp_labels=gpar(fontsize = 14, fontface = 2))


## End(Not run)
```

VII2 *CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS-MUESTRAS INDEPENDIENTES Y VARIABLES DICOTOMICAS*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, se diferencian en un determinado carácter cualitativo.

Usage

```
VII2(data, varInteres, varAgrupa=NULL, pruebas=c(1,4,5,6,7),grafica=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`varInteres`	Variable o variables a las que se quiere realizar el contraste de homogeneidad.
`varAgrupa`	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
`pruebas`	TIPO DE CONTRASTE DE HOMOGENEIDAD. 1. Prueba chi-cuadrado. 2. Prueba G sin corrección de Williams. 3. Prueba G con corrección de Williams. 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2). 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral. 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que. 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que. 8. Prueba ANOVA de Friedman. 9. Prueba de McNemar sin corrección de continuidad. 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2). 11. Prueba de Cochran.
`grafica`	GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.
`compress`	GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.
`xlim`	GRÁFICO. Vector con los límites del eje X.
`ylim`	GRÁFICO. Vector con los límites del eje Y.
`spacing`	GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.
`keep_aspect_ratio`	GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.
`xscale`	GRÁFICO. Ancho de barra de las categorías.
`yspace`	GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.
`main`	GRÁFICO. Título principal del gráfico.
`sub`	GRÁFICO. Subtítulo del gráfico.
`residuals_type`	GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.

shade	GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.
gp_axis	GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.
grid.edit	GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.
text.grid	GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.
gp.grid	GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.
labeling	GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_left", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.
labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.
legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

VII.1. MUESTRAS INDEPENDIENTES

VII.1.2. Variable dicotómicas

Cuando ambas variables son dicotómicas, es decir tienen dos únicas categorías como, por ejemplo, presencia/ausencia o mujer/hombre, entonces tenemos una tabla de contingencia de 2x2, y se usa una corrección del test χ^2, la corrección de Yates (1934) o el test de Fisher (1922).

VII.1.2.1. Corrección de Yates

La corrección de Yates se aplica cuando una tabla de contingencia es 2x2, es decir, dos filas y dos columnas, y la muestra tiene un número pequeño de casos ya que no es posible agrupar categorías.

En esta situación, el valor de χ^2 se sobreestima y, para corregirlo, se aplica la denominada corrección de Yates que modifica el cálculo del estadístico de la siguiente manera:

$$\chi^2 = \sum_{m=1}^{j} \sum_{c=1}^{i} \frac{|O_{mc} - E_{mcj}| - \frac{1}{2})^2}{E_{mc}}$$

En este caso el valor crítico de contraste sería el χ^2 para un grado de libertad. Si el valor obtenido es menor que el valor tabulado de contraste, se acepta la hipótesis nula de que las muestras son homogéneas.

No obstante, conviene mencionar que la utilización de la corrección de Yates no exime de ciertos requerimientos acerca del tamaño muestral necesario para la utilización del estadístico χ^2. Como norma general, se exigirá que el 80% de las celdas en una tabla de contingencia tengan valores esperados mayores que 5. Así, en una tabla 2x2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas muestre frecuencias esperadas ligeramente por debajo de este valor.

VII.1.2.2. Prueba de Fisher

La prueba de Fisher se utiliza con tablas de contingencia 2x2 y es la más aconsejable cuando exista una o más celdas con frecuencias esperadas menores que 5.

La prueba se basa en la determinación, a partir de una distribución hipergeométrica, de la probabilidad de obtener las frecuencias observadas cuando los totales de filas y columnas son constantes, teniendo como hipótesis nula que las variables son independientes.

FUNCIONES

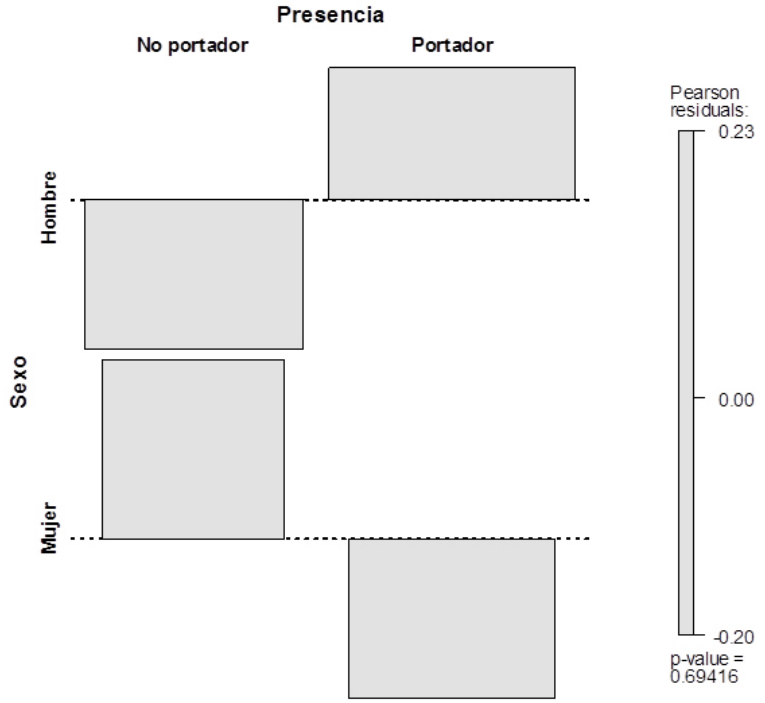
Se usó la función assoc para realizar el gráfico (Meyer et al., 2006; 2013). Para más detalles de como usar esta función consultar el ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO

Datos de la presencia del virus del herpes labial (0 es no portador y 1 es portador) en hombres y mujeres (1 es hombre y 2 es mujer). Se quiere determinar si el número de portadores del virus es diferente entre hombres y mujeres. Para ello, en el argumento *varInteres = c("Sexo", "Presencia")* se especifican las variables a comparar.

La Figura VII2.1 muestra que no hay diferencias significativas con p = 0,694. Por tanto, la presencia del virus del herpes labial es igual en hombres y mujeres.

Figura VII2.1 Representación gráfica de la tabla de contingencia de las variables sexo y presencia del virus del herpes labial.



Los resultados de todos los test se muestran a continuación:

Chi-Cuadrado

El valor χ^2 es 0,15 con p = 0,694. Por lo tanto, concluimos que no existe relación entre el sexo y la presencia del virus del herpes labial.

Prueba exacta de Fisher.

Nos devuelve el estadístico calculado tanto con una hipótesis alternativa unilateral (que informa de si es mayor el número de mujeres portadoras del virus que el número de hombres portadores) como con una hipótesis alternativa bilateral (que simplemente informa si la presencia del virus es distinta). Las probabilidades serían p = 0,425 unilateral y p = 0,838 para bilateral. En ambos casos se acepta la hipótesis de homogeneidad de variables, aunque con nuestros datos (no suponemos a priori si serán más mujeres con virus que hombres), debemos emplear la probabilidad bilateral.

```
$listaPruebas[[3]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.8379
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3541774 2.0532746
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[4]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.4248
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.804971
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[5]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.7253
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.4028887       Inf
sample estimates:
odds ratio
 0.8527715
```

Corrección de Yates.

En este caso dado, como las frecuencias de cada categoría son mayores de 5, no sería necesaria esta corrección. Suponiendo que lo fuese, tendríamos un valor de contraste de 0,035 y p = 0,851, por lo que se acepta la hipótesis de que la presencia del virus es igual en ambos sexos.

```
        Presencia
Sexo     No portador Portador
  Hombre          25       34
  Mujer           19       22

$listaPruebas
$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 0.1546, df = 1, p-value = 0.6942


$listaPruebas[[2]]

     Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 0.0355, df = 1, p-value = 0.8505
```

Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

Fisher, L.D. & van Belle, G. (1993) *Biostatistics. A methodology for the health sciences*. John Wiley y Sons, New York.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Meyer, D., Zeileis, A. & Hornik, K. (2013) Visualizing Categorical Data. R package version 1.3-1. Available at: http://CRAN.R-project.org/package=vcd.

Yates, F. (1934) Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1: 217-235.

Examples

```
## Not run:

data(ZVII2)

#Presencia del virus del herpes
VII2(data = ZVII2 , varInteres = c("Sexo","Presencia") , gp_varnames = gpar(fontsize=15,
fontface=2), gp_labels=gpar(fontsize = 13, fontface = 2))


## End(Not run)
```

VII3 *CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS-MUESTRAS RELACIONADAS Y VARIABLES POLITOMICAS*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, se diferencian en un determinado carácter cualitativo.

Usage

```
VII3(data, varInteres, varAgrupa=NULL, pruebas=c(8),grafica=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de homogeneidad.
varAgrupa	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas. En el caso del ANOVA de Friedman es necesario ponerlas en el siguiente orden: la variable que tiene los datos de la respuesta, el factor y, por último, la variable que tiene la información sobre los individuos a los que pertenece cada respuesta.
pruebas	TIPO DE CONTRASTE DE HOMOGENEIDAD. 1. Prueba chi-cuadrado. 2. Prueba G sin corrección de Williams. 3. Prueba G con corrección de Williams. 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2). 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral. 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que. 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que. 8. Prueba ANOVA de Friedman.

9. Prueba de McNemar sin corrección de continuidad.

10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2).

11. Prueba de Cochran.

`grafica` GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.

`compress` GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.

`xlim` GRÁFICO. Vector con los límites del eje X.

`ylim` GRÁFICO. Vector con los límites del eje Y.

`spacing` GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.

`keep_aspect_ratio` GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.

`xscale` GRÁFICO. Ancho de barra de las categorías.

`yspace` GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.

`main` GRÁFICO. Título principal del gráfico.

`sub` GRÁFICO. Subtítulo del gráfico.

`residuals_type` GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.

`shade` GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.

`gp_axis` GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.

`grid.edit` GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.

`text.grid` GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.

`gp.grid` GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.

`labeling` GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_left", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.

labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.
legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

VII.2. MUESTRAS RELACIONADAS

Se puede dar el caso de que las variables a comparar estén relacionadas, es decir, son datos pareados. Por ejemplo, cuando a un grupo de personas se les aplica un tratamiento y se estudia si ese fármaco provoca algún tipo de alergia, cuantificada esta alergia por medio de alguna variable cualitativa como presencia o ausencia de manchas en la piel. En este caso, las variables a comparar no son independientes, ya que son las mismas personas las que se comparan a lo largo del tiempo.

Al igual que en el caso de las muestras independientes, existen pruebas para variables politómicas y dicotómicas.

VII.2.1. Variable politómicas

VII.2.1.1. ANOVA de Friedman

El test desarrollado por Friedman (1937) es la prueba análoga al análisis de varianza (ANOVA) para muestras repetidas por lo que, a pesar de que se trata de un contraste χ^2, recibe el nombre de ANOVA.

Los pasos de este estadístico están descritos por Siegel & Castellán (1988). De forma resumida son:

1. Asignar rangos dentro de cada elemento de la muestra.

2. Sumar los rangos de cada elemento de la muestra.

3. Calcular el estadístico χ^2 de contraste y compararlo con el valor tabulado χ^2 crítico con *a* - 1 grados de libertad, siendo *a* el número de variables.

EJEMPLO

Datos sobre como evalúan 4 diferentes tipos de vinos 15 enólogos. La categoría 1 es Muy malo, 2 Malo, 3 Bueno y 4 Muy bueno. El objetivo es determinar si existen diferencias entre la calidad de los vinos. Como se compara la evaluación de cada uno de los enólogos de los distintos vinos, los datos no son independientes, están relacionados. Además cada una de las variables tiene más de 2 valores, por tanto son politómicas.

En el caso del ANOVA de Friedman es necesario ponerlas en el siguiente orden: la variable que tiene los datos de la respuesta, el factor y, por último, la variable que tiene la información sobre los individuos a los que pertenece cada respuesta, en este caso sería *varInteres=c("Quality","Wine","Oenologist")*.

Como la probabilidad es menor de 0,05, se rechaza la hipótesis nula de homogeneidad y, por tanto, hay diferencias significativas entre la calidad de los vinos según los enólogos (ANOVA χ^2 = 10,1, df = 3, p = 0,018).

```
           medias  desv.est
Vino 1 3.0666667 0.9611501
Vino 2 1.7333333 0.7037316
Vino 3 2.2000000 0.8618916
Vino 4 2.8666667 1.3020131

$listaPruebas
$listaPruebas[[1]]

     Friedman rank sum test

data:  varInt, factor(varFact) and factor(varBloque)
Friedman chi-squared = 10.1172, df = 3, p-value = 0.0176
```

Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32: 675-701.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZVII3)

#Cata de vinos
VII3(data=ZVII3, varInteres=c("Quality","Wine","Oenologist"),
grafica=FALSE)


## End(Not run)
```

VII4 *CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS-MUESTRAS RELACIONADAS Y VARIABLES DICOTOMICAS SIN REPETICION*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, se diferencian en un determinado carácter cualitativo.

Usage

```
VII4(data, varInteres, varAgrupa=NULL, pruebas=c(9,10),grafica=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de homogeneidad.
varAgrupa	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
pruebas	TIPO DE CONTRASTE DE HOMOGENEIDAD. 1. Prueba chi-cuadrado. 2. Prueba G sin corrección de Williams. 3. Prueba G con corrección de Williams. 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2). 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral. 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que. 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que. 8. Prueba ANOVA de Friedman. 9. Prueba de McNemar sin corrección de continuidad. 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2). 11. Prueba de Cochran.

`grafica` GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.

`compress` GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.

`xlim` GRÁFICO. Vector con los límites del eje X.

`ylim` GRÁFICO. Vector con los límites del eje Y.

`spacing` GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.

`keep_aspect_ratio` GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.

`xscale` GRÁFICO. Ancho de barra de las categorías.

`yspace` GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.

`main` GRÁFICO. Título principal del gráfico.

`sub` GRÁFICO. Subtítulo del gráfico.

`residuals_type` GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.

`shade` GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.

`gp_axis` GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.

`grid.edit` GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.

`text.grid` GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.

`gp.grid` GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.

`labeling` GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_left", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.

`labeling_args` GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.

`margins` GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.

legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

VII.2. MUESTRAS RELACIONADAS

VII.2.2. Variable dictómicas sin repetición

VII.2.2.1. Prueba de McNemar

Se utiliza con muestras a las que se les mide la misma variable dicotómica dos veces, es decir, son datos pareados (McNemar, 1947). Es muy utilizado en medicina para conocer si un tratamiento tiene efecto, pudiéndose representar los datos en una tabla de contingencia 2x2.

Se realiza el siguiente contraste, en el que se asume como hipótesis nula que las frecuencias esperadas b y c son iguales:

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

El valor crítico de contraste se corresponde con la distribución χ^2 para 1 grado de libertad.

EJEMPLO

El objetivo del estudio es determinar si un antibiótico provoca una respuesta alérgica en la piel. Para ello se administró el antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas entre los días 5 y 10 después del tratamiento. Lo que se contrasta es si hay diferencias en cada uno de los grupos, es decir si son homogéneos en cuanto a la respuesta alérgica, en los dos períodos considerados, 5 y 10 días. Si suponemos que 5 días es antes y 10 días después y la alergia aparece necesariamente en ese intervalo, el efecto alérgico del antibiótico puede probarse si encontramos homogeneidad en el grupo placebo y no homogeneidad en el grupo al que se le suministró antibiótico, aunque en este último caso debemos comprobar adicionalmente que la alergia es mayor después que antes y no al revés (en ambos casos el contraste puede ser significativo).

```
$NO
$NO$tabla
          Dia.10
Dia.5      Negativa Positiva
  Negativa       43        3
  Positiva        0        4

$NO$listaPruebas
$NO$listaPruebas[[1]]

     McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 3, df = 1, p-value = 0.08326


$NO$listaPruebas[[2]]

     McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 1.3333, df = 1, p-value = 0.2482
```

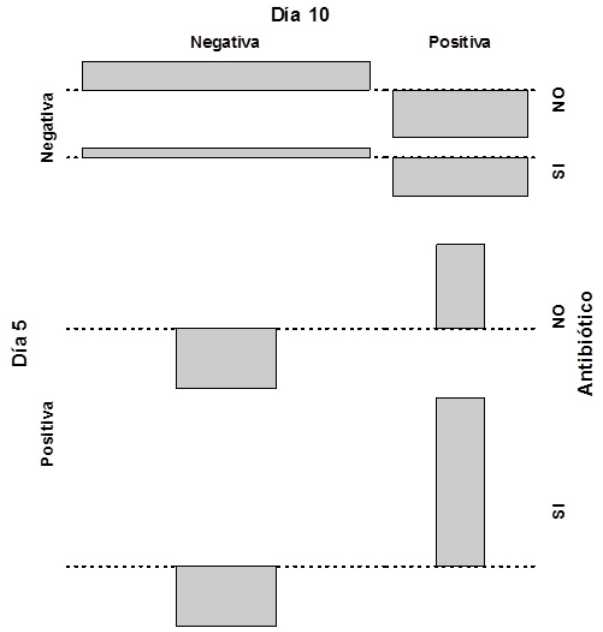
En el caso de las personas a las que no se les suministró el antibiótico, como p= 0,248 en el contraste con la corrección de continuidad (que es el que debe usarse cuando hay frecuencias menores que 5), se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica observada en la piel entre el día 5 y el día 10.

En el caso de las personas a las que se les suministró el antibiótico como p = 0,133, al ser mayor de 0,05, también se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica al antibiótico entre el día 5 y día 10.

```
$SI
$SI$tabla
          Dia.10
Dia.5      Negativa Positiva
  Negativa       39        4
  Positiva        0        7

$SI$listaPruebas
$SI$listaPruebas[[1]]

     McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 4, df = 1, p-value = 0.0455


$SI$listaPruebas[[2]]

     McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 2.25, df = 1, p-value = 0.1336
```

La Figura VII4.1 muestra los resultados. No se muestra la escala con la significación porque no se utiliza el test χ^2.

Figura VII4.1 Tabla de contingencia de las variables antibiótico y los días 5 y 10.



Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

McNema,r Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12: 153-157.

Examples

```
## Not run:

data(ZVII4)

VII4(data=ZVII4, varInteres=c("Day.5","Day.10"), varAgrupa="Antibiotic",
shade=FALSE,  labeling_args =list(set_varnames=
c(Day.5="Día 5", Day.10="Día 10", Antibiotic="Antibiótico")),
gp_varnames = gpar(fontsize=15,fontface=2),
gp_labels=gpar(fontsize = 13, fontface = 2))


## End(Not run)
```

VII5 *CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS-MUESTRAS RELACIONADAS Y VARIABLES DICOTOMICAS CON REPETICION*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, se diferencian en un determinado carácter cualitativo.

Usage

```
VII5(data, varInteres, varAgrupa=NULL, pruebas=c(11),grafica=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variable o variables a las que se quiere realizar el contraste de homogeneidad.
varAgrupa	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
pruebas	TIPO DE CONTRASTE DE HOMOGENEIDAD. 1. Prueba chi-cuadrado. 2. Prueba G sin corrección de Williams. 3. Prueba G con corrección de Williams. 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2). 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral. 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que. 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que. 8. Prueba ANOVA de Friedman. 9. Prueba de McNemar sin corrección de continuidad. 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2). 11. Prueba de Cochran.

grafica	GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.
compress	GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.
xlim	GRÁFICO. Vector con los límites del eje X.
ylim	GRÁFICO. Vector con los límites del eje Y.
spacing	GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.
keep_aspect_ratio	GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.
xscale	GRÁFICO. Ancho de barra de las categorías.
yspace	GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.
main	GRÁFICO. Título principal del gráfico.
sub	GRÁFICO. Subtítulo del gráfico.
residuals_type	GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.
shade	GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.
gp_axis	GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.
grid.edit	GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.
text.grid	GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.
gp.grid	GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.
labeling	GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_left", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.
labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.

legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

VII.2. MUESTRAS RELACIONADAS

VII.2.3. Variable dictómicas con repetición

VII.2.3.1. Q de Cochran

Es el equivalente a la prueba de McNemar cuando las mediciones se realizan más de dos veces. Para el cálculo del estadístico de contraste, los datos se presentan en una tabla de contingencia. Los valores de la tabla de contingencia deben ser transformados a 0 y 1. Por ejemplo, si nuestra variable fuese presencia/ausencia de un carácter, podríamos indicar presencia con 1 y ausencia con 0. El estadístico es el siguiente:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} c_j^2 - \left(\sum_{j=1}^{k} c_j\right)^2\right]}{k\sum_{j=1}^{k} c_j - \sum_{i=1}^{N} n_i^2}$$

Siendo *k* el número de variables dicotómicas y *N* el número de casos, c_j el número de éxitos para cada variable y n_i el número de éxitos para cada caso.

EJEMPLO

El objetivo del estudio es determinar si un antibiótico provoca una respuesta alérgica en la piel. Para ello se administró el antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas en varios días. Lo que se contrasta es si hay diferencias en cada uno de los grupos, es decir si son homogéneos en cuanto a la respuesta alérgica, durante el periodo considerado. El efecto alérgico del antibiótico puede probarse si encontramos homogeneidad en el grupo placebo y no homogeneidad en el grupo al que se le suministró antibiótico, aunque en este último caso debemos comprobar adicionalmente que la alergia es mayor después que antes y no al revés (en ambos casos el contraste puede ser significativo).

En el caso de las personas a las que no se les suministró el tratamiento, la probabilidad es menor que 0,05 por lo que rechazamos la hipótesis nula de que las muestras son iguales, es decir, hay diferencias entre los distintos días en los síntomas de alergia detectados en las personas sometidas al placebo (Test *Q* de Cochran = 9, df = 3, p = 0,03).

```
$NO
$NO$tabla
       Porcentaje Negativa Porcentaje Positiva
datos1                  92                   8
datos2                  86                  14
datos3                  86                  14
datos4                  84                  16

$NO$listaPruebas
$NO$listaPruebas[[1]]

     Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 9, df = 3, p-value = 0.02929
```

En el caso de las personas a las que se les suministró antibiótico, también se rechaza la hipótesis nula de que las muestras son iguales, es decir, hay diferencias entre los distintos días en los síntomas de alergia detectados (Test *Q* de Cochran = 50,4, df = 3, p < 0,001).

```
$SI
$SI$tabla
       Porcentaje Negativa Porcentaje Positiva
datos1                  86                  14
datos2                  78                  22
datos3                  62                  38
datos4                  38                  62

$SI$listaPruebas
$SI$listaPruebas[[1]]

     Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 50.4, df = 3, p-value = 6.566e-11
```

Es importante resaltar que, el hecho de que se observaran diferencias en el tiempo en ambos grupos (con y sin antibiótico) no significa concluir definitivamente que el antibiótico no es el responsable de la alergia. Esta prueba solo nos permite mostrar que hay diferencias en la respuesta alérgica en el tiempo. Es decir, se podría dar el caso de que en las personas a las que se les suministró placebo, la respuesta alérgica disminuyó en el tiempo y, por el contrario, en las que recibieron antibiótico aumentó. Por tanto, sería necesario a mayores realizar un contraste de independencia como se explica en la función VIII1, para poder descartar la asociación entre el antibiótico y la alergia.

Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

Examples

```
## Not run:

data(ZVII5)

#Alergia en la piel a un antibiótico en varios días
VII5(data=ZVII5, varInteres=c("Day.5","Day.10","Day.15","Day.20"),
varAgrupa="Antibiotic", grafica=FALSE)


## End(Not run)
```

VIII1 *CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUALITATIVAS-VARIABLES POLITOMICAS*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, son independientes y, en caso de dependencia, el grado de asociación en variables politómicas.

Usage

```
VIII1(data, varInteres, varAgrupa=NULL, varEstratos=NULL, pruebas=c(1,2,3),
grafica=TRUE, compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variables cuya asociación se quiere determinar.

varAgrupa — Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.

varEstratos — Variable que define los estratos.

pruebas — TIPO DE CONTRASTE DE HOMOGENEIDAD.

1. Prueba chi-cuadrado.
2. Prueba G sin corrección de Williams.
3. Prueba G con corrección de Williams.
4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
12. Medidas de asociación.
13. Coeficiente de incertidumbre.
14. Pruebas de Cochran (sin corrección) (solo tablas 2x2xk).
15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas 2x2xk).
16. Cálculo de las razones de riesgo por estrato (solo 2x2xk).
17. Pruebas Breslow & Day con la corrección de Tarone. (solo 2x2xk).

grafica — GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo se muestra como ordenar las categorías e incluir los acentos.

compress — GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.

xlim — GRÁFICO. vector con los límites del eje X.

ylim — GRÁFICO. Vector con los límites del eje Y.

spacing — GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.

keep_aspect_ratio — GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.

xscale — GRÁFICO. Ancho de barra de las categorías.

yspace — GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.

main — GRÁFICO. Título principal del gráfico.

sub — GRÁFICO. Subtítulo del gráfico.

residuals_type — GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.

shade — GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.

gp_axis — GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.

grid.edit	GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.
text.grid	GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.
gp.grid	GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.
labeling	GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_lef", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.
labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.
legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VIII. CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUALITATIVAS

Uno de los principales propósitos de la estadística es determinar si existe asociación entre dos o más caracteres o variables de una misma muestra. Además de conocer si dos o más variables son independientes o dependientes es interesante conocer, también, cuál es la fuerza de dicha asociación.

De forma general diremos que la fórmula de aproximación para el estudio de la independencia es la medición de las diferencias entre las frecuencias observadas y las frecuencias que esperaríamos encontrar si las variables fuesen independientes.

VIII.1. VARIABLES POLITÓMICAS

VIII.1.1. Pruebas de independencia

VIII.1.1.1. Chi-cuadrado de Pearson y test G de razón de verosimilitud

Son los contrastes más utilizados, aunque no suelen emplearse con variables dicotómicas, ya que hay pruebas específicas para ellas.

Operativamente el cálculo de ambos estadísticos es igual al contraste de homogeneidad de muestras (véase los apartados VII.1.1.1 y VII.1.1.2 en la sección *details* de la función VII1), con la salvedad de que en este caso filas y columnas se refieren a variables diferentes de una misma muestra.

VIII.1.2. Medidas de asociación

La medición de la intensidad de esta asociación puede estar basada en el cálculo de distancias entre variables o en la cantidad de variabilidad explicada si calculamos las frecuencias de una variable en función de la otra variable.

VIII.1.2.1. V de Cramer

El coeficiente *V* de Cramer es el siguiente:

$$V = \sqrt{\frac{\chi^2}{n(min(f,c)-1)}}$$

siendo f y c el número de categorías en filas y columnas respectivamente y n el número de casos.

Un valor 0 indica independencia total de las dos variables, mientras que un valor 1 indica asociación perfecta. Tiene la ventaja de que permite analizar tablas de contingencia de cualquier tamaño.

VIII.1.2.2. Coeficiente de Contingencia

También es una medida asociada a χ^2. La fórmula es:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

donde n es el número de elementos de la muestra.

Este coeficiente toma valores entre 0 y 1. Un valor próximo a cero indica independencia, y próximo a uno fuerte dependencia. Nunca puede tomar el valor límite 1, incluso cuando la dependencia es total.

VIII.1.2.3. Coeficiente de Incertidumbre

Es una medida de asociación basada en la reducción proporcional del error, con un rango entre 0 y 1, de las mediciones de una variable basándose en el conocimiento de la otra. Es decir, nos indica el porcentaje de la variabilidad explicado por la asociación de variables.

Existen dos métodos distintos para realizar esta prueba:

1. Considerando una variable independiente y otra dependiente. En este caso se pueden calcular dos estadísticos distintos ya que podemos considerar como variable dependiente tanto las filas como las columnas. Las medidas de asociación son, suponiendo la variable Y en columnas y X en filas:

Columnas variable dependiente

$$I_{y|x} = \frac{(I_x + I_y - I_{xy})}{I_y}$$

Filas variable dependiente

$$I_{x|y} = \frac{(I_x + I_y - I_{xy})}{I_x}$$

donde

$$I_y = -\sum_{j=1}^{k} \frac{n_j}{n} ln \frac{n_j}{n}$$

$$I_x = -\sum_{i=1}^{h} \frac{n_i}{n} ln \frac{n_i}{n}$$

$$I_{xy} = -\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{n_{ij}}{n} ln \frac{n_{ij}}{n}$$

siendo *h* y *k* el número de filas y columnas, *n* el tamaño de la muestra, n_i y n_j las frecuencias marginales de fila y de columna, y n_{ij} las frecuencias conjuntas.

2. Sin considerar dependencia entre variables

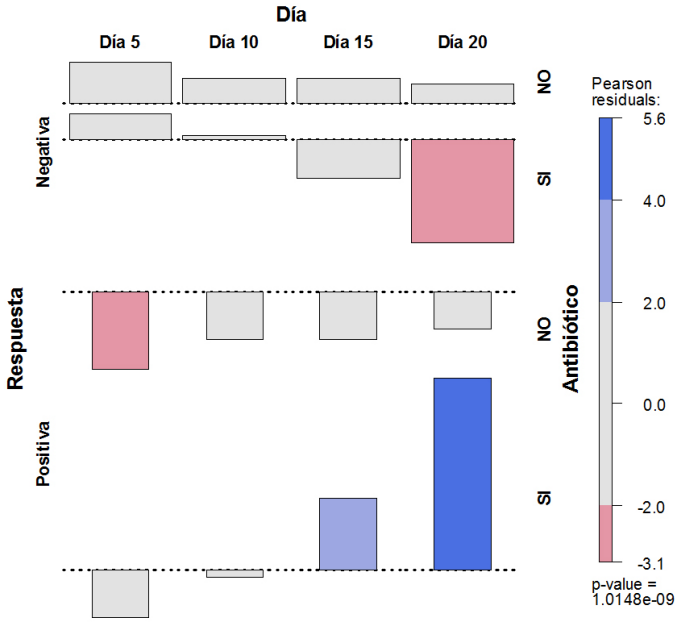
$$I_{ind} = 2\frac{(I_x + I_y - I_{xy})}{I_x + I_y}$$

EJEMPLO

En el ejemplo de la función VII5 se realizó un estudio para determinar si un antibiótico provoca una respuesta alérgica en la piel. El contraste de homogeneidad mostró que, tanto en el caso de las personas tratadas con el antibiótico como a las que no se les suministró, existían diferencias significativas en los síntomas de alergia detectados en la piel a lo largo del periodo de estudio. Ahora se quiere determinar si efectivamente existe una relación entre el tiempo y los síntomas de alergia. La variable respuesta al antibiótico es dicotómica (respuesta positiva o negativa), pero la variable tiempo no es dicotómica ya que hay datos de respuesta a los días 5, 10, 15 y 20 del estudio.

Este grafico no permite acentos y ordenada las categorías por orden alfabético. Los nombres de las columnas se ordenan con algún tipo de código en el archivo de datos de entrada, para que luego salgan estas categorías en el gráfico en el orden deseado. Para modificar las etiquetas, con el fin de incluir los acentos, se usa el argumento *labeling_args* y luego se especifica *set_varnames* para las leyendas y *set_labels* para las etiquetas de las categorías, como se muestra en el script.

Figura VIII1 Tabla de contingencia de las variables antibiótico y respuesta, a lo largo del tiempo.


Día
Día 5
Día 10
Día 15
Día 20
Respuesta
Negativa
Positiva
Antibiótico
NO
SI
NO
SI
Pearson residuals:
5.6
4.0
2.0
0.0
-2.0
-3.1
p-value = 1.0148e-09

```
$NO
$NO$tabla
          Dia
Respuesta  I.Dia.5 II.Dia.10 III.Dia.15 IV.Dia.20
  Negativa      46        43         43        42
  Positiva       4         7          7         8

$NO$listaPruebas
$NO$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 1.5915, df = 3, p-value = 0.6613


$NO$listaPruebas[[2]]
                     X^2 df P(> X^2)
Likelihood Ratio 1.7181  3  0.63293
Pearson          1.5915  3  0.66132

Phi-Coefficient   : NA
Contingency Coeff.: 0.089
Cramer's V        : 0.089

$NO$listaPruebas[[3]]
$Iydx
[1] 0.003098296

$Ixdy
[1] 0.0111162

$Iind
[1] 0.004845938
```

El análisis conjunto de los datos muestra que existen diferencias significativas en la respuesta alérgica en el tiempo en función de si se suministraba el antibiótico (Figura VIII1, $p < 0{,}001$). Estas diferencias eran significativas principalmente en el día 20 donde, a las personas que se les suministraba el antibiótico, las respuestas positivas eran significativamente mayores y las repuestas negativas eran significativamente menores (Figura VIII1).

El análisis por separado muestra que en el grupo al que no se suministró antibiótico, cuyos resultados se muestran a continuación, se acepta la hipótesis nula de independencia ($\chi^2_3 = 1{,}59$, $p = 0{,}661$). Por tanto, a pesar de que el contraste de homogeneidad mostró diferencias significativas en la respuesta alérgica observada en la piel con el tiempo, estas diferencias observadas no son suficientes para establecer una dependencia significativa entre el tiempo y la respuesta alérgica en aquellos pacientes a los que se les suministró un placebo en vez del antibiótico.

```
$SI
$SI$tabla
           Dia
Respuesta  I.Dia.5 II.Dia.10 III.Dia.15 IV.Dia.20
  Negativa      43        39         31        19
  Positiva       7        11         19        31

$SI$listaPruebas
$SI$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 29.9465, df = 3, p-value = 1.416e-06


$SI$listaPruebas[[2]]
                     X^2 df   P(> X^2)
Likelihood Ratio 30.414  3 1.1291e-06
Pearson          29.947  3 1.4163e-06

Phi-Coefficient   : 0.387
Contingency Coeff.: 0.361
Cramer's V        : 0.387

$SI$listaPruebas[[3]]
$Iydx
[1] 0.05484806

$Ixdy
[1] 0.1186136

$Iind
[1] 0.07501054
```

Por el contrario, en el grupo de personas que recibieron antibiótico, se rechaza la hipótesis nula de independencia (χ^2_3 = 29,9, p < 0,001), es decir, existe una relación significativa entre el tiempo y la respuesta alérgica observada en la piel.

Entre los coeficientes de asociación consideramos el *V* de Cramer, el coeficiente de incertidumbre, y el coeficiente de contingencia. No consideramos *phi* ya que la tabla no es 2x2.

El coeficiente *V* de Cramer es 0,39 y el de contingencia 0,36, lo que indica asociación débil. El coeficiente de incertidumbre y-dependiente es 0,1186, es decir, la variable X (el tiempo) reduce un 11,86% la incertidumbre sobre la predicción de la variable Y (la alergia). Naturalmente consideramos que la variable alergia es dependiente del tiempo, y no al revés.

Con estos resultados podemos concluir que nuestras variables están relacionadas significativamente con valor de 0,39 y, por tanto, como la alergia (variable Y) depende del tiempo (variable X), se puede concluir que el tiempo explica un 12% de la alergia observada en la piel en las personas a las que se les suministró el antibiótico.

Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Examples

```
## Not run:

data(ZVIII1)

#Respuesta alérgica de un antibiótico en el tiempo
VIII1(data=ZVIII1, varInteres=c("Respuesta","Dia"), varAgrupa="Antibiotico",
pruebas=c(1,12,13), labeling_args =list(set_varnames=c(Dia="Día",
Antibiotico="Antibiótico"), set_labels = list(Dia= c("Día 5", "Día 10",
"Día 15", "Día 20"))), gp_varnames = gpar(fontsize=15,fontface=2),
gp_labels=gpar(fontsize = 13, fontface = 2))


## End(Not run)
```

VIII2 *CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUALITATIVAS-VARIABLES DICOTOMICAS*

Description

Se determina si varias muestras, tomadas de la misma población o de poblaciones diferentes, son independientes y, en caso de dependencia, el grado de asociación en variables dicotómicas.

Usage

```
VIII2(data, varInteres, varAgrupa=NULL, varEstratos=NULL, pruebas=c(1,2,3),
grafica=TRUE, compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variables cuya asociación se quiere determinar.
varAgrupa	Variables con las que se agrupan los datos para hacer los cálculos. En el caso de seleccionar NULL, no se realizaría ninguna agrupación y se calcularía considerando todos los datos de las variables seleccionadas.
varEstratos	Variable que define los estratos.

`pruebas` TIPO DE CONTRASTE DE HOMOGENEIDAD.
1. Prueba chi-cuadrado.
2. Prueba G sin corrección de Williams.
3. Prueba G con corrección de Williams.
4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
12. Medidas de asociación.
13. Coeficiente de incertidumbre.
14. Pruebas de Cochran (sin corrección) (solo tablas 2x2xk).
15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas 2x2xk).
16. Cálculo de las razones de riesgo por estrato (solo 2x2xk).
17. Pruebas Breslow & Day con la corrección de Tarone. (solo 2x2xk).

`grafica` GRÁFICO. Si es TRUE se muestra el gráfico de la tabla de contingencia. Este gráfico no permite acentos y ordenada las categorías por orden alfabético. En el ejemplo de la función VIII1 se muestra como ordenar las categorías e incluir los acentos.

`compress` GRÁFICO. Valor lógico, que si es FALSE el espacio entre filas y columnas se elige para que el total de alturas y anchos de las filas y columnas sean iguales. Si es TRUE el espacio entre filas y columnas es fijo y, por tanto, el gráfico está más comprimido.

`xlim` GRÁFICO. vector con los límites del eje X.

`ylim` GRÁFICO. Vector con los límites del eje Y.

`spacing` GRÁFICO. Espacio entre las barras, tanto horizontal como vertical, el cual se puede ampliar cambiando el valor "sp=0" a un número mayor. Ver la función spacings para más opciones.

`keep_aspect_ratio` GRÁFICO. Valor lógico, si es TRUE el alto y ancho del gráfico son iguales.

`xscale` GRÁFICO. Ancho de barra de las categorías.

`yspace` GRÁFICO. Separación vertical entre las barras que se puede expresar en diferentes unidades como "lines", "cm", "mm", "inches", etc. Ver la función unit para más detalles.

`main` GRÁFICO. Título principal del gráfico.

`sub` GRÁFICO. Subtítulo del gráfico.

`residuals_type` GRÁFICO. Especifica el tipo de residuos, aunque por ahora solo permite los residuos de Pearson.

`shade` GRÁFICO. Valor lógico que si es TRUE aparecen los resultados del estadístico Chi cuadrado de Pearson y, además, se sombrean las categorías que son significativamente diferentes.

`gp_axis` GRÁFICO. Define el tipo de línea, grosor, color, etc., de la línea de las barras de las categorías.

`grid.edit` GRÁFICO. Valor lógico que si es TRUE permite resaltar una categoría con un color diferente.

`text.grid` GRÁFICO. Texto que permite identificar las variables y las categorías de las variables que se quiere resaltar, por ejemplo "rect:Padres=Si,Sexo=Hombre". No se usa si text.grid es FALSE. Ver uno de los ejemplos para más detalles.

gp.grid	GRÁFICO. Objeto de la clase gpar que permite definir la transparencia y el color de la categoría que se quiere resaltar.
labeling	GRÁFICO. Permite poner las etiquetas de las categorías a la izquierda con "labeling_lef", un marco con "labeling_cboxed", cambiar las de la izquierda a la derecha y las de arriba abajo y ponerles un marco con "labeling_lboxed", etc. Ver la función labeling_border para más detalles.
labeling_args	GRÁFICO. Permite especificar la distancia a la que se ponen las etiquetas de las categorías, leyendas, etc. Ver uno de los ejemplos para más detalles.
margins	GRÁFICO. Objeto de tipo unit con cuatro componentes que delimitan los márgenes de arriba, derecha, abajo e izquierda en el gráfico. Ver uno de los ejemplos para más detalles.
legend_width	GRÁFICO. Margen de la leyenda al borde derecho.
gp_varnames	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de las leyendas de las variables.
gp_labels	GRÁFICO. Tamaño, formato y tipo de fuente de la letra de cada una de las categorías de las variables.
main_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del título principal.
sub_gp	GRÁFICO. Tamaño, formato y tipo de fuente de la letra del subtítulo.
legend	GRÁFICO. Permite modificar muchos aspectos de la leyenda como el tamaño de letra, tipo de fuente, número de decimales, cambiar el texto, etc.
gp	GRÁFICO. Permite cambiar la escala de colores de la leyenda y tiene estas diferentes paletas: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" y "shading_binary".
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

VIII. CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUALITATIVAS

VIII.2. VARIABLES DICOTÓMICAS

VIII.2.1. Pruebas de independencia

VIII.2.1.1. Correción de Yates y prueba de Fisher

Para contrastar la independencia en variables dicotómicas, cuando solo se tiene una tabla 2x2, se puede usar el test χ^2 con la corrección de Yates, el cual ya se explicó en el apartado VII.1.2.1 de la función VII2.

La prueba de Fisher, la cual se explicó brevemente en el apartado VII.1.2.2 de la función VII2, también se utiliza cuando la tabla de contingencia es 2x2 y debe usarse en vez de la prueba χ^2 cuando en la tabla de contingencia exista alguna celda con frecuencias esperadas menores de 5.

VIII.2.1.2. Pruebas de Cochran y Mantel-Haenszel

Al igual que el test de Fisher, las pruebas de Cochran (1954) y Mantel & Haenszel (1959) también sirven para contrastar la independencia en una tabla de contingencia 2x2. Sin embargo, tienen la ventaja sobre la prueba de Fisher de que también se pueden usar cuando existan varias tablas 2x2. Estas tablas de contingencia son del tipo Kx2x2, es decir, hay tantas tablas 2x2 como K estratos. En

estas tablas, se trata de estudiar si existe o no asociación entre dos variables dicotómicas, cuando se dispone de información referida a varios estratos. Por ejemplo, se quiere determinar si la presencia de cáncer de cuello uterino está relacionada con la presencia del virus del papiloma humano en mujeres de diferentes edades o por el contrario son independientes. Por lo tanto, tendríamos varias tablas 2x2 (presencia o no presencia de cáncer y presencia o no presencia del virus, ambas variables dicotómicas) en varios estratos que serían las diferentes clases de edad en que se divida el grupo de mujeres estudiadas. En este ejemplo, analizar cada tabla 2x2 por separado con la prueba de Fisher no daría información sobre el efecto de la edad. El estadístico de Cochran es el siguiente:

$$\chi^2_{Cochran} = \frac{\left(\sum_k n_k - \sum_k m_k\right)^2}{\sum_k \sigma^2_{n_k}}$$

donde:

k = cada uno de los estratos

n_k = frecuencia observada en cualquiera de las casillas del estrato *k* (solo una y siempre es la misma en todos los estratos)

m_k = frecuencia esperada correspondiente a n_k

$$\sigma^2_{n_k} = \frac{n_{f1k} n_{f2k} n_{c1k} n_{c2k}}{n^3}$$

n_{f1k}, n_{f2k}, n_{c1k} y n_{c2k} son las cuatro frecuencias marginales asociadas a las tablas 2X2 de cada estrato *k*.

El estadístico Mantel-Haenszel es muy parecido al de Cochran, excepto por dos aspectos:

1) Utiliza la corrección por continuidad (resta medio punto al numerador antes de elevar al cuadrado).

2) Cambia ligeramente el denominador de la varianza, utilizando $n^2(n-1)$ en lugar de n^3.

En ambos estadísticos, el de Cochran y Mantel-Haenszel, si el valor de *p* es mayor que 0,05, se acepta la hipótesis de independencia y se concluye que, una vez tenido en cuenta el efecto de los estratos, las dos variables dicotómicas no están asociadas.

VIII.2.2. Medidas de asociación

En el caso de tablas de contingencia 2x2 se puede usar cualquiera de los tres coeficientes explicados anteriormente en el apartado VIII.1.2 de la función VIII1: *V* de Cramer, Coeficiente de Contingencia y Coeficiente de Incertidumbre. Además de estos tres coeficientes se puede usar el coeficiente Phi (ϕ).

VIII.2.2.1. Phi

El rango de este coeficiente solo es entre 0 y 1 en las tablas de contingencia 2x2 y, por ello, solo se debe usar cuando las variables a comparar son ambas dicotómicas. Phi (ϕ) es una medida basada en el estadístico χ^2. La fórmula es:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

donde *n* es el número de elementos de la muestra.

*VIII.2.2.2. Riesgo relativo, razón de ventajas (*odds ratio*) y las pruebas de Breslow-Day y de Tarone*

Cuando tenemos tablas de contingencia *K*x2x2, de las que se habló en el apartado anterior, la importancia de una asociación se estima mejor usando el riesgo relativo y la razón de ventajas (*odds ratio*). Estas medidas se pueden usar también en tablas 2x2, al igual que los coeficientes descritos anteriormente (*V* de Cramer, Coeficiente de Contingencia, Coeficiente de Incertidumbre y Phi), pero en el caso de las tablas *K*x2x2 estas medidas tienen la ventaja de que se pueden comparar entre los estratos.

Riesgo relativo nos indica cuánto más probable es que ocurra el suceso de un primer grupo frente al segundo. En una tabla 2x2 como la que se muestra en la tabla el riesgo o probabilidad del suceso considerado en el grupo A y en el grupo B serían:

	Grupo A	**Grupo B**	**Total**
Suceso	a	c	a+c
No suceso	b	d	b+d
Total	a+b	c+d	*n*

$$Riesgo(A) = \frac{a}{(a+b)} \quad Riesgo(B) = \frac{c}{(c+d)}$$

el cociente de ambas cantidades es el Riesgo relativo del suceso (RR_{suceso}) y del mismo modo se podría calcular el Riesgo relativo del suceso contrario o "no suceso" ($RR_{no-suceso}$)

$$RR_{suceso} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} \quad RR_{no-suceso} = \frac{\frac{b}{(a+b)}}{\frac{d}{(c+d)}}$$

El cociente de ambos riesgos relativos es igual a la Razón de ventajas o Razón de predominios (*OR*, *Odds Ratio*):

$$OR = \frac{RR_{suceso}}{RR_{no-suceso}} = \frac{(a*d)}{(b*c)}$$

La interpretación del Riesgo relativo es más intuitiva que la de la Razón de predominios. A pesar de ello la razón de ventajas se usa mucho, principalmente por su relación con la Regresión Logística, que se verá en otro capítulo, con importantes aplicaciones en Epidemiología y otras áreas.

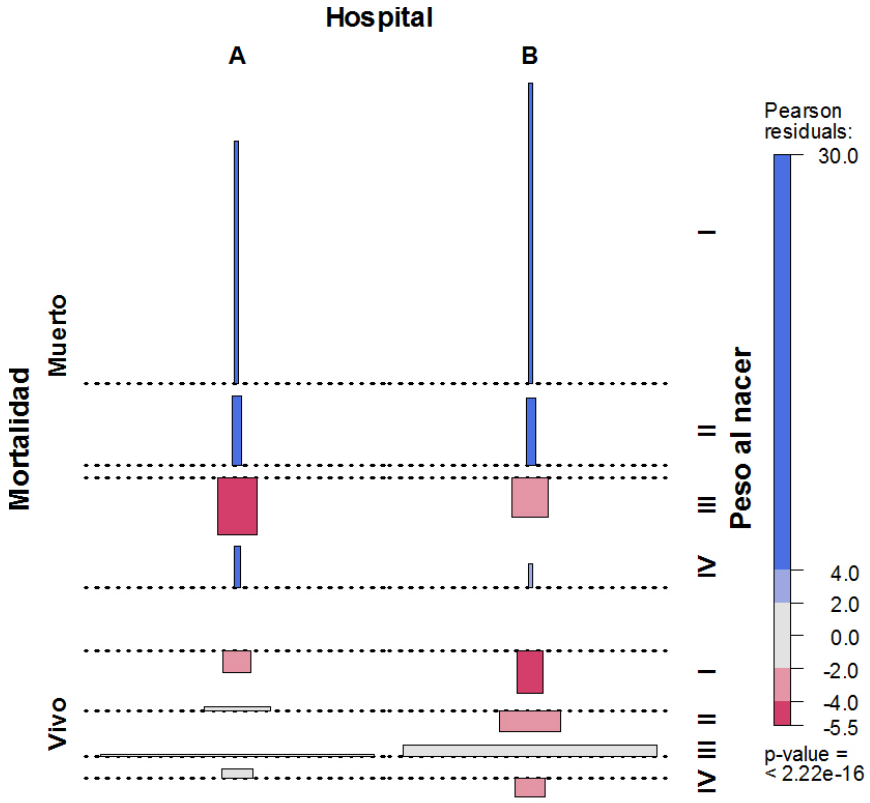
La prueba de Breslow & Day (1980) y la modificación de esta prueba de Tarone (1985) son pruebas de homogeneidad que permiten comparar las razones de ventaja entre los distintos estratos, en tablas de contingencia *K*x2x2. Con esta información se consigue determinar si el grado de asociación entre las dos variables dicotómicas varía en función de los diferentes estratos. En el ejemplo que se mencionó anteriormente, nos informaría acerca de si la posible asociación entre el virus del papiloma humano y el cáncer de cuello uterino aumenta o se hace menos intensa con la edad de las mujeres.

EJEMPLO

Se analizó el número de niños que morían al cabo de un año en relación a su peso al nacer, en dos hospitales diferentes. El grupo I eran bebés que pesaban al nacer menos de 1,5 kg, el grupo II entre 1,5 y 2,5 kg, el grupo III entre 2,5 y 4,2 kg y el grupo IV más de 4,2 kg. El objetivo era determinar si había diferencias en la mortalidad entre hospitales y se intentó eliminar el posible efecto que pueda tener el peso al nacer en la mortalidad infantil.

Figura VIII2 Representación gráfica de la tabla de contingencia de la proporción de bebés vivos

y muertos al cabo de un año, en los dos hospitales y para cada grupo de peso al nacer.



El análisis conjunto de los datos muestra que existen diferencias significativas en la mortalidad de los bebés (Figura VIII2, $p < 0{,}001$). En ambos hospitales el porcentaje de bebés muertos al cabo de un año es significativamente más alto en los grupos I y II (bebés con menos peso al nacer, barras azules por encima de la línea punteada) y también el el grupo IV (bebés con un peso muy superior a lo normal). Por el contrario, el porcentaje de bebés muertos era significativamente inferior (barras por debajo de la línea punteada) en el grupo III.

Lo primero que nos encontramos al abrir el archivo TXT con los resultados son una serie de tablas de doble entrada con los datos resumidos.

```
, , Peso.Nacer = I

          Hospital
Mortalidad    A    B
    Muerto   21   24
    Vivo     19    1

, , Peso.Nacer = II

          Hospital
Mortalidad    A    B
    Muerto   18   16
    Vivo    202  134

, , Peso.Nacer = III

          Hospital
Mortalidad    A    B
    Muerto   25   32
    Vivo   3365 2978

, , Peso.Nacer = IV

          Hospital
Mortalidad    A    B
    Muerto    5    3
    Vivo     55   26
```

A continuación podemos ver una serie de resultados de los diferentes análisis realizados. El χ^2 sin (χ^2_1 = 1,9, p = 0,168) y con corrección de Yates (χ^2_1 = 1,6, p = 0,196) no son significativos, por lo que concluimos que no existe asociación entre ambas variables, lo que quiere decir que -sin tener en cuenta el peso al nacer- no hay diferencias en el porcentaje de bebés vivos y muertos entre hospitales.

```
     Pearson's Chi-squared test

data:  tbl1
X-squared = 1.8976, df = 1, p-value = 0.1684


$listaPruebas[[2]]

     Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 1.6721, df = 1, p-value = 0.196
```

Las pruebas de Cochran (sin corrección de continuidad, χ^2_1 = 7,9, p = 0,004) y Mantel-Haenszel (con corrección de continuidad, χ^2_1 = 7,39, p = 0,006) son significativas, indicando que si tenemos en cuenta el peso al nacer la mortalidad depende del hospital, deja de ser independiente. También se muestran los valores de la razón de ventajas (*odds ratio*), que permiten explicar esa relación: en el nivel 1 del factor de estratificación (bebés con peso al nacer < 1,5kg) el predominio de fallecidos sobre no fallecidos en el hospital A es 0,046 veces el predominio del hospital B (22 veces más pequeño); en los restantes niveles de peso las diferencias son mucho menores.

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 7.9321, df = 1, p-value = 0.004857
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.3948980 0.8490013
sample estimates:
common odds ratio
        0.5790241


$listaPruebas[[4]]

        Mantel-Haenszel chi-squared test with continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 7.3942, df = 1, p-value = 0.006543
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.3948980 0.8490013
sample estimates:
common odds ratio
        0.5790241


$listaPruebas[[5]]

        Cochran-Mantel-Haenszel Chi-square Test

data:  tbl1
CMH statistic = 7.932, df = 1.000, p-value = 0.005, MH Estimate =
0.579, Pooled Odd Ratio = 0.793, Odd Ratio of level 1 = 0.046, Odd
Ratio of level 2 = 0.746, Odd Ratio of level 3 = 0.691, Odd Ratio of
level 4 = 0.788


$listaPruebas[[6]]
$X2.HBD
[1] 9.219336

$X2.HBDT
[1] 9.218351

$p
[1] 0.02652434
```

El test de Breslow (HBD = 9,2) y la correción de Tarone (HBDT = 9,2) sirven para comprobar si hay diferencias significativas entre estratos. En nuestro caso, el valor de p asociado a estos estadísticos es 0.026, con lo cual rechazamos la hipótesis nula de que no existen diferencias entre estratos. Esto quiere decir que el peso al nacer influye en la mortalidad de los bebés durante su primer año de vida.


```
$listaPruebas[[6]]
$X2.HBD
[1] 9.219336

$X2.HBDT
[1] 9.218351

$p
[1] 0.02652434
```


Value

Se obtiene un archivo TXT con los resultados de los contrastes y la gráfica que muestra la tabla de contingencia, si se seleccionó que se mostrara.

References

Breslow, N.E. & Day, N.E. (1980) *Statistical methods in cancer research.* Volume 1. The analysis or case-control studies. Lyon: International Agency for Research on Cancer.

Cochran, W.G. (1954) Some methods for strengthening the common χ^2 tests. *Biometrics*, 10: 417-451.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22: 719-748.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Tarone, R.E. (1985) On heterogeneity tests based on efficient scores. *Biometrika*, 72: 91-95.

Examples

```
## Not run:

data(ZVIII2)

#Mortalidad al nacer
VIII2(data=ZVIII2, varInteres=c("Mortalidad","Hospital"), varEstratos="Peso.Nacer",
pruebas=c(1,4,14,15,16,17), gp_labels=gpar(fontsize = 14, fontface = 2),
gp_varnames=gpar(fontsize = 16, fontface = 2),
labeling_args =list(set_varnames=c(Peso.Nacer="Peso al nacer" )))


## End(Not run)
```

X1 *CORRELACIONES BIVARIADAS SIMPLES*

Description

Se aplica un análisis de correlación entre dos variables usando los coeficientes de Pearson, Spearman y Kendall. Además se puede graficar la relación entre la variable dependiente y la independiente y también diferenciar entre grupos.

Usage

```
X1(data, varInteresX,  varInteresY, pruebas=c("Pearson","Spearman", "Kendall"),
varAgrupacion=NULL, ResetPAR=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL,
CEXPCH=1.3, COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteresX	Variable independiente.
varInteresY	Variable dependiente.
pruebas	Se puede elegir uno, dos o los tres coeficientes de correlación: Pearson, Spearman y/o Kendall.
varAgrupacion	Variable en el caso de querer estimar las correlaciones por grupos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
CEXPCH	Tamaño de los símbolos del gráfico.
COLOR	Permite modificar los colores de los símbolos, pero en el caso de que *varAgrupacion* no sea NULL, deben ser tantos como grupos diferentes tenga la variable *varAgrupacion*.
PCH	Vector con los símbolos del gráfico. Si es NULL se calculan automáticamente empezando por el símbolo 15.
LEGEND	Permite añadir una leyenda al gráfico.
AXIS	Permite añadir ejes al gráfico.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

X. CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUANTITATIVAS

Uno de los objetivos más frecuentes a la hora de analizar datos es el de estudiar el grado de dependencia o asociación entre variables. La correlación es el marco teórico general en el que podremos estudiar estos aspectos.

Decimos que una variable y depende de otra variable x cuando el valor de x determina en cierta medida el valor de y. Puede darse el caso de que el valor de y quede perfectamente determinado por el valor de x. Entonces, si conocemos el valor de la variable x, hay una relación funcional que nos define de forma perfecta el valor que tomaría y. En este caso hablaríamos de que entre ambas variables existe una relación de dependencia funcional. Pero en la práctica, con mayor frecuencia nos encontraremos con situaciones en las que el valor de la variable y no queda perfectamente definido por el valor de x, sino que existe un cierto grado de relación entre las dos que nos permite aproximar el valor de y. En este caso hablamos de una relación entre ambas variables de dependencia aleatoria.

Para analizar el grado de dependencia o asociación entre variables utilizamos una serie de estadísticos que se denominan coeficientes de correlación. Los coeficientes de correlación nos indican si las

variables están asociadas o no (si son independientes) y además, en qué sentido se da esa asociación o correlación entre variables (positiva o negativa). También nos informan acerca de la intensidad de la correlación. A través de diversos estadísticos, también podemos saber si la correlación entre variables es significativa o no.

X.1. CORRELACIONES BIVARIADAS SIMPLES

X.1.1. Coeficiente de correlación de Pearson (r)

El coeficiente de correlación de Pearson (*r*) indica la fuerza de la relación lineal entre las dos variables. Es necesario asumir, por tanto, que cualquier relación de dependencia que exista entre ambas variables es de tipo lineal. Solo sirve para variables cuantitativas.

El coeficiente *r* de Pearson se calcula según la siguiente fórmula, que se basa en la covarianza entre las dos variables:

$$r_{y_1,y_2} = \frac{\sum_{i=1}^{n}[(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2))]}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}}$$

o lo que es lo mismo

$$\frac{S_{xy}}{S_x S_y}$$

es decir, la covarianza dividida por el producto de las desviaciones típicas.

Los valores de este coeficiente se extienden en un rango entre -1 y 1. Cuanto más cerca de -1 o 1, más fuerte es la correlación entre las variables. Valores cercanos a 0 indican ausencia de correlación y, por tanto, de dependencia.

Para la prueba de significación del coeficiente de Pearson suele establecerse como H_0 que *R* = 0, utilizando el coeficiente de correlación *r* obtenido en la muestra como estimador del verdadero coeficiente de correlación poblacional *R*. Además es necesario para esta prueba que la distribución de probabilidad conjunta de ambas variables sea Normal.

$$t = \frac{r}{s_r}$$

donde s_r es el error estándar del estimador *r*:

$$s_r = \sqrt{\frac{(1-r^2)}{(n-2)}}$$

X.1.2. Coeficiente de correlación de Searman (ρ)

Entre los métodos no paramétricos más utilizados nos encontramos con el coeficiente de Spearman (ρ). Como en todos los métodos no paramétricos, no necesitamos hacer ninguna hipótesis acerca de la distribución que siguen nuestros datos. Se trata de un estadígrafo que se calcula de un modo similar al coeficiente de Pearson, pero asignando rangos a los datos. Cuando hay varios datos con el mismo valor se asigna a todos ellos el promedio de los rangos correspondientes.

Este coeficiente funciona para cualquier tipo de asociación entre variables que siga una función de tipo monotónico. Obviamente, esto se debe al hecho de utilizar una escala ordinal, con lo que esta propiedad sirve para cualquier otro coeficiente que emplee rangos para ser calculado. Esta es una diferencia importante frente al coeficiente de Pearson, que sólo responde bien si la relación entre

ambas variables es lineal. Además, sirve tanto para variables cuantitativas como para cualitativas ordinales.

Se calcula por medio de la siguiente fórmula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

donde d_i es la diferencia, para cada dato u observación, entre los rangos que toman los valores de las dos variables. El rango de valores que toma este coeficiente va desde -1 a 1. Cuanto más cerca de estos extremos, más fuerte es el grado de asociación, y el signo indica si la relación funcional entre ambas variables es descendente o ascendente, de la misma forma que el coeficiente de Pearson.

Para determinar la significación de este coeficiente se utiliza el siguiente estadístico, cuya distribución se aproxima a la *t* de Student cuando el tamaño de muestra es mayor que 20 unidades:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

Al igual que con el coeficiente de Pearson, se establece como H_0 que $\rho = 0$.

X.1.3. Coeficiente tau de Kendall

Este coeficiente es especialmente indicado para variables expresadas de forma ordinal. Por lo tanto también se puede utilizar tras asignar rangos a los valores de las variables.

Existe una fórmula para su cálculo que no tiene en cuenta "empates" entre rangos de datos, que se denota por τ_A, y se calcula de la siguiente manera:

$$\tau_A = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

donde n_c es el número de pares de rangos concordantes y n_d es el número de pares de rangos discordantes.

En el caso de que tengamos datos en los cuales existen "empates" entre rangos, utilizaríamos esta otra fórmula:

$$\tau_B = \frac{n_p - n_q}{\sqrt{(n_p + n_q + n_{E(X)})(n_p + n_q + n_{E(Y)})}}$$

donde *p* indica concordancia o no inversión, *q* discordancia o inversión y *E* empate considerando todos los pares de casos.

Los valores de este coeficiente se encuentran entre -1 y 1, y sus propiedades e interpretación son análogas a los coeficientes de Pearson y Spearman.

La significación se contrasta por medio de estadísticos que se distribuyen siguiendo una normal. En el caso de τ_A tendríamos la siguiente fórmula:

$$z = \frac{3(n_c - n_d)}{\sqrt{\frac{n(n-1)(2n+5)}{2}}}$$

Para el caso de τ_B tendríamos la siguiente fórmula:

$$z = \frac{n_c - n_d}{\sqrt{v}}$$

donde:

$$\upsilon = \frac{\upsilon_0 - \upsilon_t - \upsilon_u}{18 + \upsilon_1 + \upsilon_2}$$

donde a su vez:

$$\upsilon_0 = n(n-1)(2n+5)$$

$$\upsilon_t = \sum t_i(t_i-1)(2t_i+5)$$

$$\upsilon_u = \sum_j u_j(u_j-1)(2u_j+5)$$

$$\upsilon_1 = \sum_i t_i(t_i-1) \sum_j \frac{u_j(u_j-1)}{2n(n-1)}$$

$$\upsilon_2 = \sum_i t_i(t_i-1)(t_i-2) \sum_j \frac{u_j(u_j-1)(u_j-2)}{9n(n-1)(n-2)}$$

La H_0 para la significación de este coeficiente es que su valor es 0 y, por tanto, no hay asociación entre variables.

X.1.4. Coeficiente Gamma γ

Es el coeficiente más adecuado cuando los datos presentan un alto solapamiento de rangos. De nuevo, el intervalo de valores que puede tomar este coeficiente es entre -1 y 1, siendo -1 asociación perfecta negativa, 1 asociación perfecta positiva y 0 no asociación.

La fórmula para su cálculo es la siguiente:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

donde n_c es el número de pares de rangos concordantes y n_d es el número de pares discordantes.

La significación se comprueba utilizando el valor del siguiente estadístico, que se distribuye siguiendo una Normal:

$$z = \gamma \sqrt{\frac{n_c - d_d}{N(1-\gamma^2)}}$$

Al igual que con el resto de coeficientes, la H_0 que se establece es que el valor del coeficiente es igual a 0, y por lo tanto no existe asociación entre las variables.

FUNCIONES

Para estimar las correlaciones se utiliza la función rcor.test (Rizopoulos 2006; 2013).

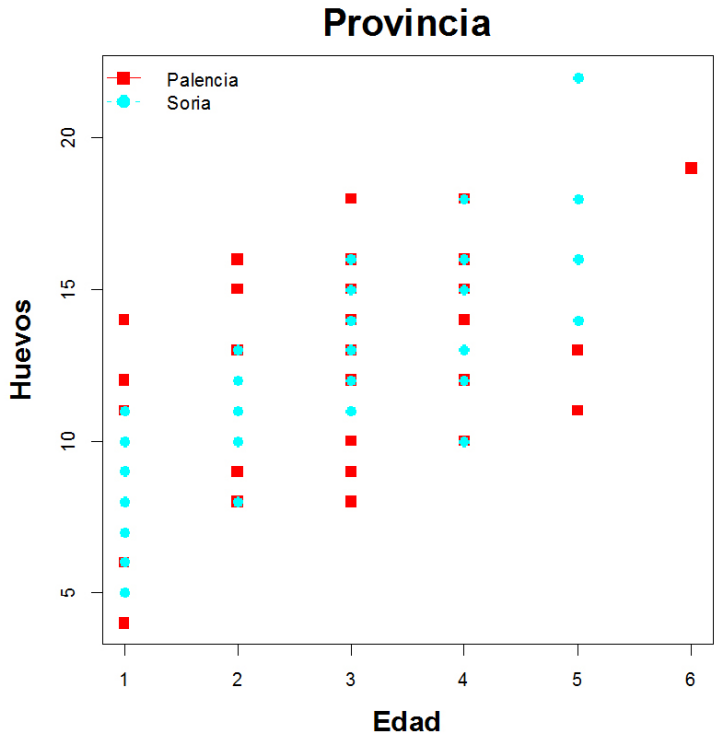
EJEMPLO

Datos de un muestreo realizado en nidadas de una pequeña especie de ave, la codorniz común. En este muestreo se contó el número de huevos por hembra, puestos durante una época de cría, en 50 hembras marcadas cuya edad era conocida. Al mismo tiempo se registraron la temperatura media

durante la misma época de cría y la disponibilidad de alimento para las hembras (cuyos valores son porcentajes respecto al máximo medido). El mismo muestreo se repitió en la siguiente época de cría con una muestra diferente de hembras en la misma población. El objetivo de la investigación es descubrir si existe relación entre la edad de la hembra y el número de huevos por puesta. Las estimaciones se realizan para cada una de las Provincias muestreadas.

La Figura X.1 muestra que existe una clara relación positiva entre la edad de la hembra y el número de huevos por puesta en las dos provincias.

Figura X.1 Relación entre la edad de la hembra de codorniz y el número de huevos por puesta, para cada una de las provincias.



La tabla de resultados muestra cada uno de los coeficientes de correlación en la mitad superior derecha de la tabla, y en la mitad inferior izquierda la probabilidad. Se observa que, en todos los casos, se rechaza la hipótesis de independencia, ya que la probabilidad es siempre menor que 0,05. Por lo tanto, en ambas provincias y considerando los tres coeficientes, se concluye que existe una relación de dependencia entre la edad de la hembra y el número de huevos. Esta asociación parece ser más fuerte en la provincia de Soria donde, por ejemplo, el coeficiente de correlación de Pearson es 0,78, mientras que en Palencia es 0,45. Los coeficientes de Spearman y Kendall también muestran esta mayor asociación entre ambas variables en una provincia que en la otra.

```
[1] "Pearson"

[[2]]
[[2]]$Palencia

       Edad   Huevos
Edad    *****  0.448
Huevos  0.004  *****

upper diagonal part
lower diagonal part


[[2]]$Soria

       Edad   Huevos
Edad    *****  0.777
Huevos <0.001  *****
```

```
[1] "Spearman"

[[4]]
[[4]]$Palencia

       Edad   Huevos
Edad    *****  0.352
Huevos  0.026  *****

upper diagonal part
lower diagonal part


[[4]]$Soria

       Edad   Huevos
Edad    *****  0.775
Huevos <0.001  *****
```

```
[1] "Kendall"

[[6]]
[[6]]$Palencia

       Edad   Huevos
Edad    *****  0.282
Huevos  0.022  *****

upper diagonal part
lower diagonal part


[[6]]$Soria

       Edad   Huevos
Edad    *****  0.663
Huevos <0.001  *****
```

Value

Se obtiene un archivo TXT con los resultados de las correlaciones y se puede representar un gráfico relancionado la variable dependiente con la independiente, diferenciando además entre los grupos de la variable agrupación.

References

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Examples

```
## Not run:

data(ZX1)

X1(data=ZX1,varInteresX="Edad", varInteresY="Huevos", varAgrupacion="Provincia")


## End(Not run)
```

X2 *CORRELACIONES PARCIALES*

Description

Se aplica un análisis de correlación parcial entre dos variables usando los coeficientes de Pearson y Kendall, y teniendo en cuenta el efecto de otra variable. Además se puede graficar la relación entre la variable dependiente y la independiente y también diferenciar entre grupos.

Usage

```
X2(data, varInteresX,  varInteresY, varInteresZ, pruebas=c("Pearson","Kendall"),
varAgrupacion=NULL, ResetPAR=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL,
CEXPCH=1.3, COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteresX	Variable independiente.
varInteresY	Variable dependiente.
varInteresZ	Variable que se quiere contrastar si afecta a la correlación.
pruebas	Se puede elegir uno o los dos coeficientes de correlación: Pearson y/o Kendall.
varAgrupacion	Variable en el caso de querer estimar las correlaciones por grupos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
CEXPCH	Tamaño de los símbolos del gráfico.
COLOR	Permite modificar los colores de los símbolos, pero en el caso de que *varAgrupacion* no sea NULL, deben ser tantos como grupos diferentes tenga la variable *varAgrupacion*.
PCH	Vector con los símbolos del gráfico. Si es NULL se calculan automáticamente empezando por el símbolo 15.
LEGEND	Permite añadir una leyenda al gráfico.
AXIS	Permite añadir ejes al gráfico.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida.

Details

X. CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUANTITATIVAS

X.2. CORRELACIONES PARCIALES

X.2.1. Coeficiente de correlación de Pearson

En ocasiones, una correlación lineal bivariada puede estar enmascarada por el efecto de otras variables que no controlamos. Por ejemplo, puede existir una relación fuerte entre el tamaño de la nidada de una determinada especie de pájaro y la edad de la hembra y, sin embargo, esta correlación se enmascara por la influencia en nuestros datos de otras variables como podrían ser la temperatura

media durante la época de cría. Para estudiar una correlación simple entre dos variables de interés, a la vez que controlamos el efecto de otras variables, utilizamos las correlaciones parciales.

Si tenemos dos variables *X* e *Y* sobre las que queremos calcular la correlación parcial controlando *k* variables *Z*, el método se basa en calcular el coeficiente de correlación de Pearson entre dos variables derivadas d_x y d_y tales que:

d_x = residuos de la regresión lineal de *x* con z_1, ..., z_k
d_y = residuos de la regresión lineal de *y* con z_1, ..., z_k

Es posible controlar al mismo tiempo el efecto de una o más variables *Z*. Si controlamos sólo una, hablaríamos de correlación parcial de primer orden. Si controlamos dos, correlación parcial de segundo orden, y así sucesivamente.

El cálculo de las correlaciones parciales de primer orden se realiza por medio de la siguiente fórmula:

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

donde *r* es el coeficiente de correlación de Pearson entre las variables indicadas.

Como ejemplo de mayor orden, mostramos la fórmula del coeficiente en el caso de que se estén controlando dos variables (z_i y z_j) con lo cual es una correlación parcial de segundo orden. Las de mayor orden se obtendrían siguiendo la misma lógica:

$$r_{xy,z_iz_j} = \frac{r_{xy,z_i} - r_{xz_j,z_i}r_{yz_j,z_i}}{\sqrt{(1 - r_{xz_j,z_i}^2)(1 - r_{yz_j,z_i}^2)}}$$

La prueba de significación de este coeficiente, con la hipótesis nula de que su valor poblacional es cero, se realiza calculando el estadístico *t* por medio de la siguiente fórmula:

$$T = r_{xy,k}\frac{\sqrt{m - k - 2}}{\sqrt{(1 - r_{xy,k}^2)}}$$

donde *m* es el número de casos (si no es igual para todos los pares de variables debe tomarse el menor de ellos) y *k* es el número de variables controladas que hemos empleado.

X.2.2. Coeficiente de correlación parcial de Kendall ($T_{xy,z}$)

El fundamento de las correlaciones parciales es aquí el mismo que en el caso de la estadística paramétrica pero necesitamos un estadístico diferente, el cual nos permita trabajar sin hacer ninguna asunción acerca de la distribución que siguen nuestros datos.

El principal estadístico que cumple esta función es la *T* de Kendall, que es en esencia la misma que la τ descrita en el apartado X.1.3 de la función X1, pero adaptada al cálculo de correlaciones parciales.

La fórmula para computarla es la siguiente:

$$T_{xy,z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{(1 - T_{xz}^2)(1 - T_{yz}^2)}}$$

donde cada correlación bivariada T se calcula de la misma forma que la τ descrita en el apartado X.1.3 de la función X1.

Para el cálculo de su significación se emplea el siguiente estadístico, que sigue una distribución Normal:

$$z = \frac{3T_{xy,z}\sqrt{N(N-1)}}{\sqrt{2(2N+5)}}$$

La H_0 que se establece, como es habitual para estos estadísticos, es que el valor del coeficiente poblacional es 0.

FUNCIONES

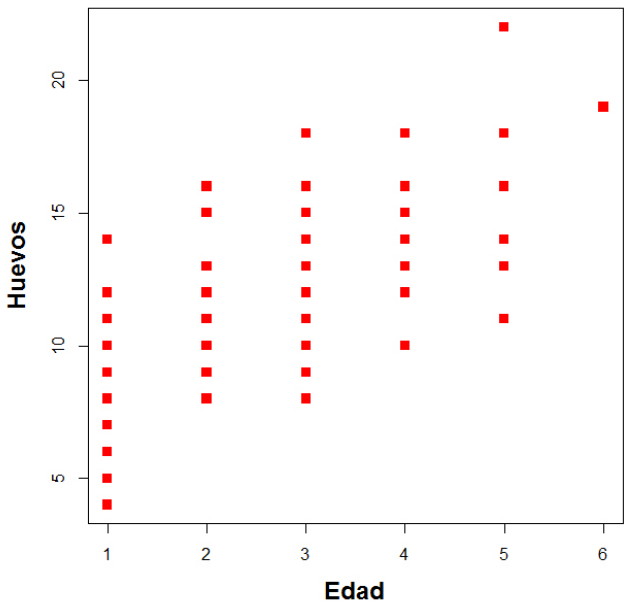
Para estimar las correlaciones se utiliza la función rcor.test (Rizopoulos 2006; 2013).

EJEMPLO

Datos de un muestreo realizado en nidadas de una pequeña especie de ave, la codorniz común. En este muestreo se contó el número de huevos por hembra, puestos durante una época de cría, en 50 hembras marcadas cuya edad era conocida. Al mismo tiempo se registraron la temperatura media durante la misma época de cría y la disponibilidad de alimento para las hembras (cuyos valores son porcentajes respecto al máximo medido). El mismo muestreo se repitió en la siguiente época de cría con una muestra diferente de hembras en la misma población. El objetivo de la investigación es determinar si la temperatura afecta a la posible relación entre la edad de la hembra y el número de huevos por puesta. Los cálculos se realizan sin diferenciar entre provincias.

La Figura X.2 muestra que existe una clara relación positiva entre la edad de la hembra y el número de huevos por puesta, considerando el conjunto de todas las provincias.

Figura X.2 Relación entre la edad de la hembra de codorniz y el número de huevos por puesta, considerando todas las provincias.



En el coeficiente de Pearson, la matriz superior derecha muestra que existe una correlación entre ambas variables de 0,637, y la matriz inferior izquierda es la probabilidad. Se observa que se rechaza la hipótesis de independencia, ya que p < 0,001. Por tanto, existe una correlación significativa entre la edad de la hembra y el número de huevos por hembra. En el coeficiente de Kendall es 0,498 y también p < 0,001.

```
[[1]]
[1] "Pearson"

[[2]]
$cor.mat
                Edad    Huevos Temperatura
Edad         1.0000000 0.6371661  -0.1068197
Huevos       0.6371661 1.0000000   0.5552461
Temperatura -0.1068197 0.5552461   1.0000000

$p.values
              pvals
[1,] 1 2 1.025775e-12
[2,] 1 3 2.901482e-01
[3,] 2 3 2.030009e-09
```

```
[[3]]
[1] "Kendall"

[[4]]
$cor.mat
                Edad    Huevos Temperatura
Edad         1.00000000 0.4983081 -0.08312776
Huevos       0.49830809 1.0000000  0.49301219
Temperatura -0.08312776 0.4930122  1.00000000

$p.values
              pvals
[1,] 1 2 1.255461e-10
[2,] 1 3 3.586914e-01
[3,] 2 3 9.474623e-09
```

A continuación se muestran los resultados incluyendo la variable temperatura durante la época de cría. De nuevo solo hay que fijarse en el coeficiente entre la variable dependiente y la independiente, que para Pearson es 0,842 y para Kendall 0,622. En ambos casos la probabilidad, que se ve en la matriz *pvalue*, es significativa para las dos pruebas con p < 0,001. Por tanto, como ambos coeficientes de correlación aumentan con respecto a cuando no se consideraba la variable control, en este ejemplo la temperatura, 0,637 para Pearson y 0,498 para Kendall, se concluye que la relación entre la edad de la hembra y el número de huevos por hembra varía en función de la temperatura durante la época de cría.

```
[[5]]
[1] "Pearson partial correlation"

[[6]]
[[6]]$estimate
                Edad    Huevos Temperatura
Edad         1.0000000 0.8422472  -0.7185677
Huevos       0.8422472 1.0000000   0.8133817
Temperatura -0.7185677 0.8133817   1.0000000

[[6]]$p.value
                  Edad       Huevos  Temperatura
Edad        0.000000e+00 8.942973e-28 5.586822e-17
Huevos      8.942973e-28 0.000000e+00 1.489113e-24
Temperatura 5.586822e-17 1.489113e-24 0.000000e+00
```

```
[[7]]
[1] "Kendall partial correlation"

[[8]]
[[8]]$estimate
                 Edad    Huevos Temperatura
Edad         1.0000000 0.6220117  -0.4358950
Huevos       0.6220117 1.0000000   0.6185601
Temperatura -0.4358950 0.6185601   1.0000000

[[8]]$p.value
                    Edad       Huevos  Temperatura
Edad        0.000000e+00 7.378824e-20 1.631946e-10
Huevos      7.378824e-20 0.000000e+00 1.175779e-19
Temperatura 1.631946e-10 1.175779e-19 0.000000e+00
```

Value

Se obtiene un archivo TXT con los resultados de las correlaciones y se puede representar un gráfico relacionando la variable dependiente con la independiente, diferenciando además entre los grupos de la variable agrupación.

References

Kim, S. (2013) Partial and Semi-partial (Part) correlation. R package version 1.0. Available at: http://CRAN.R-project.org/package=ppcor.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Examples

```
## Not run:

data(ZX1)

X2(data=ZX1,varInteresX="Edad", varInteresY="Huevos", varInteresZ="Temperatura")


## End(Not run)
```

X3 *CORRELACIONES MULTIPLES*

Description

Se aplica un análisis de correlación entre todas las variables seleccionadas por el usuario y se calculan también los valores de probabilidad. Además se realizan dos gráficos matriciales de correlaciones.

Usage

```
X3(data, varInteres, ResetPAR=TRUE, PAR=NULL, CORRPLOT=NULL, PAIRS=NULL,
colhist="#00FFFFFF", file="Output.txt")
```

Arguments

data	Archivo de datos.
varInteres	Variables para las que se calculan la correlaciones.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes de ambos gráficos.
CORRPLOT	Se accede a la función corrplot que permite especificar los argumentos del primer gráfico matricial.
PAIRS	Se accede a la función pairs que permite especificar los argumentos del segundo gráfico matricial.
colhist	Color de las barras de los histogramas en el segundo gráfico matricial.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

X. CONTRASTES DE INDEPENDENCIA Y ASOCIACION EN VARIABLES CUANTITATIVAS

X.3. CORRELACIÓN MÚLTIPLE

X.3.1. Coeficiente de correlación múltiple de Pearson

Se trata de una extensión del método de correlación simple de Pearson a casos en los que tengamos más de una variable independiente. Este coeficiente indica el grado de asociación entre varias variables independientes x_i y una variable dependiente y. Este coeficiente suele denotarse como *R*, y su cuadrado R^2 es el "coeficiente de determinación", que se interpreta como la proporción de varianza explicada por la regresión lineal múltiple:

$$R^2 = c' R_{xx}^{-1} c$$

donde c es una matriz o vector columna con todas las correlaciones posibles entre las variables independientes y la dependiente. *c'* es la traspuesta de esta matriz o vector fila. R_{xx} es una matriz de correlaciones entre todas las variables independientes, y el superíndice -1 indica que en la fórmula se utiliza la inversa de esta matriz. También se puede calcular como el coeficiente de correlación simple -bivariado- entre la variable dependiente y la estimación de *y* obtenida mediante el modelo de regresión lineal múltiple, *y'*.

La significación se puede obtener por medio del estadístico *F*, calculado con la siguiente fórmula:

$$F = \frac{\frac{R^2_{yx_{1\ldots k}}}{k}}{\frac{(1-R^2_{yx_{1\ldots k}})}{(n-k-1)}}$$

donde el término $R^2_{yx_{1\ldots k}}$ es el valor del coeficiente de correlación múltiple, *k* es el número de variables independientes, y *n* el número de datos. Los grados de libertad de la *F* son *k* y (*n* - *k* - 1).

X.3.2. Coeficiente de concordancia de Kendall (W)

Todos los coeficientes no paramétricos para medir asociación entre variables que hemos visto hasta ahora eran aplicables para datos bivariados y se basan en la comparación de dos grupos de rangos o de variables ordinales. El coeficiente de concordancia de Kendall (*W*), sin embargo, nos permitirá trabajar con más de dos grupos de rangos o variables ordinales. Es por tanto, en cierta medida, análogo al coeficiente de correlación múltiple de Pearson, aunque en este caso no existe una única variable dependiente. Supongamos que tenemos las puntuaciones de *m* jueces (por ejemplo catadores) sobre *n* objetos (por ejemplo vinos) y queremos medir el grado de acuerdo entre jueces. La fórmula para computarlo es la siguiente:

$$W = \frac{12S}{m^2(n^3 - n)}$$

donde:

$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2$$

R_i es el rango medio para el objeto *i* y $\bar{R}$ es la media de los rangos:

$$R_i = \sum_{j=1}^{m} r_{i,j} \qquad \bar{R} = \tfrac{1}{2}m(n+1)$$

En caso de "empates" entre varios datos se asigna a cada dato la media de los rangos (si el tercero y el cuarto son iguales, su rango será 3,5 para ambos).

Para probar la significación de este estadístico tendríamos dos casos posibles:

- Si $3 \leq N \leq 7$ se emplea una distribución generada a partir de todas las combinaciones posibles de rangos, que existe tabulada para diferente número de variables o grupos ($3 \leq k \leq 20$), y amplio número de rangos (Siegel & Castellan, 1988).

- Si *N*>7 utilizaremos este otro estadístico, que sigue una distribución aproximadamente χ^2 con n - 1 grados de libertad:

$$\chi^2 = k(n-1)W$$

En cualquier caso, la H_0 que se establece será que el valor de *W* es igual a cero.

X.3.3. Cuadrado del coeficiente de correlación lineal múltiple

Otro procedimiento de cálculo es el cuadrado del coeficiente de correlación lineal múltiple. Ese cuadrado se interpreta como el coeficiente de determinación o proporción de varianza explicada por la regresión lineal sobre el conjunto de todas las variables explicativas.

Se calcula la matriz de correlaciones para todo el conjunto, se invierte esa matriz, y finalmente se restan de la unidad los inversos de los elementos de la diagonal, obteniendo así simultáneamente todos los coeficientes de correlación múltiple al cuadrado. Este es el método que usaremos para estimar la correlación de una variable con varias variables.

El cuadrado del coeficiente de correlación lineal múltiple es igual al coeficiente de determinación r^2 que se obtiene en las regresiones.

FUNCIONES

Se utiliza la función rcor.test del paquete ltm (Rizopoulos 2006;2013) para el cálculo de las correlaciones y probabilidades entre las variables y la función smc del paquete psych (Revelle, 2014) para

calcular las correlaciones múltiples. Para el primer gráfico se utiliza la función corrplot del paquete corrplot (Wei, 2013). Para el segundo gráfico se usan las funciones panel.hist, panel.cor y panel.reg del paquete SciViews (Grosjean, 2014) y pairs del paquete base graphics.

EJEMPLO

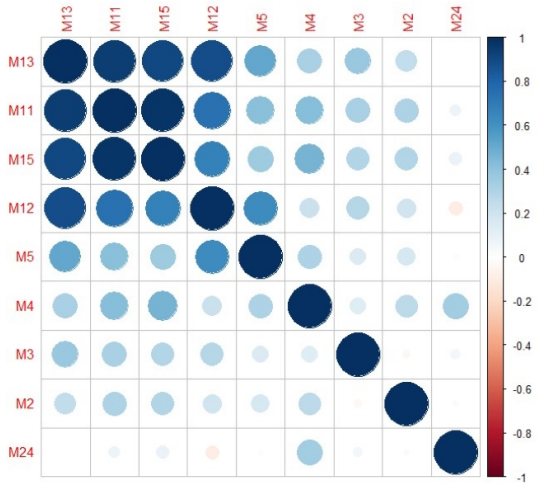
Datos morfométricos de varias especies del orden Characiformes, como la longitud de la base de la aleta dorsal (M12), altura del cuerpo (M11), etc. Para más detalles ver Guisande et al. (2010). El objetivo es estimar la correlación entre varias variables.

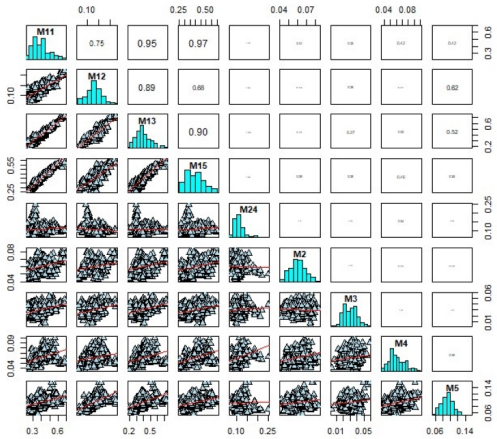
Los resultados se muestran en la siguiente tabla, donde la matriz superior derecha son las correlaciones entre variables, y la inferior izquierda las probabilidades.

	M11	M12	M13	M15	M24	M2	M3	M4	M5
M11	*****	0.747	0.945	0.975	0.076	0.307	0.327	0.424	0.418
M12	<0.001	*****	0.887	0.678	-0.099	0.194	0.285	0.216	0.621
M13	<0.001	<0.001	*****	0.903	-0.006	0.246	0.370	0.324	0.518
M15	<0.001	<0.001	<0.001	*****	0.088	0.292	0.292	0.461	0.354
M24	0.167	0.071	0.906	0.107	*****	-0.020	0.054	0.342	0.018
M2	<0.001	<0.001	<0.001	<0.001	0.709	*****	-0.035	0.261	0.177
M3	<0.001	<0.001	<0.001	<0.001	0.327	0.527	*****	0.138	0.151
M4	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.011	*****	0.301
M5	<0.001	<0.001	<0.001	<0.001	0.744	0.001	0.006	<0.001	*****

En las Figuras X.3 y X.4 se representan las correlaciones entre las medidas morfométricas de los peces en dos formatos diferentes. En el caso de la Figura X.3 las variables se ordenan por defecto usando un Análisis de Componentes Principales, aunque hay otros métodos disponibles.

Figuras X.3 y X.4 Correlación entre medidas morfométricas de varias especies de peces del orden Characiformes.





Por último, en la siguiente tabla se muestran las correlaciones múltiples de cada variable con todas las restantes. Por ejemplo, valor de 0,975 para la variable M11 significa que el resto de variables explican el 97,5% de la variabilidad observada en M11.

```
[1] "Squared Multiple Correlation (SMC)"

[[4]]
      M11       M12       M13       M15       M24        M2        M3        M4
0.9756485 0.8923650 0.9728759 0.9611285 0.1751850 0.1691844 0.2116277 0.4207534
       M5
0.4620176
```

Value

Se obtiene un archivo TXT con los resultados de las correlaciones y los valores de probabilidad, además de realizar dos gráficos matriciales de correlaciones.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Grosjean, P. (2014) SciViews GUI API - Main package. R package version 0.9-5. Available at: http://CRAN.R-project.org/package=SciViews.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17(5)**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Revelle, W. (2014) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.4.5. Available at: http://CRAN.R-project.org/package=psych.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Wei, T. (2013) Visualization of a correlation matrix). R package version 0.73. Available at: http://CRAN.R-project.org/package=corrplot.

Examples

```
## Not run:

data(ZX3)

#Correlación entre variables morfométricas de especies de peces
#del orden Characiformes

X3(data=ZX3, varInteres=c("M11","M12","M13","M15","M24","M2","M3","M4","M5"))


## End(Not run)
```

XI1 *REGRESION SIMPLE*

Description

Se aplican diferentes modelos de regresión para describir el tipo de función que mejor se ajusta a la posible relación que existe entre dos variables.

Usage

```
XI1(data, varInteresY, varInteresX, modelo=c("Linear", "Log", "S-curve",
"Power", "Exp",  "Quadratic", "Cubic", "Inverse"), outliers=NULL,
quant1=0.05, quant2 = 0.95, ResetPAR=TRUE, mfrow=c(2,4), PAR=NULL,
PLOT=NULL, PLOTR=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1.3, COLABLINE="#48D1CCFF",
COLOR="#C0FF3EFF", PCH=15, resPlot=TRUE, LEGEND=NULL, AXIS=NULL,
MTEXT= NULL, TEXT=NULL, file1="Output.txt", file2="Coefficients.csv",
file3="Residuals.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteresY	Variable dependiente.
varInteresX	Variable independiente.
modelo	Se puede elegir uno, o varios modelos de regresión: "Linear","Log", "S-curve", "Power", "Exp", "Quadratic", "Cubic", "Inverse". No se estima el modelo en aquellos casos que hay que aplicar el logaritmo a alguna de las variables, si algún valor de la variable, a la que se aplica el logaritmo, es cero o negativo. El modelo inverso no se calcula si algún valor de la variable independiente es cero.

outliers	Si es NULL no se eliminan los valores atípicos (*outliers*). Si se elige algún modelo de regresión, "Linear","Log", "S-curve", "Power", "Exp", "Quadratic", "Cubic" o "Inverse", entonces se eliminan los valores atípicos usando el modelo de regresión seleccionado.
quant1	Cuantil del extremo inferior para la eliminación de los datos atípicos.
quant2	Cuantil del extremo superior para la eliminación de los datos atípicos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
mfrow	Si es NULL y hay varios modelos de regresión seleccionados, éstos salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
PLOT	Se accede a la función plot.default que permite modificar el gráfico de las regresiones.
PLOTR	Se accede a la función plot.default que permite modificar el gráfico de los residuos.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
CEXPCH	Tamaño de los símbolos del gráfico.
COLABLINE	Color de la línea del modelo de regresión.
COLOR	Color de los símbolos.
PCH	Tipo de símbolo.
resPlot	Si es TRUE se muestran los gráficos que representan la relación entre el valor pronosticado y el residuo tipificado.
LEGEND	Permite añadir una leyenda al gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
AXIS	Permite añadir ejes al gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
MTEXT	Permite añadir textos en los márgenes del gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
file1	ARCHIVO TXT. Nombre del archivo de salida con los resultados de los análisis.
file2	ARCHIVOS CSV. Nombre del archivo con los coeficientes de los modelos de regresión.
file3	ARCHIVOS CSV. Nombre del archivo con los residuos de los modelos de regresión.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.

`dec`	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

Se define regresión como la teoría que trata de expresar mediante una función matemática la relación que existe entre una variable dependiente y una (regresión simple) o varias (regresión múltiple) variables independientes. La obtención de esta función permite predecir cual será el valor de la variable dependiente en función del valor que tome la variable o variables independientes.

La regresión se diferencia de la correlación en que esta última estudia el grado de asociación entre las variables, y determina si la relación es o no es significativa, mientras que la regresión, como se mencionó anteriormente, trata de definir la función que mejor explica la relación entre las variables.

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

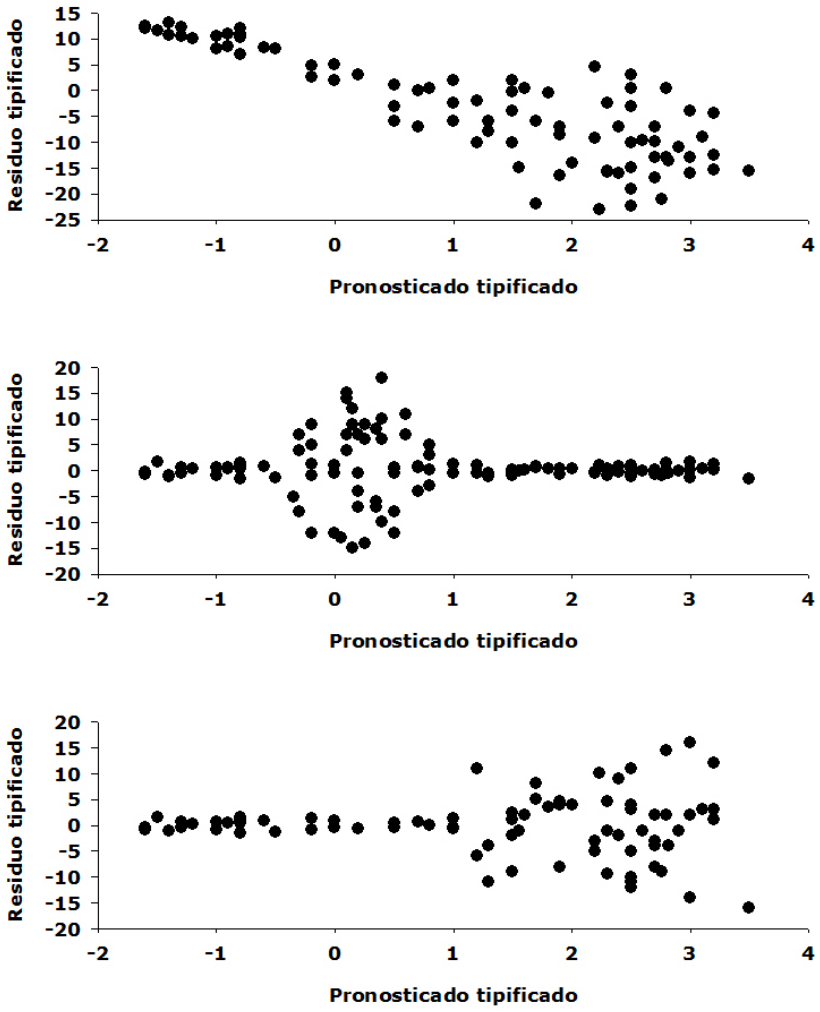
XI.1.1. Requisitos

Para aplicar un modelo de regresión entre variables no se requiere que los datos presenten una distribución normal o que exista homogeneidad de varianzas; sin embargo, para poder determinar si la función obtenida con el modelo de regresión es significativa es necesario aplicar contrastes, y para ello se tienen que cumplir los siguientes requisitos, que en gran medida están relacionados con los residuos (diferencia entre el valor observado de la variable dependiente y el valor ajustado por la función):

1. Los residuos obtenidos del modelo de regresión deben presentar una distribución Normal.

2. Debe existir homocedasticidad en los residuos, es decir, la varianza de los mismos debe ser constante. Las gráficas que se muestran a continuación (Figura XI1.1) son ejemplos en los que no se cumple la homocedasticidad de los residuos, porque su variabilidad no es constante, sino que cambia dependiendo de los valores pronosticados. En la primera gráfica, obtenida en el ajuste de una recta, se observa además, que la relación no es lineal, ya que no se mantiene la tendencia horizontal.

Figura XI1.1. Ejemplos en los que no se cumple la homocedasticidad.



3. No debe existir autocorrelación en la serie de residuos (deben ser independientes). Si este requisito no se cumple, no es posible saber el grado de relación exacta entre la variable independiente y la dependiente, ya que parte de la predicción se debe a los propios valores de la variable dependiente.

4. En el caso del modelo de regresión múltiple, no debe existir relación entre las variables independientes, es decir, no debe existir multicolinealidad.

XI.1.2. Regresión simple

Las funciones más comunes, y que se calculan en esta función, son las siguientes, que se muestran como están expresadas en el archivo de salida TXT con los resultados de los modelos y en el CSV donde se muestran los coeficientes de los modelos:

Modelo	**CSV**	**TXT**
Lineal	$y = a + bx$	Igual
Logarítmico	$y = e^{a+\frac{b}{x}}$	Igual
Curva-S	$y = e^{a+\frac{b}{x}}$	Igual

Potencial	$y = ax^b$	$y = e^{ln(a)+bln(x)}$
Exponencial	$y = ae^{bx}$	$y = e^{ln(a)+bx}$
Cuadrática	$y = a + bx + cx^2$	Igual
Cúbica	$y = a + bx + cx^2 + dx^3$	Igual
Inversa	$y = a + \frac{b}{x}$	Igual

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors, la función dwtest del paquete lmtest (Hothorn et al., 2013) para analizar la autocorrelación con el test de Durbin-Watson y la función bptest del paquete lmtest (Hothorn et al., 2013) para realizar el test Breusch-Pagan de homocedasticidad.

EJEMPLO

Los datos corresponden a un estudio que tiene como objetivo encontrar el mejor modelo de regresión que existe entre el peso y el ancho de la concha para un conjunto de tres especies de moluscos.

Paso 1.

En primer lugar se analiza la relación entre el peso y el ancho de la concha considerando las tres especies de moluscos y todos los modelos de regresión.

Los residuos de los modelos, los cuales están en uno de los archivos CSV que genera esta función, se muestran parcialmente en la siguiente tabla. Son útiles para identificar valores atípicos, para ver si se cumple la homocedasticidad, como se mostró en la Figura XI1.1 y para analizar la Normalidad.

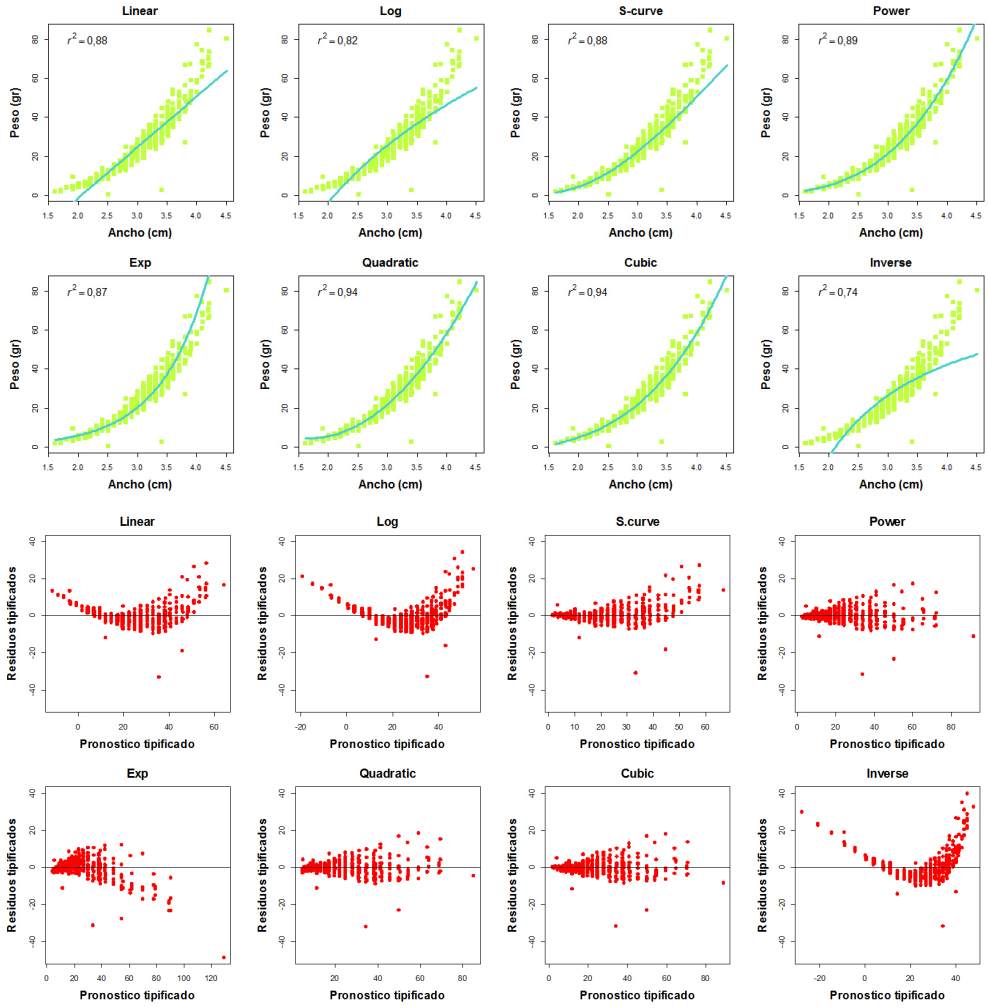
Peso	Ancho	Linear	Log	S.curve	Power	Exp	Quadratic	Cubic	Inverse
32,14	3,42	-3,39451695	-2,99745376	-1,27288609	-1,90681061	-1,63325628	-2,56620918	-2,13843131	-2,07132384
13,33	2,61	-1,06074769	-2,21955868	-0,40767149	0,36408798	0,84216812	0,29493773	-0,07414737	-3,69431857
46,99	3,71	3,88549159	5,95413047	5,19380668	1,45657075	-1,23479096	1,30346209	1,61001891	8,44975389
28,5	3,1	1,31857708	0,48186956	3,63870546	4,52882464	5,70328152	3,93872629	4,0830398	0,00535748
28,79	3,32	-4,13417506	-4,19685579	-1,86142856	-1,83270091	-1,0795719	-2,52374284	-2,15263685	-3,753241

La siguiente tabla muestra los coeficientes de las ecuaciones de los modelos de regresión y los r^2.

Model	a	b	c	d	r2
Linear	-53,739	26,103	NA	NA	0,88
Log	-53,974	72,47	NA	NA	0,82
S-curve	6,3729	-9,7946	NA	NA	0,88
Power	0,42137	3,5718	NA	NA	0,89
Exp	0,50607	1,2283	NA	NA	0,87
Quadratic	33,35	-34,141	10,099	NA	0,94
Cubic	-10,668	12,317	-5,7827	1,7614	0,94
Inverse	89,592	-189,4	NA	NA	0,74

La Figura XI1.2, que genera la función, muestra los ocho modelos de regresión en un solo panel y en la figura Figura XI1.3 los residuos (la línea negra indica un residuo de cero).

Figuras XI1.2 y XI1.3. Relación entre ancho de la concha y peso en tres especies de moluscos y los residuos de las regresiones.



Aunque los valores de r^2 son mayores en los modelos cuadrático y cúbico, vemos como en el modelo potencial los residuos son igual de pequeños, tienen un comportamiento lineal y son muy parecidos a lo largo de todo el rango de valores pronosticados de la variable dependiente, a los obtenidos con los modelos cuadrático y cúbico (Figura XI1.3). Por tanto, elegimos el modelo potencial, ya que es preferible elegir un modelo más sencillo, siempre que los residuos sean semejantes a los obtenidos en los modelos cuadrático y cúbico.

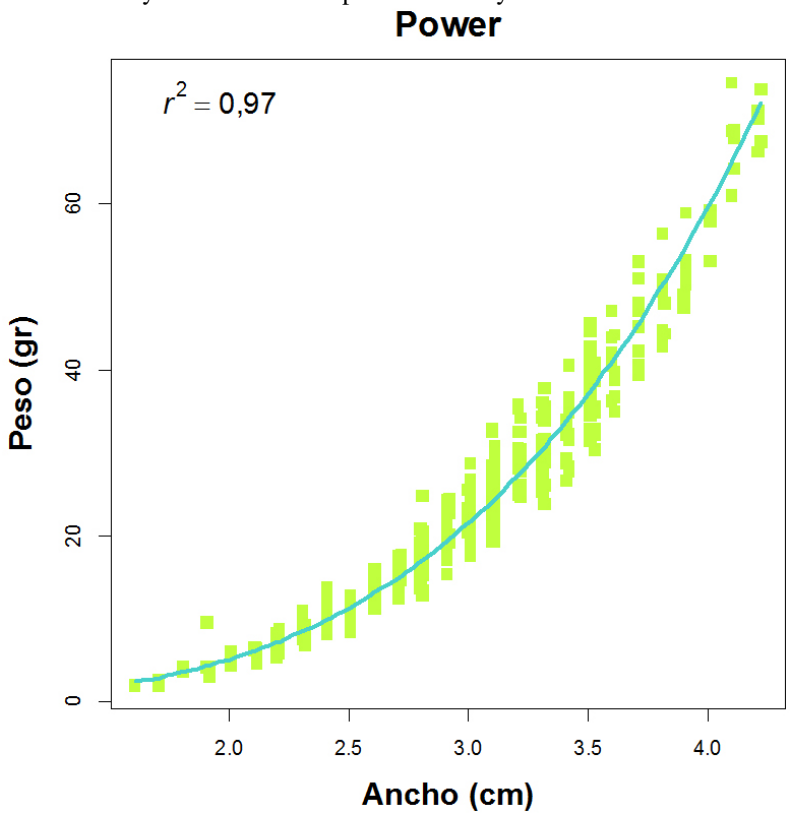
Paso 2.

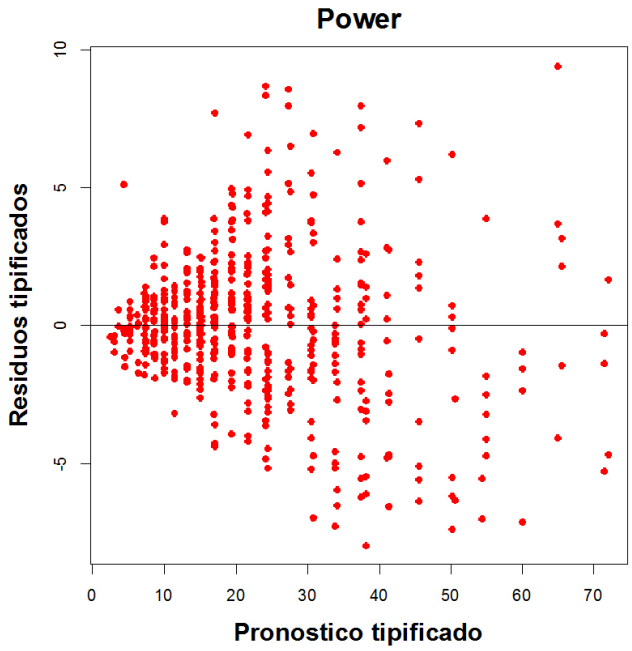
En la Figura XI1.2 se observó que había algunos valores atípicos, los cuales se pueden eliminar fácilmente usando el argumento *outliers*. Como anteriormente vimos que el mejor modelo era el potencial, ahora definimos con el argumento *modelo="Power"*. Además especificamos que se eliminen datos atípicos usando el mismo modelo *outliers="Power"* y como solamente sacamos una gráfica, anulamos el panel con *mfrow=NULL*.

El nuevo modelo que se obtiene (Figura XI1.4) muestra que han desaparecido los valores atípicos y debido a ello el $r^2 = 0,97$, es mayor que en la Figura XI1.2 (r^2 = 0,89) donde los datos atípicos no

habían sido eliminados. La Figura XI1.5 muestra los residuos, que a continuación veremos si son homogéneos a lo largo de todo el rango de valores pronosticados de la variable dependiente.

Figuras XI1.4 y XI1.5 Relación potencial entre ancho de la concha y peso y entre los valores pronosticados y los residuos.





En el archivo TXT que genera la función, se muestra el modelo de regresión potencial, donde se observa que la variable ancho de la concha es significativa (p < 0,001, véase *Pr(>|t|)*) y, por tanto, el modelo en su conjunto también es significativo (p < 0,001, véase *p-value* al final de los resultados). El r^2 (véase *Multiple R-squared*) muestra que el ancho de la concha explica un 97% de la varianza observada en el peso de los individuos. El r^2 ajustado (véase *Ajusted R-squared*) tiene en cuenta el tamaño de la muestra a la hora de determinar la proporción anterior y, en este caso, muestra el mismo valor. El r^2 ajustado se debe usar para comparar modelos con diferente número de observaciones o variables independientes. La ecuación del modelo de regresión potencial se podría expresar de cualquiera de estas dos formas:

$$Peso = 0,436 * Ancho^{3,547} \quad \text{ó} \quad Peso = e^{ln(-0,829)+3,547*ln(Ancho)}$$

```
[1] "POWER REGRESSION"

[[26]]

Call:
lm(formula = fo, data = datos2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40561 -0.07072  0.00126  0.06916  0.78052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.82971    0.03044  -27.26   <2e-16 ***
Ancho        3.54689    0.02835  125.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1205 on 564 degrees of freedom
Multiple R-squared:  0.9652,    Adjusted R-squared:  0.9652
F-statistic: 1.565e+04 on 1 and 564 DF,  p-value: < 2.2e-16



[1] "Normality"

[[28]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.0756, p-value = 3.494e-08


[[29]]
[1] "Autocorrelation"

[[30]]

     Durbin-Watson test

data:  reg
DW = 1.8129, p-value = 0.012
alternative hypothesis: true autocorrelation is greater than 0


[[31]]
[1] "Homocedasticity"

[[32]]

     studentized Breusch-Pagan test

data:  reg
BP = 17.1018, df = 1, p-value = 3.543e-05
```

En la tabla anterior se muestran los resultados del test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors, del test de autocorrelación de Durbin-Watson y del test Breusch-Pagan de homocedasticidad.

Normalidad Los residuos no tienen una distribución Normal con p < 0,001. Aunque no se cumple el supuesto de Normalidad, no invalida el modelo ya que es muy predictivo con un r^2 muy alto. El problema derivado de que los residuos no son Normales, es que no se puede asegurar que el grado de significación, valor de probabilidad que nos muestra el modelo, es el correcto.

Autocorrelación No se cumple el requisito de que no debe existir autocorrelación ya que en la prueba de Durbin-Watson p = 0,012. Esto significa que el valor de r^2 del 97% no es todo debido a la variable independiente, el ancho de la concha, sino que también es en parte debido a la propia

variable dependiente que se auto explica y, por tanto, no podemos saber exactamente cuanto es la varianza explicada por la variable independiente. De todas formas es necesario mencionar que el valor de probabilidad del test de Durbin-Watson puede ser menor de 0,05 fácilmente cuando hay muchos datos, como en este ejemplo. El estadístico DW, cuyo valor es 1,81 en este ejemplo, es un mejor indicador de la autocorrelación cuando el número de datos es muy grande. Según Durbin & Watson (1951), un DW menor que 1 significa una fuerte autocorrelation positiva, un DW mayor que 4 una fuerte autocorrelación negativa, valores entre 1 y 3 una autocorrelación moderada, y un valor próximo a 2 implica que no hay autocorrelación. Por lo tanto, se puede concluir que la autocorrelation en este ejemplo es muy pequeña y la varianza explicada por la variable independiente puede estar cercana al 97%.

Homocedasticidad Por último, no se cumple el requisito de homocedasticidad de los residuos, ya que la probabilidad del test Breusch-Pagan es $p < 0{,}001$. El no cumplirse este requisito significa que el modelo no es igual de predictivo para todo el rango de valores de la variable dependiente. En los residuos del modelo (Figura XI1.5) se observa que a medida que el ancho es mayor, los residuos son más grandes. Por tanto, el modelo predice bien el peso en individuos pequeños, y un poco peor en los moluscos de mayor talla.

Value

Se obtiene un archivo TXT con los resultados de las regresiones y los test de Normalidad, homogeneidad de varianzas y homocedasticidad para cada uno de los modelos de regresión, así como gráficos con la relación entre la variable dependiente e independiente para cada uno de los modelos de regresión, además de gráficos con la relación entre el valor pronosticado y el residuo tipificado.

References

Durbin, J. & Watson G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, **38**, 159-171.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T. et al., (2013) Testing Linear Regression Models R package version 0.9-33. Available at: http://CRAN.R-project.org/package=lmtest.

Examples

```
## Not run:

data(ZXI1)

#Paso 1. Regresión entre el ancho de la concha (variable
#independiente) y el peso (variable dependiente)
#considerando varias especies de moluscos

XI1(data=ZXI1, varInteresX="Ancho",varInteresY="Peso",
XLAB= "Ancho (cm)", YLAB="Peso (gr)")

#Paso 2. La misma relación anterior pero eliminando los datos
#atípicos (outliers) y realizando solo el modelo potencial
```

```
XI1(data=ZXI1, varInteresX="Ancho",varInteresY="Peso",
modelo="Power", outliers="Power", mfrow=NULL,
XLAB= "Ancho (cm)", YLAB="Peso (gr)")


## End(Not run)
```

XI10 *REGRESION DE POISSON-ESTIMACION*

Description

Se aplica una regresión de Poisson que permite buscar el modelo que mejor se ajusta a la posible relación que existe entre el número de veces que ocurre cierto fenómeno aleatorio (variable dependiente) y varias variables independientes.

Usage

```
XI10(data, varInteresY, varInteresX, varplot=NULL, offset=NULL, offsetlog=TRUE,
stepwise=TRUE, ResetPAR=TRUE, PAR=NULL,  YLAB=NULL, XLAB=NULL, PLOT.TS=NULL,
OrderCat=NULL, LabelCat=NULL, COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
TEXT=NULL, file1="Coefficients.csv", file2="Predictions.csv", na="NA", dec=",",
row.names=FALSE, file3="Output.txt")
```

Arguments

data	Archivo de datos.
varInteresY	Variable dependiente.
varInteresX	Variables independientes.
varplot	Vector con tres variables. La primera variable es el eje Y, la segunda el eje X y la tercera es como un factor, de tal forma que se muestra la relación entre la primera y la segunda variable para cada una de las diferentes categorías de la tercera variable de este vector *varplot*.
offset	Se puede especificar una variable de ponderación.
offsetlog	Si es TRUE se transforma logarítmicamente la variable *offset*
stepwise	Si es TRUE se aplica una regresión por pasos para eliminar aquellas variables que no son significativas. Utiliza el criterio de información de Akaike (*Akaike Information Criterion, AIC*)
.	
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
YLAB	Leyenda del eje Y.

`XLAB`	Leyenda del eje X.
`PLOT.TS`	Se accede a la función plot.ts del paquete stats.
`OrderCat`	Permite especificar un vector con el orden en que se muestran las categorías del eje X.
`LabelCat`	Permite especificar un vector con los nombres de las categorías de eje X.
`COLOR`	Colores que se usan en el gráfico de series de tiempo, para identificar las categorías de la variable número tres que se especifica en el vector del argumento *varplot*.
`PCH`	Símbolos que se usan en el gráfico de series de tiempo, para identificar las categorías de la variable número tres que se especifica en el vector del argumento *varplot*.
`LEGEND`	Permite incorporar una leyenda.
`AXIS`	Permite añadir ejes.
`MTEXT`	Permite añadir textos en los márgenes del gráfico.
`TEXT`	Permite añadir texto en cualquier zona de la parte interna del gráfico.
`file1`	ARCHIVOS CSV. Nombre del archivo con los coeficientes de la regresión de Poisson.
`file2`	ARCHIVOS CSV. Nombre del archivo con las predicciones de la regresión de Poisson.
`na`	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.
`file3`	ARCHIVO TXT. Nombre del archivo de salida con los resultados de los análisis.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

XI.1.6. Regresión de Poisson

XI.1.6.1. Estimación del modelo

Se usa para modelar el número de veces que ocurre cierto fenóneno aleatorio. Por ejemplo, el número de casos nuevos de una determinada enfermedad a lo largo de un período de tiempo dado, en condiciones bastante generales se puede modelizar adecuadamente mediante una distribución de Poisson. Basta para ello con el supuesto de que los casos se producen aleatoriamente a lo largo del tiempo y de forma independiente. Esto es válido para muchos otros acontecimientos, como las llamadas a una centralita telefónica, las averías en una máquina, o los accidentes de tráfico, pero en el ámbito epidemiológico el número de casos nuevos se utiliza para medir la tasa de incidencia, por lo que la distribución de Poisson es importante en el estudio de la incidencia y sus causas.

Cuando se utiliza uno o más factores explicativos, que dividen a la población en estratos (fumadores y no fumadores, grupos que reciben medicación diferente, grupos de edad, etc), se suele construir

un modelo para tratar de explicar las diferencias entre grupos, admitiendo que dentro de cada grupo la variabilidad se produce de forma aleatoria.

La distribución de Poisson es sencilla: solo tiene un parámetro, generalmente representado por λ. Este valor es al mismo tiempo la media y la varianza de la variable, por lo que su estimación es muy simple: basta con tomar una muestra representativa y calcular la media. Si en una determinada zona aparecen cada año en promedio 7,2 casos de una enfermedad (valor medio, por ejemplo, de los últimos 10 años), ese es el valor estimado del parámetro lambda.

Si en la muestra observamos que la varianza es muy distinta de la media, ello indicará que la distribución de Poisson no es adecuada para modelizar esa situación, pero en general la regresión de Poisson es apropiada para variables que miden recuentos, en cuyo caso la regresión lineal no es adecuada, ya que la estimación podría producir recuentos negativos, los residuos no tienen distribución Normal, y la varianza de la respuesta suele incrementarse con la media.

En el modelo de regresión de Poisson se supone que la variable dependiente, dados los valores de las variables explicativas, tiene una distribución de Poisson (distribución condicionada por aquellos valores) cuyo parámetro λ tiene el valor:

$$\lambda = e^{P(X_i)}$$

siendo $P(X_i)$ una función lineal de las variables explicativas.

FUNCIONES

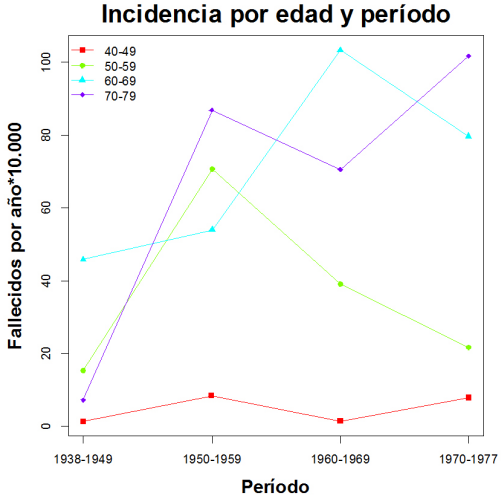
Para estimar la regresión de Poisson se usa la función glm del paquete stats. Para estimar el modelo por pasos se usó la función stepAIC del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2014). El gráfico de series de tiempo se realiza con la función plot.ts del paquete stats. El constraste para determinar si la distribución es de Poisson se realiza con la función *epicalc*.

EJEMPLO

Los datos corresponden a un estudio de cohortes, sobre la exposición al arsénico de la industria y las muertes por enfermedades respiratorias, disponible en la base de datos Montana del antiguo paquete *epicalc* (Chongsuvivatwong, 2012). Los datos son el número de fallecidos en cada grupo (*respdeath*), número de personas por años de exposición (*personyrs*), cociente entre el número de fallecidos y el número de personas por años de exposición * 10.000 (*respdeath.personyrs*), grupos de edad(*agegr*), período del empleo(*period*), comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*).

La variable principal de este estudio es (*respdeath*), número de fallecidos por causa respiratoria, y tratamos de encontrar la relación que pudiera existir entre esa variable y las restantes. Como la variable objetivo es un recuento, puede ser adecuado el modelo de Poisson. Para el modelo de regresión la variable dependiente es el número de fallecidos en cada grupo (*varInteresY=respdeath*). Las independientes son grupo de edad(*agegr*), período del empleo(*period*), comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*), las cuales se definen con el argumento *varInteresX=c("agegr", "period", "start", "arsenic")*. El argumento *offset* permite incluir una constante o una variable en la función. Como a priori sabemos que el número de fallecidos debe ser proporcional en cada grupo al número de personas expuestas, en el ejemplo la variable *personyrs*, incluimos esa variable en el modelo, siendo necesario transformar logarítmicamente la variable para que se corresponda con la transformación general realizada con las restantes variables, para lo cual dejamos el argumento por defecto *offsetlog=TRUE*. Con el argumento *varplot=c("respdeath.personyrs","period", "agegr")* se definen las variables a representar en el gráfico de series de tiempo.

Figura XI11.1. Evolución de la tasa de incidencia (fallecidos dividido por número de personas expuestas) para diferentes períodos de trabajo y distintos grupos de edad.



En el gráfico (Figura XI11.1) se observa que la incidencia, número de fallecidos dividido por el número de personas que estuvieron expuestas, es más alta en los dos grupos de mayor edad.

Los resultados siguientes muestran el modelo obtenido, cuyo $r^2 = 0,8$ y el contraste, cuya hipótesis nula es que la variable sigue una distribución de Poisson, tiene una p = 0,26. Por tanto, se concluye que la distribución es de Poisson.

```
[1] "Poisson regression"

[[2]]

Call:
glm(formula = respdeath ~ agegr + period + start + arsenic, family = poisson,
    data = datos3, offset = personyrs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7604  -0.8329  -0.2377   0.5007   2.1480

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -8.2074     0.3161 -25.965  < 2e-16 ***
agegr50-59          1.6018     0.2749   5.827 5.64e-09 ***
agegr60-69          2.3213     0.2672   8.687  < 2e-16 ***
agegr70-79          2.5090     0.2955   8.491  < 2e-16 ***
period1950-1959     0.4982     0.2326   2.142  0.03221 *
period1960-1969     0.6618     0.2345   2.822  0.00478 **
period1970-1977     0.7288     0.2731   2.669  0.00761 **
startbefore 1925   -0.4972     0.1643  -3.027  0.00247 **
arsenic>15 years    0.9539     0.1859   5.131 2.89e-07 ***
arsenic1-5 years    0.8050     0.1735   4.639 3.50e-06 ***
arsenic5-15 years   0.5486     0.2191   2.504  0.01229 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 340.525  on 95  degrees of freedom
Residual deviance:  92.929  on 85  degrees of freedom
AIC: 301.23
```

La tabla siguiente muestra las predicciones para cada caso que permiten analizar el riesgo relativo.

respdeath	agegr	period	start	arsenic	personyrs	Predicción
19	60-69	1970-1977	before 1925	<1 year	8,54	17,86
14	70-79	1970-1977	before 1925	<1 year	7,48	7,45
13	50-59	1960-1969	before 1925	<1 year	9,13	14,66
12	60-69	1960-1969	before 1925	<1 year	8,32	13,39
10	60-69	1960-1969	1925 or later	<1 year	7,31	8,09
9	60-69	1938-1949	1925 or later	>15 years	6,51	4,85
7	50-59	1950-1959	before 1925	1-5 years	6,84	2,83
7	60-69	1960-1969	before 1925	1-5 years	6,57	5,22
7	60-69	1950-1959	1925 or later	<1 year	7,42	7,66
7	40-49	1960-1969	before 1925	<1 year	9,43	4,00
7	60-69	1950-1959	1925 or later	>15 years	6,09	5,23
6	60-69	1970-1977	before 1925	1-5 years	6,96	8,25

Value

Se obtiene un archivo TXT con los resultados de la regresión, dos archivos CSV con los coeficientes y las predicciones de la regresión, así como un gráfico de series de tiempo si se ha seleccionado.

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI11)

XI10(data = ZXI11, varInteresY = "respdeath", varInteresX = c("agegr", "period",
"start", "arsenic"), offset = "personyrs", varplot = c("respdeath.personyrs",
"period", "agegr"), YLAB = "Fallecidos por año*10.000", XLAB = "Período",
MTEXT = c("text = 'Incidencia por edad y período'", "line = 1", "cex = 2",
"font = 2"))


## End(Not run)
```

XI11 *REGRESION DE POISSON-PREDICCION*

Description

Se aplica un modelo de regresión de Poisson ya calculado con la función XI10.

Usage

```
XI11(data, varInteresY,  varInteresX, offset=NULL,  offsetlog=TRUE,
file="Model predictions.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`varInteresY`	Variable dependiente.
`varInteresX`	Variables independientes.
`offset`	Se puede especificar una variable de ponderación.
`offsetlog`	Si es TRUE se transforma logarítmicamente la variable *offset*.
`file`	ARCHIVOS CSV. Nombre del archivo con las predicciones de la regresión de Poisson.
`na`	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

XI.1.6. Regresión de Poisson

XI.1.6.2. Predicción del modelo

El mejor método para evaluar la calidad del modelo de regresión es comprobar el grado de acierto usando unos datos que no se habían incluido dentro de la matriz original que se usó para estimar el modelo.

Esta función permite de forma fácil aplicar un modelo de regresión de Poisson ya estimado lo cual, además de servir para evaluar la calidad del modelo, también sirve para poder usar el modelo, es decir, pronosticar los valores que se predicen para un determinado caso.

Para poder usar esta función es necesario haber, en primer lugar, calculado el modelo de regresión de Poisson con la función XI10.

EJEMPLO

Los datos deben tener el mismo formato, es decir, las mismas variables y en la misma posición, que la matriz de datos que se usó cuando se estimo el modelo con la función XI10. Además, el directorio de trabajo debe ser el mismo que se usó cuando se estimo el modelo con la función XI10.

Los datos corresponden a un estudio de cohortes, sobre la exposición al arsénico de la industria y las muertes por enfermedades respiratorias, disponible en la base de datos del antiguo paquete *epicalc*. Los datos son el número de fallecidos en cada grupo (*respdeath*), número de personas por años de exposición (*personyrs*), cociente entre el número de fallecidos y el número de personas por años de exposición * 10.000 (*respdeath.personyrs*), grupos de edad(*agegr*), período del empleo(*period*),

comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*). Sin embargo, estos datos no fueron incluidos cuando se estimó el modelo original con la función XI10.

La variable principal de este estudio es (*respdeath*), número de fallecidos por causa respiratoria, y tratamos de encontrar la relación que pudiera existir entre esa variable y las restantes. Como la variable objetivo es un recuento, puede ser adecuado el modelo de Poisson.

Para el modelo de regresión la variable dependiente es el número de fallecidos en cada grupo (*varInteresY=respdeath*). Las independientes son grupo de edad(*agegr*), período del empleo(*period*), comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*), las cuales se definen con el argumento *varInteresX=c("agegr", "period", "start", "arsenic")*.

El argumento *offset* permite incluir una constante o una variable en la función. Como a priori sabemos que el número de fallecidos debe ser proporcional en cada grupo al número de personas expuestas, en el ejemplo la variable *personyrs*, incluimos esa variable en el modelo, siendo necesario transformar logarítmicamente la variable para que se corresponda con la transformación general realizada con las restantes variables, para lo cual dejamos el argumento por defecto *offsetlog=TRUE*.

Los resultados muestran que $r^2 = 0,84$. Por tanto, parece ser que el modelo obtenido es bastante fiable.

Value

Se obtiene un archivo CSV con las predicciones del modelo.

Examples

```
## Not run:

data(ZXI12)

XI11(data=ZXI12, varInteresY="respdeath", varInteresX=c("agegr", "period", "start",
"arsenic"), offset="personyrs")


## End(Not run)
```

XI2 *FUNCIONES NO LINEALES*

Description

Se aplican diferentes modelos para describir el tipo de función no lineal que mejor se ajusta a la posible relación que existe entre dos variables.

Usage

```
XI2(data, varInteresY, varInteresX, modelo=c("Clench", "Nexponential",
"Saturation", "Rational", "Bertalanffy", "Gompertz", "Logistic"),
outliers=NULL, quant1=0.05, quant2 = 0.95, ResetPAR=TRUE, mfrow=c(2,4),
PAR=NULL, PLOT=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1.3, COLABLINE="#48D1CCFF",
```

```
COLOR="#C0FF3EFF", PCH=15, resPlot=TRUE, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file1="Output.txt", file2="Coefficients.csv", file3="Residuals.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
varInteresY	Variable dependiente.
varInteresX	Variable independiente.
modelo	Se puede elegir uno, o varios modelos de regresión: "Clench", "Nexponential", "Saturation", "Rational", "Bertalanffy", "Gompertz" y/o "Logistic".
outliers	Si es NULL no se eliminan los valores atípicos (*outliers*). Al elegir algún modelo de regresión, "Clench", "Nexponential", "Saturation", "Rational", "Bertalanffy", "Gompertz" o "Logistic", entonces se eliminan los valores atípicos usando el modelo de regresión seleccionado (para más detalles véase la sección *details*).
quant1	Cuantil del extremo inferior para la eliminación de los datos atípicos.
quant2	Cuantil del extremo superior para la eliminación de los datos atípicos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
mfrow	Si es NULL y hay varias funciones seleccionadas, éstas salen en ventanas independientes. Si se quiere que los gráficos salgan en paneles, este argumento es un vector con el formato c(nr, nc) que indica el número de figuras por fila (nr) y por columna (nc), rellenando primero las filas.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
PLOT	Se accede a la función plot.default que permite modificar muchos aspectos diferentes del gráfico plot.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
CEXPCH	Tamaño de los símbolos del gráfico.
COLABLINE	Color de la línea del modelo de regresión.
COLOR	Color de los símbolos.
PCH	Tipo de símbolo.
resPlot	Si es TRUE se muestran los gráficos que representan la relación entre el valor pronosticado y el residuo tipificado.
LEGEND	Permite añadir una leyenda al gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
AXIS	Permite añadir ejes al gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
MTEXT	Permite añadir textos en los márgenes del gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.

TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico. Solo tiene sentido si se hace un solo gráfico, ya que se selecciona un único modelo de regresión.
file1	ARCHIVO TXT. Nombre del archivo de salida con los resultados de los análisis.
file2	ARCHIVOS CSV. Nombre del archivo con los coeficientes de los modelos de regresión.
file3	ARCHIVOS CSV. Nombre del archivo con los residuos de los modelos de regresión.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

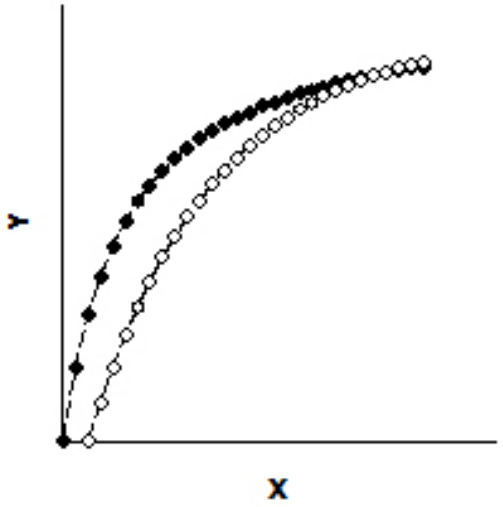
XI.1.3. Funciones no lineales

Muchas veces las variables se ajustan a un tipo de función no lineal que no es ninguna de las comunes que vienen descritas en la mayoría de los programas estadísticos y se vieron en la función XI1. A continuación se describen algunas de estas funciones no lineales y como calcularlas.

XI.1.3.1. Funciones de acumulación

La tasa de crecimiento de una población bacteriana, de levaduras o fitoplanctónica a distintas concentraciones de nutrientes o de temperatura, la tasa fotosintética de plantas a diferentes intensidades de luz, la riqueza acumulada en función del número de muestreos realizados, entre otros muchos ejemplos, se ajustan a las llamadas funciones de acumulación (Figura XI2.1). Existen diferentes tipos pero se pueden dividir entre aquellas en las que el cero de Y y el cero de X siempre coinciden y las que permiten un valor diferente de cero para Y cuando X es cero (Figura XI2.1).

Figura XI2.1. Curvas de acumulación en las que el cero de Y y el cero de X siempre coinciden (círculos negros) y curvas que permiten un valor diferente de cero para X cuando Y es cero (círculos blancos).



En este *script* se calcula: 1) la curva de Clench (Clench, 1979), la cual es una modificación de la función de Monod (Monod, 1950), y se propuso para mariposas; 2) la exponencial negativa (Miller & Wiegert, 1989) que se propuso para especies de plantas raras; 3) la curva de saturación que se utilizó para mostrar la relación entre crecimiento de un alga tóxica fitoplanctónica del género *Alexandrium* y la concentración de fosfato (Frangópulos et al., 2004), la cual es identica a la función de crecimiento de von Bertalanffy que veremos a continuación, pero adaptando los coeficientes para explicar mejor el patrón de una función de acumulación y, por último; 4) la función Racional (Ratkowski, 1990) que se puede usar cuando no existe un criterio claro de que modelo usar (Falther, 1996).

Nombre	**Función**	**Referencia**
Clench	$y = \frac{ax}{1+bx}$	(Clench, 1979)
Exponencial negativa	$y = a\left(1 - e^{-bx}\right)$	(Miller & Wiegert, 1989)
Saturación	$y = a\left(1 - e^{-b(x-c)}\right)$	(Frangópulos et al., 2004)
Racional	$y = \frac{(a+bx)}{(1+cx)}$	(Ratkowski, 1990)

XI.1.3.2. Funciones de crecimiento

Las curvas de crecimiento se aplican generalmente al crecimiento de los individuos aunque la función de von Bertalanffy se puede ajustar también al crecimiento de poblaciones. Es por ello que anteriormente se mostró la curva de Saturación, cuya fórmula es identica a la de von Bertalanffy, pero con sus coeficientes modificados.

XI.1.3.2.1. Curva de crecimiento de von Bertalanffy

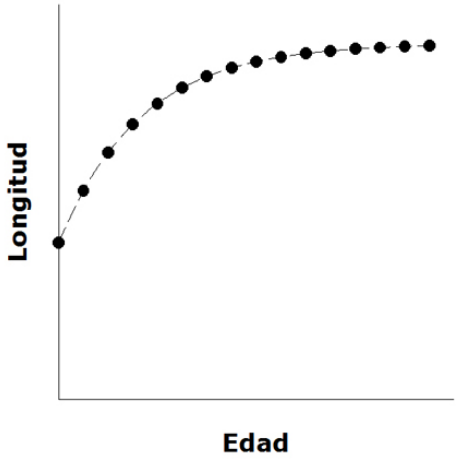
La curva de crecimiento de muchos organismos se ajusta bien a la ecuación de von Bertalanffy (Figura XI2.2):

$$L_t = L_\infty \left(1 - e^{-k(t-t_0)}\right)$$

donde L_t es la longitud del individuo a la edad *t*, L_∞ es la longitud máxima que alcanza el individuo cuando el crecimiento cesa, *k* es la constante de crecimiento expresada en $tiempo^{-1}$ y t_0 es la edad

hipotética que tendría un individuo que tuviese un tamaño cero. Por lo tanto, la variable dependiente es L_t, la variable independiente es la edad del individuo (*t*) y las constantes de esta ecuación son *k*, L_∞ y t_0.

Figura XI2.2. Curva de crecimiento de von Bertalanffy



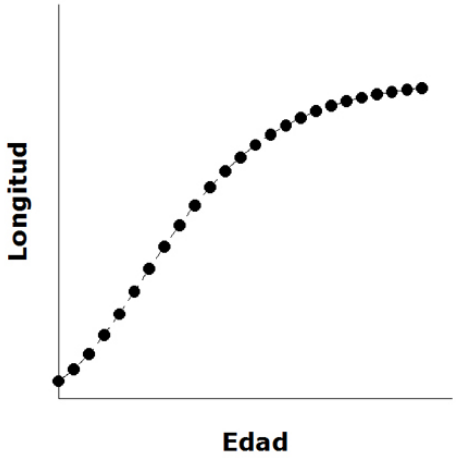
XI.1.3.2.2. Curva de crecimiento de Gompertz

En algunas especies, como por ejemplo los peces, crustáceos y moluscos, en las primeras fases del ciclo de vida (fase larvaria) el crecimiento puede ajustarse mejor a la ecuación de Gompertz (Hernandez-Llamas & Ratkowsky, 2004) (Figura XI2.3):

$$L_t = ae^{-be^{-ct}}$$

donde L_t es la longitud del individuo a la edad *t*, *a* es una constante que representa el valor máximo asintótico de la curva, *t* es la edad del individuo y, *b* y *c* son las otras dos constantes de la ecuación.

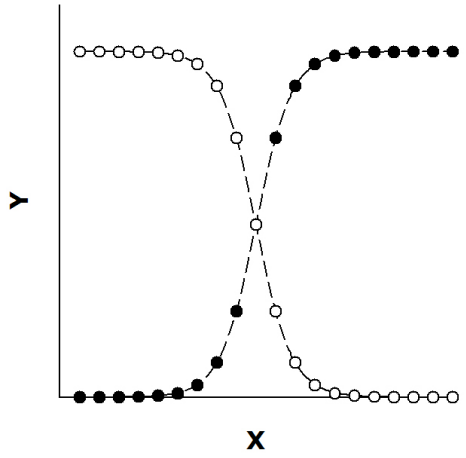
Figura XI2.3. Curva de crecimiento de Gompertz



XI.1.3.2. Función logística

Muchos fenómenos en la naturaleza se ajustan a una curva logística (Figura XI2.4), como por ejemplo los cambios en la abundancia de una población en el tiempo (Smith & Smith, 2000), el avance científico a lo largo del tiempo dentro de una determinada línea de investigación (Solla Price, Little Science, Big Science, 1963 -citado en Callon et al., 1995), etc.

Figura XI2.4. Función logística con pendiente negativa (círculos blancos) y positiva (círculos negros).



La curva logística viene definida por la siguiente ecuación:

$$Y = \frac{a}{1 + e^{(b-cX)}}$$

donde la constante *a* establece el límite superior de la curva y es igual a *K* o capacidad de carga de la población (número máximo de individuos de esa población), en el caso de que se modele a una población. La constante *c* determina la pendiente, y en el caso de poblaciones es igual a *r* o tasa de crecimiento de la población. La constante *b* define el tamaño de la llamada fase de latencia en el

caso de poblaciones, la fase con valores más pequeños antes de iniciar la pendiente hasta alcanzar el valor máximo.

FUNCIONES

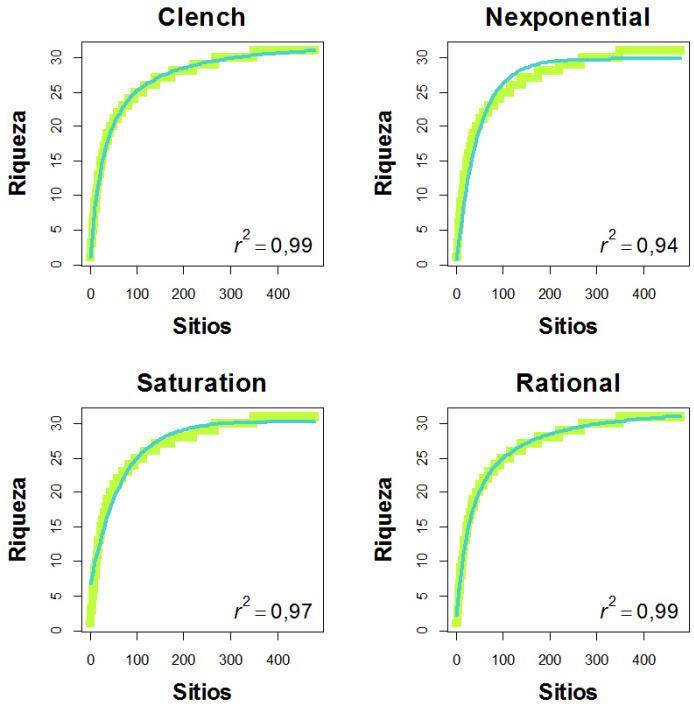
Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors y la función nls del paquete base stats para estimar los modelos.

EJEMPLO 1

Se realizaron muestreos en varios sitios y se registró el número de especies presentes. Se hizo una relación entre el número de sitios muestreados y la riqueza acumulada considerando todos los sitios muestreados anteriormente. El objetivo es determinar el número de especies totales teórico que existe en la zona.

La Figura XI2.5 muestra las funciones obtenidas, donde se observa que las funciones de Clench y Racional son las que tienen los r^2 mayores.

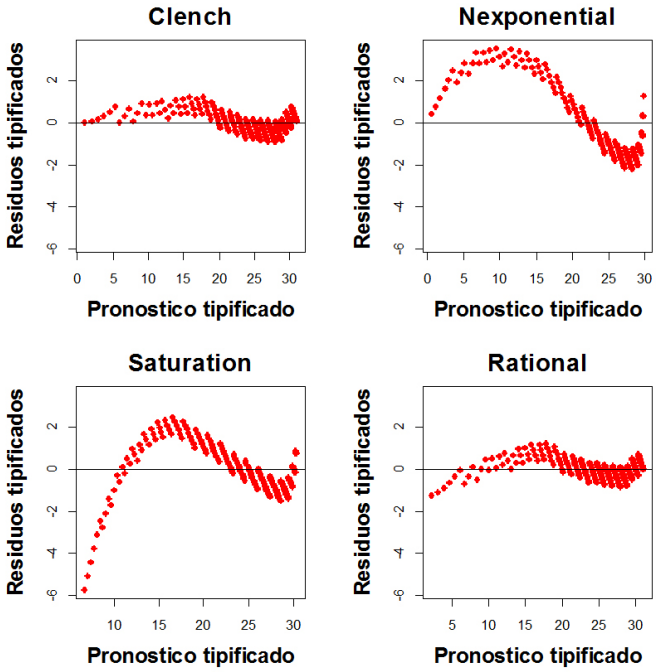
Figura XI2.5. Relación entre el número de sitios muestreados y la riqueza de especies acumulada considerando todos los sitios muestreados anteriormente.



La Figura XI2.6 muestra que las funciones de Clench y Racional son las que también tienen los residuos más pequeños, tienen un comportamiento casi lineal y son iguales a lo largo de todo el rango de valores pronosticados de la variable dependiente. Por tanto, ambas funciones son las que dan un mejor ajuste.

Figura XI2.6. Relación entre los valores pronosticados y los residuos

tipificados de todas las funciones de acumulación.



A continuación se muestra la función de Clench y el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors. Los residuos tienen una distribución Normal con p = 0,26.

```
[1] "CLENCH"

[[2]]

Formula: Riqueza ~ a * Sitios/(1 + b * Sitios)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 1.0499175  0.0072771   144.3   <2e-16 ***
b 0.0318252  0.0002529   125.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.437 on 477 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 5.778e-06


[[3]]
[1] "Normality"

[[4]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.032316, p-value = 0.2601
```

La ecuación quedaría como se muestra a continuación, de tal forma que considerando 500 sitios

muestreados la riqueza teórica total de la zona sería de 31 especies.

$$Riqueza = \frac{1,05 * Sitios}{1 + 0,032 * Sitios}$$

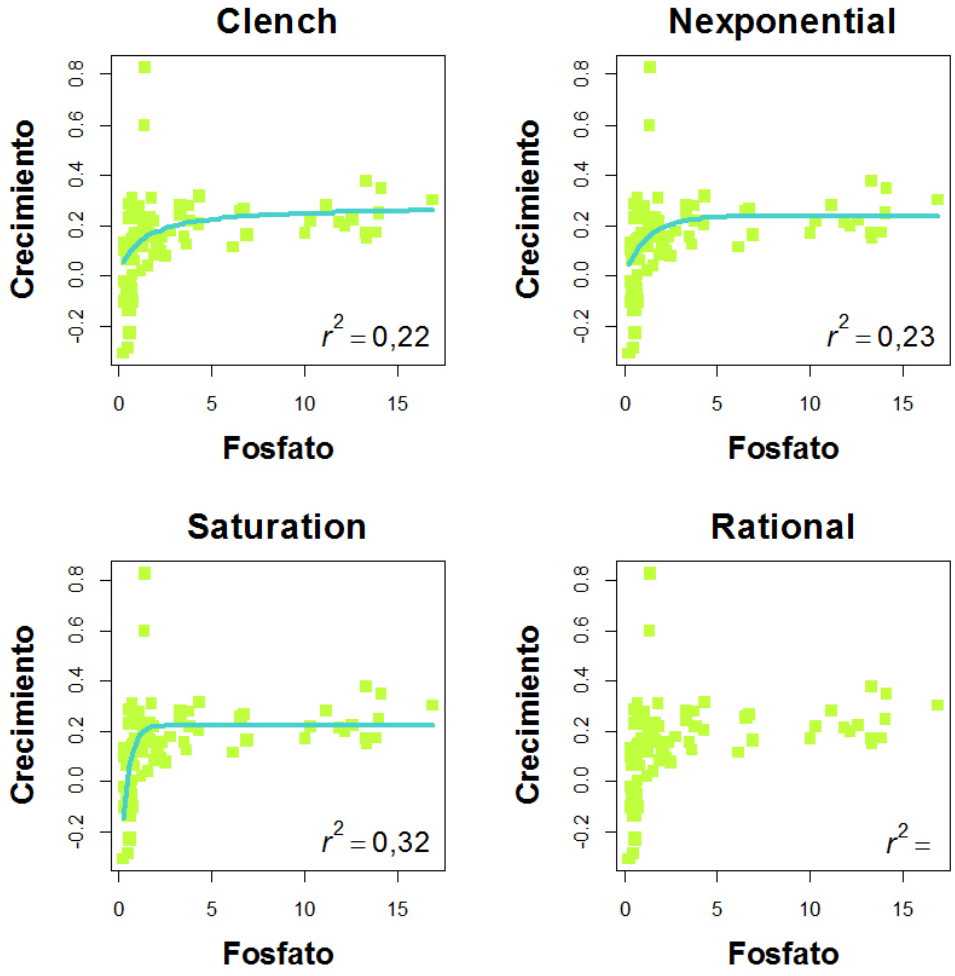
EJEMPLO 2

Paso 1.

El estudio consistió en cultivar especies del género *Alexandrium* a distintas concentraciones de fosfato. El objetivo era determinar la concentración mínima de fosfato necesaria para que exista crecimiento y la tasa máxima de crecimiento de la población.

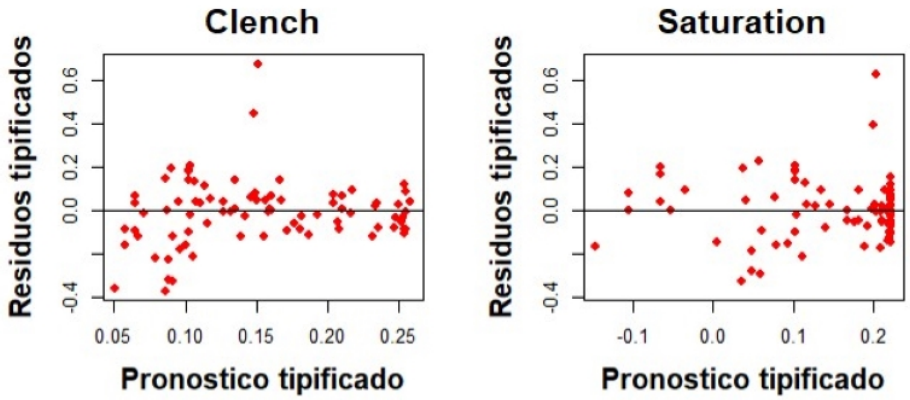
La Figura XI2.7 muestra las funciones obtenidas, donde se observa que la función de Saturación es la que tiene un r^2 mayor. También se observa que no fue posible encontrar un ajuste para la función Racional, lo cual puede ocurrir para cualquier función dependiendo del tipo de datos.

Figura XI2.7. Relación entre la concentración de fosfato y la tasa de crecimiento diaria de población de especies del género *Alexandrium*



La Figura XI2.8 muestra que para las tres funciones los residuos son pequeños, tienen un comportamiento casi lineal y son iguales a lo largo de todo el rango de valores pronosticados de la variable dependiente. Sin embargo, parece que existen valores atípicos.

Figura XI2.8. Relación entre los valores pronosticados y los residuos tipificados de todas las funciones de acumulación.

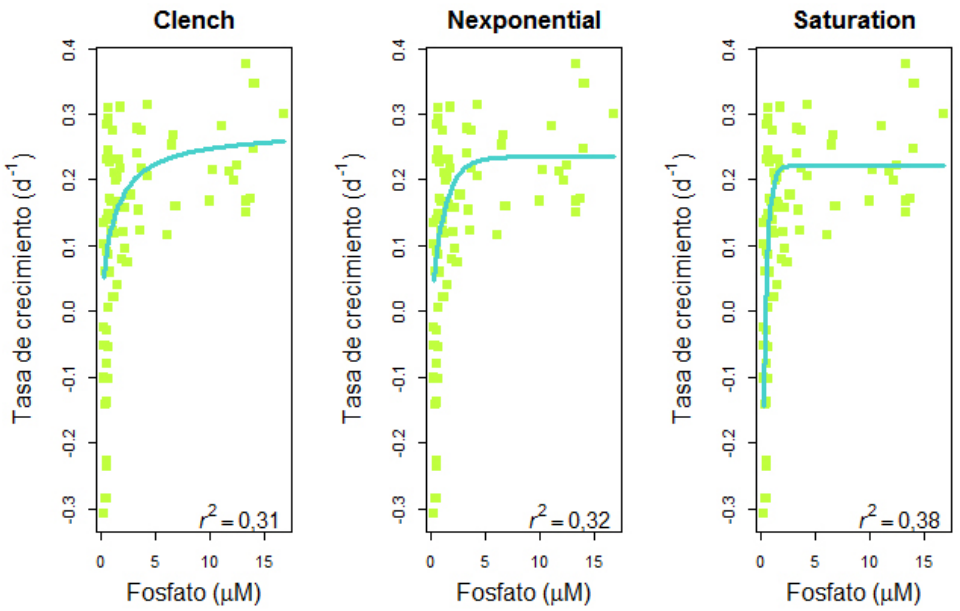


Paso 2.

En la Figura XI2.8 se observó que había algunos valores atípicos, los cuales se pueden eliminar fácilmente usando el argumento *outliers*. Como anteriormente vimos que el mejor modelo era la función de Saturación, ahora especificamos que se eliminen datos atípicos usando el argumento *outliers="Saturación"*. Además, como no era posible el ajuste a la función Racional, la eliminamos del argumento *modelo=c("Clench","Nexponential", "Saturation")*. Por último, como solamente tenemos tres gráficos, utilizamos el argumento *mfrow=c(1,3)*, para que los represente en una fila y tres columnas.

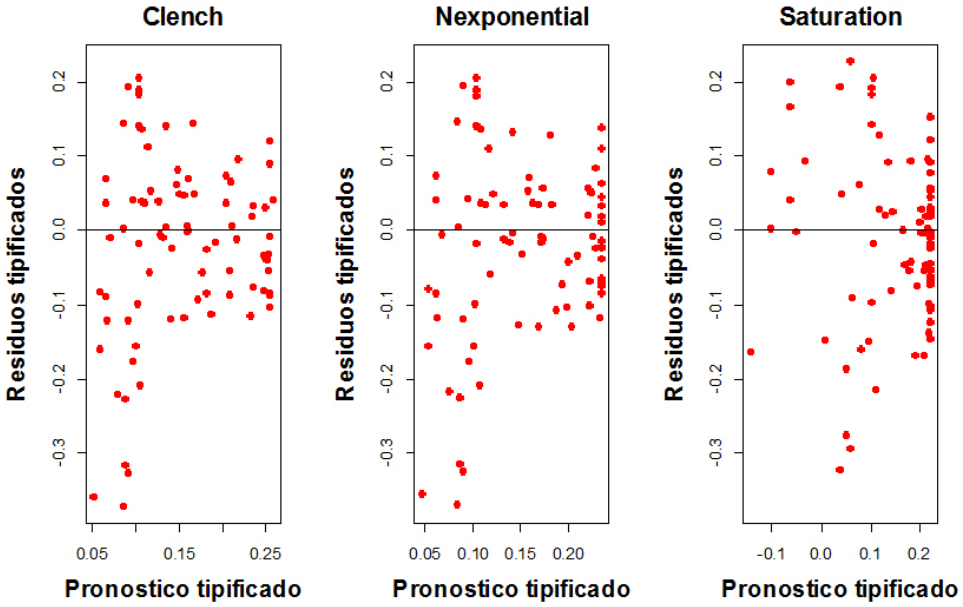
La Figura XI2.9 muestra las funciones obtenidas, donde se observa que la función de Saturación es la que sigue teniendo un r^2 mayor.

Figura XI2.9. Relación entre la concentración de fosfato y la tasa de crecimiento diaria de población de especies del género *Alexandrium*



La Figura XI2.10 muestra que para las tres funciones los residuos son homogéneos y, además, se eliminaron los datos atípicos.

Figura XI2.10. Relación entre los valores pronosticados y los residuos tipificados de todas las funciones de acumulación.



A continuación se muestra la función de Saturación y el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors, que muestra que los residuos tienen una distribución Normal con p = 0,801.

```
[1] "SATURATION"

[[10]]

Formula: Crecimiento ~ a * (1 - exp(-b * (Fosfato - c)))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  0.22210    0.02339   9.496 8.34e-15 ***
b  2.55269    0.86471   2.952  0.00413 **
c  0.46778    0.05011   9.335 1.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.142 on 81 degrees of freedom

Number of iterations to convergence: 13
Achieved convergence tolerance: 3.781e-06


[[11]]
[1] "Normality"

[[12]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.0541, p-value = 0.8012
```

La ecuación quedaría como se muestra a continuación, de tal forma que la tasa de crecimiento máxima es 0,222 d^{-1} y 0,468 μM es el umbral mínimo de concentración de fosfato que necesita la especie para crecer, ya que a esa concentración de fosfato el crecimiento es cero.

$$y = 0,222\left(1 - e^{-2,553(Fosfato - 0,468)}\right)$$

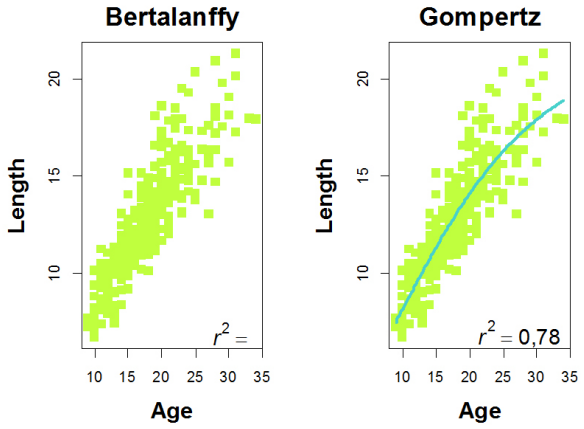
EJEMPLO 3

Un estudio realizado con la anchoa (*Engraulis encrasicolus*), en el Mar de Alborán y en el estrecho de Sicilia, consistió en analizar la relación entre la longitud de los individuos y su edad, la cual se cuantificó por medio del recuento de los anillos del otolito (Basilone et al., 2004). El objetivo era comparar el crecimiento en ambas zonas y determinar la longitud máxima. Para ello, se intentó encontrar la función de crecimiento que mejor se ajustaba a los datos.

Paso. 1

En primer lugar se seleccionaron las dos funciones de crecimiento con el argumento *modelo=c("Bertalanffy", "Gompertz")*. La Figura XI2.11 muestra que solo se pudo encontrar un ajuste con la función de Gompertz.

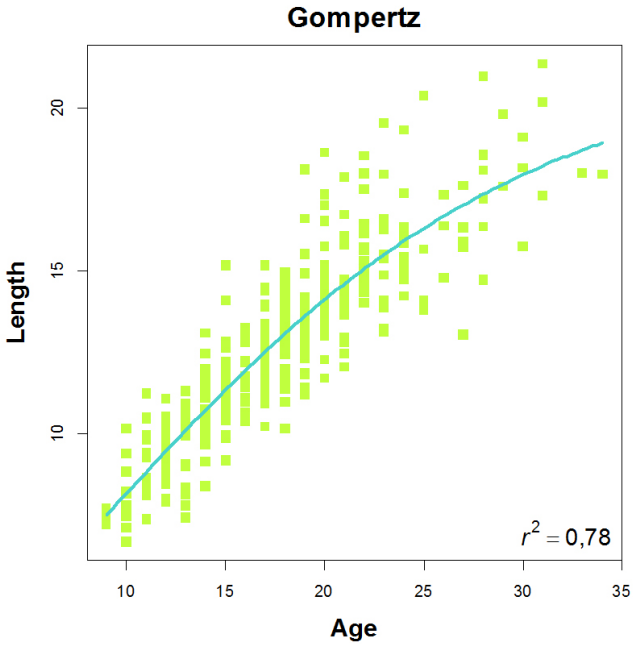
Figura XI2.11. Relación entre la longitud y la edad en la anchoa (*Engraulis encrasicolus*), en el Mar de Alborán y en el estrecho de Sicilia.



Paso. 2

A continuación se realizó el modelo solo con *modelo="Gompertz"* y para eliminar posibles valores atípicos se usó el argumento *outliers="Gompertz"*. La Figura XI2.12 muestra la función de Gompertz.

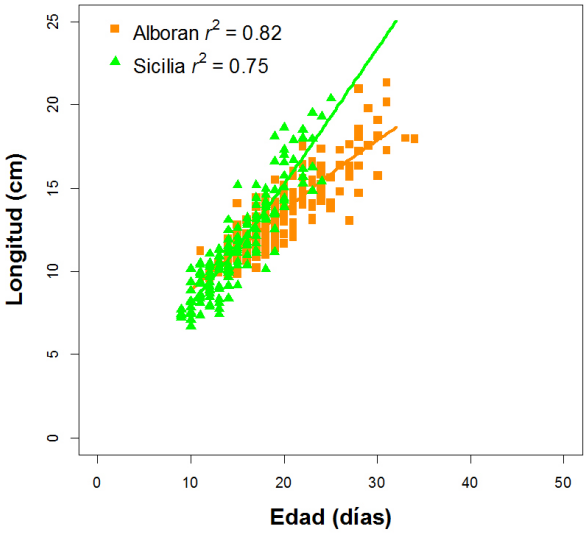
Figura XI2.12. Relación entre la longitud y la edad en la anchoa (*Engraulis encrasicolus*), en el Mar de Alborán y en el estrecho de Sicilia.



Paso. 3

Por último, se realizó un script para poder mostrar la tendencia en ambas zonas. Con la función *subset* se selecciona primero los datos del Mar de Alborán y se realiza el primer gráfico, y con el segundo *subset* se selecciona luego los datos de Sicilia. Con la función *seq* se genera una secuencia de 10 a 32 con 100 divisiones, que permite luego obtener los valores de *Longitud* aplicando la función de Gompertz correspondiente a los datos de cada una de las áreas. Con la función *lines* se representa la linea de tendencia. Para poder tener los coeficientes de las funciones de cada una de las zonas, en el paso 2 se correría el script, pero haciendo por separado el Mar del Alborán y el estrecho de Sicilia. Por último, con la función *legend* se ponen las leyendas. La gráfica resultante se muestra en la Figura XI2.13.

Figura XI2.13. Relación entre la longitud y la edad en la anchoa (*Engraulis encrasicolus*), en el Mar de Alborán y en el estrecho de Sicilia.



Las ecuaciones obtenidas para cada una de las zonas son las siguientes:

Zona	**Función**
Mar de Alborán	$Longitud = 28,994 * e^{-1,8457 * e^{(-0,044718 * Edad)}}$
Sicilia	$Longitud = 57,613 * e^{-2,8974 * e^{(-0,038967 * Edad)}}$

EJEMPLO 4

El crecimiento de una población de fitoplancton sigue normalmente una curva de crecimiento logístico. El objetivo es estimar la curva logística de crecimiento, para determinar la abundancia máxima que puede alcanzar la especie, bajo esas condiciones de cultivo.

La Figura XI2.14 muestra que el ajuste a la función logística es bueno, con un $r^2 = 0,98$, y los residuos son homogéneos a los largo de todo el rango de valores pronosticados (Figura XI2.15).

Figura XI2.14. Evolución de la abundancia de una especie del fitoplancton en el tiempo

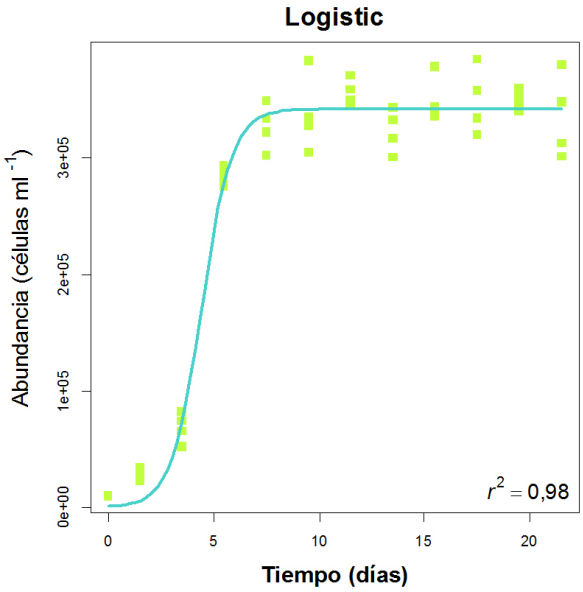
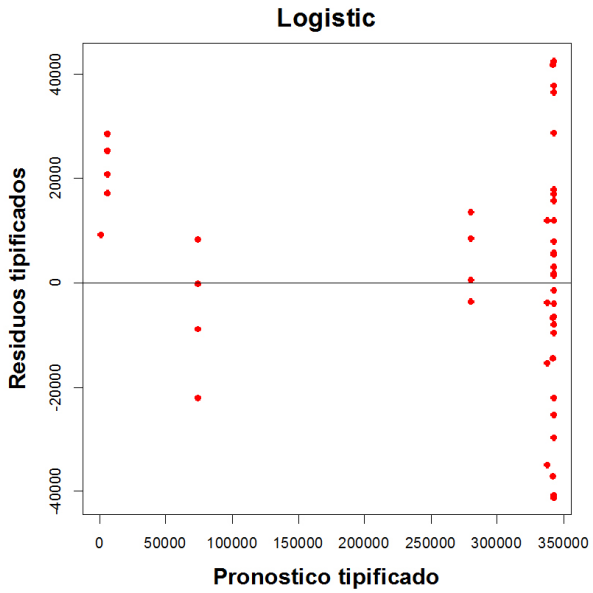


Figura XI2.15. Relación entre los valores pronosticados y los residuos tipificados de la función mostrada en la Figura XI2.14.



Los resultados muestran que los residuos son normales con una p = 0,734 y la curva logística viene definida por la siguiente función, donde se observa que la abundancia máxima es de $3,422 * 10^5$ células por ml:

$$Abundancia = \frac{3,422 * 10^5}{1 + e^{(6,152 - 1,391 * Tiempo)}}$$

```
Formula: Abundance ~ a/(1 + exp(b - c * Day))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 3.422e+05  3.853e+03   88.82  < 2e-16 ***
b 6.152e+00  6.080e-01   10.12 3.59e-13 ***
c 1.391e+00  1.373e-01   10.13 3.46e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21360 on 45 degrees of freedom

Number of iterations to convergence: 15
Achieved convergence tolerance: 4.519e-06


[[27]]
[1] "Normality"

[[28]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.074, p-value = 0.7341
```

Value

Se obtiene un archivo TXT con los resultados de las funciones y el test de Normalidad para cada una de ellas, así como gráficos con la relación entre la variable dependiente e independiente, además de gráficos con la relación entre el valor pronosticado y el residuo tipificado para cada función.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A. & Kallianiotis, A. (2004). Linking habitat conditions and growth in the European anchovy (*Engraulis encrasicolus*). *Fisheries Research*, 68: 9-19.

Callon, M., Courtial, J.P. & Penan, H. (1995) *Cienciometría: La medición de la actividad científica, de la bibliometría a la vigilancia tecnológica*. Trea, D.L., Gijón.

Clench, H.K. (1979) How to make regional lists of butterflies: some invoking empirically based criteria in selecting among thoughts. *The Journal of the Lepidopterists' Society*, 33: 216-231.

Flather, C.H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, 23: 155-168.

Frangópulos, M., Guisande, C., deBlas, E. y Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae*, 3: 131-139.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hernández-Llamas, A. & Ratkowsky, D.A. (2004) Growth of fishes, crustaceans and molluscs: estimation of the von Bertalanffy, Logistic, Gompertz and Richards curves and a new growth model. *Marine Ecology Progress Series* 282: 237-244.

Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, 70: 16-22.

Monod, J. (1950) La technique de culture continue, théorie et applications. *Annales de l'Institut Pasteur (Paris)*, 79: 390-410.

Ratkowski, D.A. (1990) *Handbook of nonlinear regression models*. Marcel Dekker, New York, 241 pp.

Smith, R.L. & Smith, T.M. (2000) *Ecología*. Pearson Educación, S.A., Madrid.

Examples

```
## Not run:

data(ZXI2)

#EJEMPLO 1. Riqueza acumulada y sitios de muestreo

XI2(data=ZXI2, varInteresX="Sitios",varInteresY="Riqueza",
modelo=c("Clench","Nexponential", "Saturation","Rational"), mfrow=c(2,2))

#EJEMPLO 2. Tasa de crecimiento de un alga tóxica en función
#de la concentración de fosfato

data(ZXI3)

#Paso 1.
XI2(data=ZXI3, varInteresX="Fosfato",varInteresY="Crecimiento",
modelo=c("Clench","Nexponential", "Saturation","Rational"),
mfrow=c(2,2), YLAB=expression(paste("Tasa de crecimiento (d "^"-1",")")),
XLAB=expression(paste("Fosfato (",mu,"M)")))

#Paso 2.
XI2(data=ZXI3, varInteresX="Fosfato",varInteresY="Crecimiento",
modelo=c("Clench","Nexponential", "Saturation"),
outliers="Saturation",  mfrow=c(1,3),
YLAB=expression(paste("Tasa de crecimiento (d "^"-1",")")),
XLAB=expression(paste("Fosfato (",mu,"M)")))


#EJEMPLO 3. Relación entre la edad y la longitud en la anchoa

data(ZXI4)

#Paso 1.
XI2(data=ZXI4, varInteresX="Age",varInteresY="Length",
modelo=c("Bertalanffy", "Gompertz"), mfrow=c(1,2))

#Paso 2.

XI2(data=ZXI4, varInteresX="Age",varInteresY="Length",modelo="Gompertz",
outliers=c("Gompertz"), mfrow=NULL)

#Paso 3

data(ZXI4)

#Se selecciona Alboran
```

```
data1<-subset(ZXI4,ZXI4[,"Area"]
plot( formula = Length~Age, xlim = c(0,50), ylim = c(0,25),
xlab="Edad (días)", ylab="Longitud (cm)", font.lab=2,
cex.lab=1.5, col = "#FF8C00FF" , pch = 15 , cex = 1.2 ,
data = data1 )

#Linea de la función
cx<-seq(10,32, length.out = 100)
cy<-28.994*exp(-1.8457*exp(-0.044718*cx))
lines(cx,cy, col="#FF8C00FF",lwd=3 )

#Se selecciona Sicilia
data2<-subset(ZXI4,ZXI4[,"Area"]
points(formula=Length~Age , col="#00FF00FF" , data=data2,
pch=17 , cex=1.2)

#Linea de la función
cx<-seq(10,32, length.out = 100)
cy<-57.613*exp(-2.8974*exp(-0.038967*cx))
lines(cx,cy,  col="#00FF00FF",   lwd=3 )

#Leyendas
legend(x=0, y= 26, legend=expression(paste("Alboran ",
italic("r")^"2"," = 0.82")),pch=c(15), bty="n", cex=1.3,
col="#FF8C00FF")
legend(x=0, y= 24, legend=expression(paste("Sicilia ",
italic("r")^"2"," = 0.75")),pch=c(17), bty="n", cex=1.3,
col="#00FF00FF")

#EJEMPLO 4. Evolución de la abundancia de un cultivo de
#una especie de fitoplancton en el tiempo

data(ZXI5)

XI2(data=ZXI5, varInteresX="Day",varInteresY="Abundance",
modelo=c("Logistic"), mfrow=NULL,
YLAB=expression(paste("Abundancia (células ml"^"-1",")")),
XLAB="Tiempo (días)")


## End(Not run)
```

XI3 *SELECCION DE VARIABLES NO CORRELACIONADAS USANDO EL VIF*

Description

Se utiliza el factor de inflación de la varianza (VIF) para seleccionar las variables no correlacionadas, a partir de una matriz inicial con múltiples variables.

Usage

```
XI3(data, variables, method="vifstep", threshold=10, varCode=NULL, file1="VIF.csv",
file2="Variables.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
variables	Variables a partir de las cuales se quiere seleccionar aquellas no correlacionadas.
method	Hay tres métodos para seleccionar las variables no correlacionadas: "vif", "vifcor" y "vifstep" (véase la sección *details* para ampliar la información). El método "vif" no se pone por defecto, a pesar de ser la opción aconsejable, ya que solamente funciona con RWizard y, por tanto, no funciona en otras plataformas, como por ejemplo RGUI.
threshold	Valor de corte para el VIF.
varCode	Opcionalmente se puede seleccionar variables de la matriz original, que se exportarán en el archivo de salida con las variables no correlacionadas. Por ejemplo, permite seleccionar variables que son códigos de las filas.
file1	ARCHIVOS CSV. Nombre del archivo con los valores VIF para todas las variables.
file2	ARCHIVOS CSV. Nombre del archivo donde se exportan las variables no correlacionadas.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

XI.1.4. Selección de variables usando el factor de inflación de la varianza (VIF)

XI.1.4.1. Seleción de variables no correlacionadas

En muchos modelos estadísticos multivariados es necesario utilizar variables que no estén correlacionadas. Con esta función es fácil generar un archivo con las variables no correlacionadas a partir de un conjunto de variables inicial. El método de selección se basa en el Factor de Inflación de la Varianza (VIF).

Una variable es redundante si está muy correlacionada con otra u otras, en cuyo caso contiene prácticamente la misma información y debe ser excluida. Las variables redundantes en un modelo producen el efecto de aumentar la varianza de las estimaciones de los coeficientes (los coeficientes son más inestables), distorsionando así los intervalos de confianza para esos coeficientes y los contrastes de significación (puede aparecer como no significativo un coeficiente que en realidad si lo

es). El problema se conoce como "multicolinealidad", los regresores están relacionados entre sí y no son independientes.

En general un cierto grado de multicolinealidad es inevitable, y debe aceptarse como tal. Únicamente cuando la multicolinealidad es grave deben tomarse acciones eliminando del modelo las variables que la causan.

El método más utilizado para detectar multicolinealidad y variables redundantes es el que utiliza el llamado "factor de inflación de la varianza" VIF: $VIF(x) = \frac{1}{(1-R^2)}$ siendo R^2 el coeficiente de determinación de la regresión de la variable *x* con las restantes (cuadrado del coeficiente de correlación múltiple). Como $1 - R^2$ representa la fracción de varianza de *x* no explicada por las demás variables, su inversa indica el grado de información redundante: un VIF = 40 significa que la información de esa variable ya contenida en otras variables es 40 veces mayor que la información distinta o nueva que aporta. El valor mínimo de VIF es 1, y no tiene máximo (puede ser arbitrariamente grande).

Se suelen utilizar como valores de referencia de VIF los siguientes: VIF < 10 leve; 10 < VIF < 30 moderado; VIF > 30 grave. Es aconsejable suprimir por redundantes las variables con VIF > 30.

Como suprimir variables puede hacer que el modelo sea sesgado o incompleto, generalmente se recomienda prudencia al hacerlo: es preferible mantener una variable que aporta información (a costa de aumentar la variabilidad o inestabilidad) que suprimirla, ya que el efecto podría ser más grave, sobre todo si la muestra es grande, ya que el tamaño de muestra corrige el problema haciendo que se reduzca la variabilidad de las estimaciones.

En la función se puede elegir entre tres métodos diferentes para seleccionar las variables no correlacionadas. Si se selecciona "vif" sale una ventana interactiva donde se va informando de las variables con VIF por encima del umbral indicado por el usuario en el argumento *threshold*, y el usuario es el que va decidiendo que variables elimina. La ventaja del método manual es que el usuario toma las decisiones de las variables a eliminar entre los pares correlacionados. En las otras dos opciones "vifcor" y "vifstep" se hace una selección automática de las variables a excluir, lo cual es lógicamente mucho más rápido y cómodo, pero es siempre preferible usar la opción manual, para que el usuario siempre tenga el control a la hora de elegir que variable se selecciona entre dos correlacionadas. Si se selecciona "vifcor", primero busca el par de variables que tiene la correlación lineal máxima, y elimina la que tenga el mayor VIF, repitiendo el proceso hasta que no quede ninguna variable con un alto coeficiente de correlación con otra variable. Si selecciona "vifstep" se calcula el VIF para todas las variables, y se excluye la que tenga un VIF mayor, siempre considerando el umbral definido por el usuario en el argumento *threshold*, repitiendo el proceso hasta que ninguna variable esté correlacionada con otra.

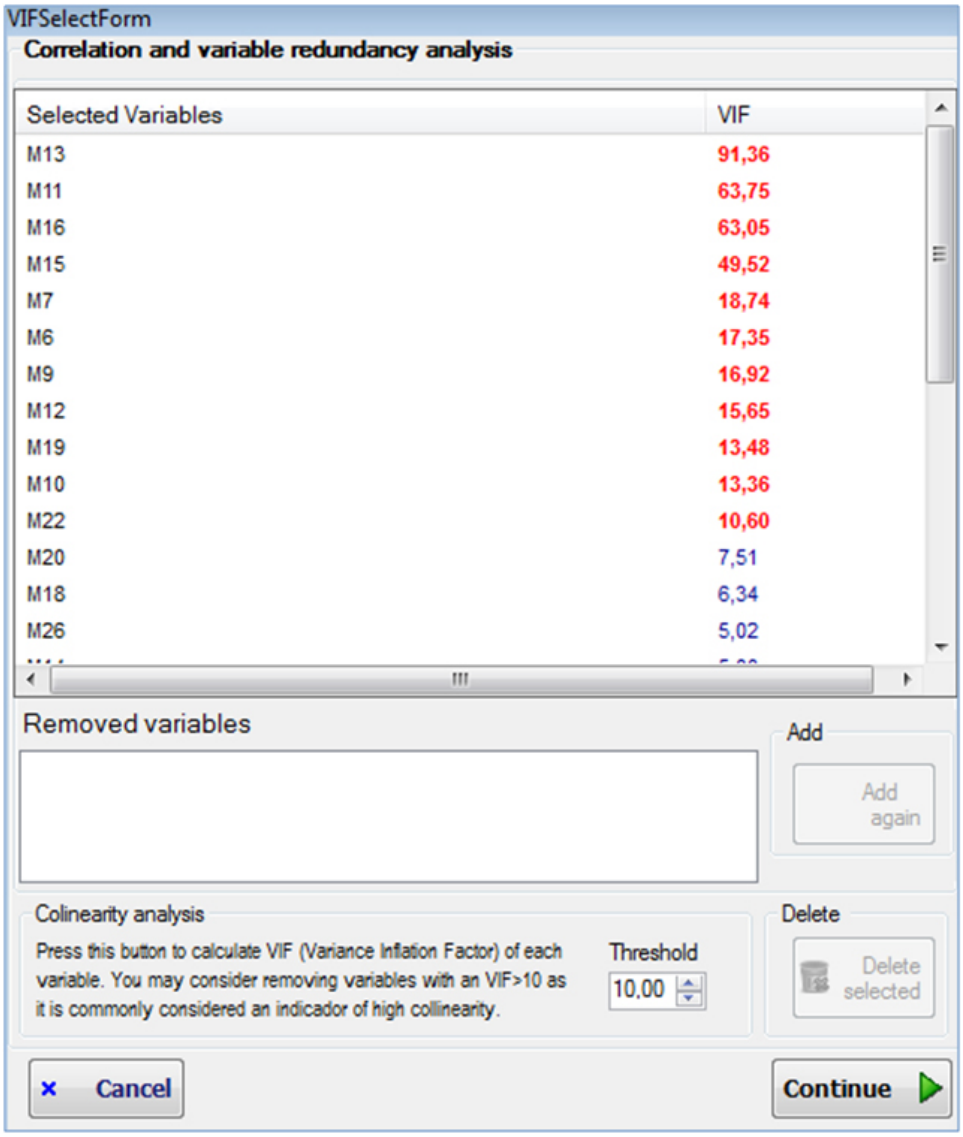
FUNCIONES

Se usaron las funciones vif, vifcor y vifstep del paquete usdm (Naimi, 2013; Naimi et al., 2014).

EJEMPLO

Los datos corresponden a medidas morfométricas de peces. El objetivo es seleccionar aquellas variables que no estén correlacionadas de un conjunto de variables. El método de selección elegido es "vif", lo cual significa que saldrá un menú, el cual se muestra a continuación (Figura XI3.1), donde el usuario va decidiendo que variables elimina. Además, con el argumento *varCode* se indica que se incluyan las variables "Order", "Family", "Genus" y "Species" en el archivo de exportación con la selección de las variables no correlacionadas.

Figura XI3.1. Menú que permite seleccionar que variables se eliminan, entre aquellas que están correlacionadas.



Value

Se exportan dos archivos CSV: uno con los valores VIF de las variables que se han seleccionado y no están correlacionadas y otro con las variables incluidas en el argumento *varCode* y las seleccionadas que no están correlacionadas.

References

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. 2014. Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:

data(ZX3)
XI3(data=ZX3, variables = c("M2", "M3", "M4", "M5", "M6", "M7", "M8",
"M9", "M10", "M11", "M12", "M13", "M14", "M15", "M16", "M17", "M18",
"M19", "M20", "M21", "M22", "M23", "M24", "M25", "M26", "M27", "M28"),
method="vif",varCode=c("Order","Family","Genus","Species"))


## End(Not run)
```

XI4 *SELECCION DE VARIABLES CORRELACIONADAS USANDO EL VIF*

Description

Se utiliza el factor de inflación de la varianza (VIF) para seleccionar las variables altamente correlacionadas, a partir de una matriz inicial con múltiples variables.

Usage

```
XI4(data, variables, threshold=10, varCode=NULL, file1="VIF.csv",
file2="Variables.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`variables`	Variables a partir de las cuales se quiere seleccionar aquellas correlacionadas.
`threshold`	Valor de corte para el VIF.
`varCode`	Opcionalmente se puede seleccionar variables de la matriz original, que se exportarán en el archivo de salida con las variables correlacionadas. Por ejemplo, permite seleccionar variables que son códigos de las filas.
`file1`	ARCHIVOS CSV. Nombre del archivo con los valores VIF para todas las variables.
`file2`	ARCHIVOS CSV. Nombre del archivo donde se exportan las variables correlacionadas.
`na`	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

XI.1.4. Selección de variables usando el factor de inflación de la varianza (VIF)

XI.1.4.2. Seleción de variables correlacionadas

No es frecuente, pero existen modelos estadísticos multivariados donde es útil que las variables estén correlacionadas, por ejemplo, el Análisis de Componentes Principales (PCA). Con esta función es fácil generar un archivo con las variables correlacionadas a partir de un conjunto de variables inicial. El método de selección se basa en el Factor de Inflación de la Varianza (VIF) que se explicó en la sección *details* de la función XI3.

FUNCIONES

Se usó la función vif del paquete usdm (Naimi, 2013; Naimi et al., 2014).

EJEMPLO

Los datos corresponden a medidas morfométricas de peces. El objetivo es seleccionar aquellas variables que estén altamente correlacionadas de un conjunto de variables. Además, se indica que se incluyan las variables "Familia" y "Género" en el archivo de exportación con la selección de las variables correlacionadas.

Value

Se exportan dos archivos CSV: uno con los valores VIF de las variables que se han seleccionado y están correlacionadas, y el otro archivo con las variables incluidas en el argumento *varCode* y las seleccionadas que están correlacionadas.

References

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:

data(ZX3)

XI4(data=ZX3, variables=c("M11","M12","M13","M15","M24","M2", "M3", "M4", "M5"),
varCode=c("Family","Genus"))


## End(Not run)
```

XI5 *REGRESION MULTIPLE*

Description

Se aplican diferentes técnicas estadísticas para obtener el tipo de función que mejor se ajusta a la posible relación que existe entre una variable dependiente cuantitativa y varias variables independientes cuantitativas.

Usage

```
XI5(data, varInteresY, varInteresX, outliers=FALSE, quant1=0.05, quant2 = 0.95,
stepwise=FALSE, direction="both", deplog=FALSE, indlog=FALSE, method="hier.part",
threshold=10, CoVa=FALSE, typeV="lmg", rela=TRUE, b=1000, names.abbrev=15,
ylimV=NULL, mainV=NULL, cex.title=1.5, PAIRS=TRUE, lower.panel=panel.smooth,
upper.panel=panel.reg, diag.panel=panel.hist, main=NULL, cex.main=2, cex=1.2,
pch=24, colhist="#00FFFFFF", bg="#B2DFEEFF", cex.labels=1, font.labels=2,
file1="Output.txt", file2="Residuals.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`varInteresY`	Variable dependiente.
`varInteresX`	Variables independientes.
`outliers`	Si es TRUE se eliminan los valores atípicos (*outliers*) usando el modelo de regresión seleccionado (para más detalles véase la sección *details* de la función XI1).
`quant1`	Cuantil del extremo inferior para la eliminación de los datos atípicos.
`quant2`	Cuantil del extremo superior para la eliminación de los datos atípicos.
`stepwise`	Si es TRUE se aplica una regresión por pasos para eliminar aquellas variables que no son significativas. Utiliza el criterio de información de Akaike (*Akaike Information Criterion, AIC)*) para definir cuales son las variables que se excluyen (véase la sección *details* para más detalles).
`direction`	Modo de búsqueda por pasos, que puede ser uno de los siguientes: "both", "backward", o "forward", siendo el valor por defecto "both".
`deplog`	Si es TRUE se aplica el logaritmo a la variable dependiente utilizando la ecuación $ln(variable - (min(variable) - 1))$.
`indlog`	Si es TRUE se calcula el logaritmo de todas las variables independientes y se añaden al resto de variables, para así realizar el modelo de regresión con las variables independientes transformadas y sin transformar, para obtener el mejor modelo. Se aplica la misma ecuación que en el argumento *deplog*.
`method`	Hay cuatro métodos para seleccionar las variables no correlacionadas: "vif", "vifcor" "vifstep" y "hier.part" (valor por defecto). El método "vif" no se pone por defecto, ya que solamente funciona con RWizard y, por tanto, no funciona

	en otras plataformas, como por ejemplo RGUI. Si este argumento es NULL no se eliminan las variables que presentan colinearidad. Para más detalles véase la sección *details*.
`threshold`	Valor de corte para el VIF (véase la sección *details* de la función XI3 para más información).
`CoVa`	Si es TRUE se muestra el gráfico con la contribución de las variables independientes al modelo de regresión, siempre y cuando el número de variables independientes sea mayor de dos.
`typeV`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Puede ser uno o un vector con los siguientes métodos: "lmg", "pmvd" (sólo versión fuera de Estados Unidos , "last", "first", "betasq", "pratt", "genizi" or "car". Para más detalles ver la sección *details* de la función calc.relimp.
`rela`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Si es TRUE la relativa importancia de todas las variables suman 100%.
`b`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Es el número de ciclos de arranque solicitados en boot.relimp. Asegúrese de ajustar a un número más bajo, si usted simplemente está probando código.
`names.abbrev`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Un número entero que define el número de letras de las etiquetas.
`ylimV`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Un vector con los límites del eje Y.
`mainV`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Título del gráfico.
`cex.title`	GRÁFICO CONTRIBUCIÓN DE LAS VARIABLES DEL ARGUMENTO *CoVa*. Tamaño del texto de título.
`PAIRS`	Si es TRUE se muestra el gráfico matricial. Si el número de variables es superior a 10, es mejor definir el argumento como FALSE, ya que pierde la utilidad al no observarse bien las relaciones entre las variables.
`lower.panel`	GRÁFICO MATRICIAL. Tipo de gráfico del panel inferior izquierdo.
`upper.panel`	GRÁFICO MATRICIAL. Tipo de gráfico del panel superior derecho.
`diag.panel`	GRÁFICO MATRICIAL. Tipo de gráfico de la diagonal.
`main`	GRÁFICO MATRICIAL. Título del gráfico.
`cex.main`	GRÁFICO MATRICIAL. Tamaño de la letra del título.
`cex`	GRÁFICO MATRICIAL. Tamaño de los símbolos.
`pch`	GRÁFICO MATRICIAL. Tipo de símbolo.
`colhist`	GRÁFICO MATRICIAL. Color de las barras del histograma.
`bg`	GRÁFICO MATRICIAL. Color de los símbolos.
`cex.labels`	GRÁFICO MATRICIAL. Tamaño de las etiquetas.
`font.labels`	GRÁFICO MATRICIAL. Tipo de fuente de las etiquetas.
`file1`	ARCHIVO TXT. Nombre del archivo de salida con los resultados de los análisis.
`file2`	ARCHIVOS CSV. Nombre del archivo con los residuos de los modelos de regresión.

na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI. REGRESIONES

XI.1. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS

XI.1.5. Regresión múltiple

Hasta ahora hemos visto regresiones en las que había una sola variable independiente. Sin embargo, es muy frecuente que necesitemos estudiar si nuestra variable dependiente está relacionada con más de una variable. En este caso, es necesario usar la regresión múltiple.

Además de la Normalidad y homocedasticidad de los residuos, y que no debe existir autocorrelación, requisitos de las regresiones simples (véase la función XI1), otro requisito de las regresiones múltiples es que no debe existir correlación entre las variables independientes, es decir, no debe existir multicolinealidad.

En la función se puede elegir entre cuatro métodos diferentes para seleccionar las variables no correlacionadas. Si se selecciona "vif" sale una ventana interactiva donde se va informando de las variables con VIF por encima del umbral indicado por el usuario en el argumento *threshold*, y el usuario es el que va decidiendo que variables elimina. La ventaja del método manual es que el usuario toma las decisiones de las variables a eliminar entre los pares correlacionados.

En las opciones "vifcor" y "vifstep" se hace una selección automática de las variables a excluir, lo cual es lógicamente mucho más rápido y cómodo. Si se selecciona "vifcor", primero busca el par de variables que tiene la correlación lineal máxima, y elimina la que tenga el mayor VIF, repitiendo el proceso hasta que no quede ninguna variable con un alto coeficiente de correlación con otra variable. Si selecciona "vifstep" se calcula el VIF para todas las variables, y se excluye la que tenga un VIF mayor, siempre considerando el umbral definido por el usuario en el argumento *threshold*, repitiendo el proceso hasta que ninguna variable esté correlacionada con otra. El problema de los métodos "vifcor" y "vifstep" es que no se tiene en cuenta, a la hora de eliminar las variables, cual es la que tiene una mayor contribución a la hora de explicar la variabilidad observada en la variable dependiente y, por tanto, no son los mejores métodos para obtener el mejor modelo de regresión.

La opción por defecto es "hier.part" en la cual se calcula la contribución de todas las variables independientes al modelo de regresión por el método de particionamiento hierárquico (Chevan & Sutherland, 1991) y, posteriormente, entre las variables que estén por encima del umbral del VIF, se eliminan aquellas que contribuyan menos al modelo de regresión. Por tanto, entre las variables auto-correlacionadas, se seleccionan las que más explican la variabilidad observada en la variable dependiente. El problema es que solamente se puede usar este método si el número de variables dependientes, incluyendo las transformadas logarítmicamente y no transformadas, no es mayor de 12. En el caso de que el número sea mayor, entonces se aplica automáticamente el método "vifstep".

La opción por defecto del argumento *outliers=TRUE*, significa que se eliminan los valores atípicos como se explicó en la función XI1.

Si se selecciona el argumento *stepwise=TRUE* significa que se realiza una regresión por pasos utilizando el criterio de información de Akaike (*Akaike Information Criterion, AIC)*) para definir

cuales son las variables que se excluyen del modelo de regresión, ya que no contribuyen a explicar la variable dependiente. Se utiliza el criterio de Akaike (AIC) para decidir cuantas variables explicativas se eligen y cuáles deben ser éstas. El método de *Akaike* valora la bondad del modelo estableciendo una penalización por su complejidad, de modo que es preferible un modelo más sencillo a otro con más variables independientes que explica solamente una pequeña parte adicional de la variabilidad de la variable dependiente. Con carácter general $AIC = 2k - 2ln(L)$, donde *k* es el número de variables explicativas o parámetros del modelo y *L* la verosimilitud o probabilidad asociada a la muestra utilizada de acuerdo con el modelo, de modo que *AIC* es más pequeño cuanto menor sea el número de variables y cuanto mayor sea la verosimilitud; entre varios modelos alternativos se debe elegir el que tiene menor valor de *AIC*. En cada paso se elige la variable que cuando se incorpora al modelo hace más pequeño el valor *AIC*, hasta que ninguna variable de las no utilizadas permite reducirlo.

Si el argumento *indlog=TRUE*, se calcula el logaritmo de todas las variables independientes usando la siguiente ecuación $LnVariable = ln(variable - (min(variable) - 1))$, y todas estas variables se añaden a las no transformadas. El objetivo es realizar un modelo de regresión en el que se incluyan las variables transformadas y no transformadas, para encontrar el mejor modelo, ya que a veces la variable transformada muestra una mejor relación con la variable dependiente que la no transformada logarítmicamente.

FUNCIONES

El particionamiento hierárquico se realiza con la función hier.part del paquete hier.part (Walsh & Mac Nally, 2014). Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors, la función dwtest del paquete lmtest (Hothorn et al., 2013) para analizar la autocorrelación con el test de Durbin-Watson, la función bptest del paquete lmtest (Hothorn et al., 2013) para realizar el test Breusch-Pagan de homocedasticidad, las funciones vif, vifcor y vifstep del paquete usdm (Naimi, 2013; Naimi et al., 2014) para el análisis del VIF, la función calc.relimp para el gráfico de la contribución de las variables independientes (Grömping, 2006; 2013) y las funciones panels y panels.diag para el gráfico matricial (Grosjean, 2013).

EJEMPLO

Los datos corresponden a un estudio realizado en áreas de afloramiento donde se obtuvo información sobre la concentración de clorofila (mg^{-3}), porcentaje de plataforma continental en la celda, transporte de Ekman ($m^3km^{-1}s^{-1}$) que es una medida de afloramiento, temperatura, turbulencia (m^3s^{-1}), estabilidad de la columna de agua ($cicloss^{-1}$) y las concentraciones de fosfatos, nitratos y silicatos en μMl^{-1}. El objetivo es determinar cuales son las variables que mejor explican las variaciones observadas en la concentración de clorofila.

Paso 1.

En primer lugar se trabaja con la variable dependiente sin transformar y, para ello, dejamos la opción que viene por defecto en el argumento *deplog=FALSE*. Sin embargo, el argumento *outliers=TRUE* para eliminar los datos atípicos,*stepwise=TRUE* para realizar la regresión por pasos y eliminar del modelo las variables no significativas, *indlog=TRUE* para aplicar el logaritmo a las variables independientes y añadirlas al resto de variables para estimar el modelo, y dejamos la opción por defecto de *method="hier.part"* para eliminar las variables independientes correlacionadas utilizando el método de particionamiento hierárquico.

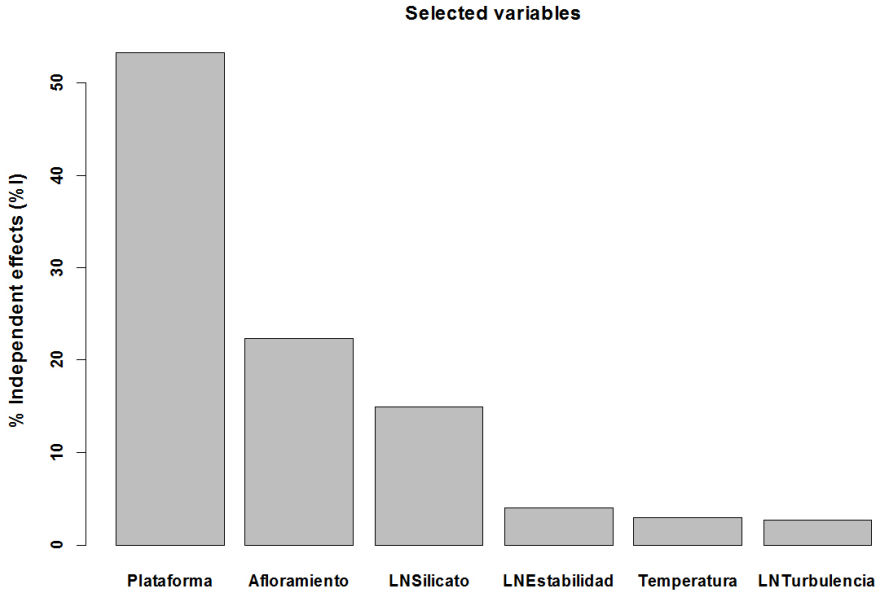
La primera tabla muestra valores altos de VIF, lo cual significa que existe multicolinearidad entre las variables independientes, lo cual es lógico ya que están la misma variable transformada y sin transformar, por ejemplo «LNTemperature» y «Temperature». La segunda tabla muestra la

contribución de cada una de las variables al modelo de regresión estimada por el método de particionamiento hierárquico. Como en el script se usa la opción por defecto *method="hier.part"*, se eliminan las variables auto-correlacionadas (variables con valores de VIF mayores que el umbral) siguiendo el criterio de eliminar aquellas con menor contribución al modelo entre las auto-correlacionadas. La tercera tabla muestra las variables que se han seleccionado, una vez eliminadas las auto-correlacionadas de menor contribución.

```
[1] "VIF FOR ALL VARIABLES"

[[2]]
         Variables       VIF
2       Plataforma  7.864145
3     Afloramiento  6.995529
4      Temperatura 29.628284
5      Turbulencia 11.497157
6      Estabilidad 94.916634
7         Silicato 26.167858
8     LNPlataforma 13.080808
9   LNAfloramiento  3.015055
10   LNTemperatura 31.908229
11   LNTurbulencia  4.744479
12   LNEstabilidad 71.230677
13      LNSilicato 30.271760


[1] "CONTRIBUTION OF VARIABLES ESTIMATED"

[[4]]
[1] "BY USING HIERARCHICHAL PARTITIONING (I)"

[[5]]
                     I
Plataforma     16.682711
Afloramiento   10.356266
Temperatura     5.096797
Turbulencia     3.752280
Estabilidad     5.423239
Silicato        9.064665
LNPlataforma   18.179903
LNAfloramiento  9.780946
LNTemperatura   4.354167
LNTurbulencia   3.904649
LNEstabilidad   4.922652
LNSilicato      8.481723

[[6]]
[1] "VIF FOR SELECTED VARIABLES"

[[7]]
      Variables      VIF
2     Plataforma 5.585788
3   Afloramiento 4.763180
9     LNSilicato 6.942991
8  LNEstabilidad 3.320566
4    Temperatura 5.535428
7  LNTurbulencia 3.856257
```

La Figura XI5.1 muestra la contribución de cada una de las variables seleccionadas, estimada por medio del método de particionamiento hierárquico. El porcentaje de plataforma continental presente en la celda fue la variable que más contribuye a explicar la variabilidad observada en la concentración de clorofila y en segundo lugar la intensidad del afloramiento.

Figura XI5.1. Contribución relativa de las variables independientes estimada por particionamiento hierárquico.



Las variables seleccionadas y mostradas en la Figura XI5.1, no todas fueron significativas y en el modelo de regresión por pasos solamente se seleccionaron el porcentaje de plataforma continental que existe en la celda, la intensidad de afloramiento y la temperatura. Estas tres variables explican el 60,7% de la varianza observada en la concentración de clorofila ($r^2 = 0,61$).

```
[1] "MULTIPLE REGRESSION"

[[16]]

Call:
lm(formula = Clorofila ~ Plataforma + Afloramiento + Temperatura,
    data = datos2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2852 -0.4641 -0.1930  0.4070  1.3108

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2218527  1.0460765   2.124  0.04264 *
Plataforma    0.0199161  0.0041346   4.817 4.58e-05 ***
Afloramiento  0.0008112  0.0002223   3.649  0.00107 **
Temperatura  -0.1297266  0.0619057  -2.096  0.04529 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7336 on 28 degrees of freedom
Multiple R-squared:  0.6073, Adjusted R-squared:  0.5653
F-statistic: 14.44 on 3 and 28 DF,  p-value: 7.142e-06
```

Las tres variables seleccionadas presentan unos valores de VIF menores de 10 y, por tanto, no existe problema con la colinearidad.

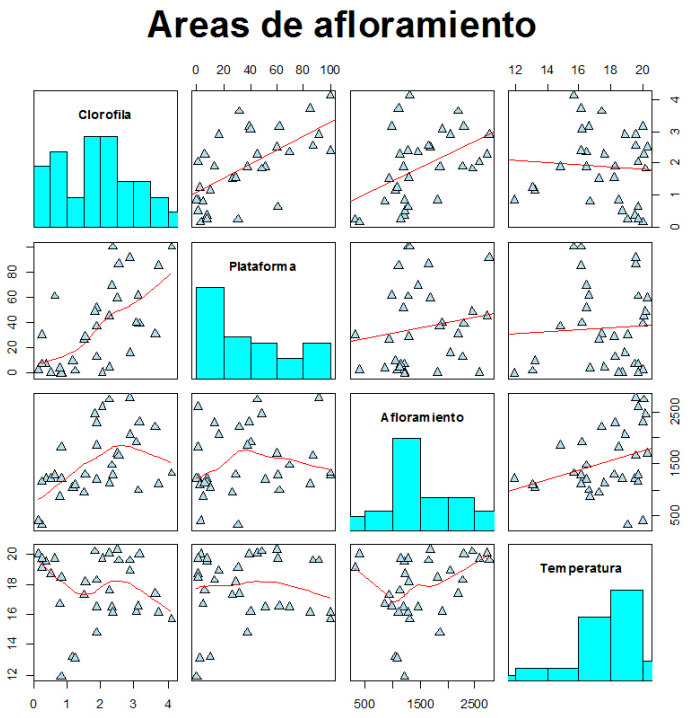
```
[1] "VIF of model"

[[6]]
  Plataforma Afloramiento  Temperatura
    1.026996     1.149107     1.122363
```

También se observa, que se cumplen los supuestos de Normalidad (p = 0,134), no existe autocorrelación (p = 0,115) y hay homocedasticidad en los residuos (p = 0,549). En caso de no cumplirse algunos de estos supuestos véase la sección *details* de la función XI1, para más detalles.

```
[1] "Normality"

[[8]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.1366, p-value = 0.1341


[[9]]
[1] "Autocorrelation"

[[10]]

     Durbin-Watson test

data:  reg
DW = 1.7319, p-value = 0.1146
alternative hypothesis: true autocorrelation is greater than 0


[[11]]
[1] "Homocedasticity"

[[12]]

     studentized Breusch-Pagan test

data:  reg
BP = 2.115, df = 3, p-value = 0.5489
```

La Figura XI5.2 muestra la relación que existe entre todas las variables seleccionadas, incluyendo la clorofila y las variables independientes.

Figura XI5.2. Gráfico matricial que muestra la relación entre las variables independientes y la concentración de clorofila.



Paso 2.

A continuación repetimos el proceso pero transformando la variable dependiente y, para ello, modificamos el argumento *deplog=TRUE*. El modelo de regresión que se obtiene muestra que se seleccionan las mismas variables que antes: el porcentaje de plataforma y el afloramiento están relacionadas positivamente y la temperatura negativamente con la concentración de clorofila.

```
[1] "MULTIPLE REGRESSION"

[[16]]

Call:
lm(formula = LNClorofila ~ Plataforma + Afloramiento + Temperatura,
    data = datos2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47223 -0.21366 -0.02371  0.18069  0.61229

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.208e+00  4.009e-01   3.013  0.00544 **
Plataforma    7.600e-03  1.585e-03   4.796 4.84e-05 ***
Afloramiento  4.021e-04  8.521e-05   4.720 5.97e-05 ***
Temperatura  -6.692e-02  2.373e-02  -2.820  0.00871 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2812 on 28 degrees of freedom
Multiple R-squared:  0.6607, Adjusted R-squared:  0.6244
F-statistic: 18.18 on 3 and 28 DF,  p-value: 9.595e-07
```

Las variables seleccionadas explican un 66% de la variabilidad observada en la concentración de

clorofila ($r^2 = 0,66$). Por tanto, este modelo es mejor que el anterior, en el cual la variable clorofila no estaba transformada. Es posible que se incluyan en el modelo variables explicativas con una p > 0,05. Con el criterio de Akaike utilizado esto significa que la mayor complejidad del modelo al añadir esta variable está compensada con el aumento de verosimilitud (la muestra observada es más probable o verosímil con el nuevo modelo). Debe indicarse no obstante que otros métodos de selección paso a paso descartan todas las variables no significativas, por lo que la decisión del investigador de excluir una variable con p > 0,05 podría considerarse también razonable.

También se observa, que se cumplen los supuestos de Normalidad (p = 0,894), existe una cierta autocorrelación (p = 0,007, véase la sección *details* de la función XI1 para más detalles) y hay homocedasticidad en los residuos (p = 0,995).

```
[1] "Normality"

[[8]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.0776, p-value = 0.8943


[[9]]
[1] "Autocorrelation"

[[10]]

     Durbin-Watson test

data:  reg
DW = 1.3158, p-value = 0.006952
alternative hypothesis: true autocorrelation is greater than 0


[[11]]
[1] "Homocedasticity"

[[12]]

     studentized Breusch-Pagan test

data:  reg
BP = 0.0686, df = 3, p-value = 0.9953
```

El modelo de regresión viene definido por la siguiente fórmula:

$$ln(Clorofila - (min(Clorofila) - 1)) = 1,2 + 0,007 * Plataforma +$$

$$0,0004 * Afloramiento - 0,069 * Temperatura$$

Value

Se obtiene un archivo TXT con los valores VIF de todas las variables, los VIF de las variables seleccionadas después de eliminar las que tengan colinearidad, los VIF de las variables que finalmente quedan en el modelo de regresión, los test de Normalidad, homogeneidad de varianzas y homocedasticidad y el modelo de regresión. También se obtiene un archivo CSV con los residuos tipificados, así como un gráfico con la relación entre la variable dependiente y todas las independientes, además de un gráfico con la contribución de cada una de las variables independientes estimada por particionamiento hierárquico y, por último si se selecciona, un gráfico con la contribución de las variables independientes al modelo de regresión.

References

Chevan, A. & Sutherland, M. (1991) Hierarchical Partitioning. *The American Statistician* 45, 90-96.

Durbin, J. & Watson G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, 38: 159-171.

Grömping, U. (2006) Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, 17: 1-27.

Grömping, U. (2013)Relative importance of regressors in linear models. R package version 2.2-2. Available at: http://CRAN.R-project.org/package=relaimpo.

Grosjean, P. (2013) SciViews GUI API - Main package. R package version 0.9-5. Available at: http://CRAN.R-project.org/package=SciViews.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T. et al., (2013) Testing Linear Regression Models R package version 0.9-33. Available at: http://CRAN.R-project.org/package=lmtest.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Walsh, C. & Mac Nally, R. (2014) Hierarchical Partitioning. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=hier.part.

Examples

```
## Not run:
data(ZXI6)

#Paso 1

XI5(data=ZXI6, varInteresY="Clorofila",varInteresX = c("Plataforma",
"Afloramiento", "Temperatura", "Turbulencia", "Estabilidad", "Silicato"),
outliers=TRUE, stepwise=TRUE, indlog=TRUE, main="Areas de afloramiento")

#Paso 2

XI5(data=ZXI6, varInteresY="Clorofila",varInteresX=c("Plataforma",
"Afloramiento", "Temperatura", "Turbulencia","Estabilidad", "Silicato"),
outliers=TRUE, stepwise=TRUE, deplog=TRUE, indlog=TRUE, main="Areas de afloramiento")

## End(Not run)
```

Description

Se aplican diferentes técnicas estadísticas para obtener el tipo de función que mejor se ajusta a la posible relación que existe entre una variable dependiente cualitativa con dos categorías y varias variables independientes cualitativas o cuantitativas.

Usage

```
XI6(data, cat, var, gruporesp, stepwise=TRUE, cart=TRUE,
minsplit=2, minbucket=2, cp=0.0001, surrogatestyle=0, type=0, extra=2,
varlen=0, ycompress= TRUE, main=NULL, cex.main=2, cex=1.2, model="Model.rda",
file1="Coefficients.csv", file2="Predictions.csv",  na="NA", dec=",",
row.names=FALSE, file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`cat`	Variable dependiente.
`var`	Variables independientes.
`gruporesp`	Categoría que será referencia en la variable dependiente.
`stepwise`	Si es TRUE se aplica una regresión logística por pasos, para eliminar aquellas variables que no son significativas. Utiliza el criterio de información de Akaike (*Akaike Information Criterion, AIC*) para definir cuales son las variables que se excluyen (véase la sección *details* de la función XI5 para más detalles).
`cart`	Si es TRUE se crea un gráfico de árbol de clasificación y regresión (*CART*).
`minsplit`	CART. Se especifica el número mínimo de observaciones que deben existir en un nodo para que se haga la partición, es decir, conforme sea más pequeño el árbol tendrá más nodos.
`minbucket`	CART. Se define el número mínimo de observaciones en un nodo terminal y, por tanto, al igual que el argumento anterior, conforme es más pequeño se profundiza más en el árbol, es decir, tiene más nodos y se hace más complejo.
`cp`	CART. Controla la complejidad y permite ahorrar tiempo en el proceso de poda. Es la fracción (por defecto 0,0001) en la que debe mejorar el indicador de ajuste para que continúe la construcción del árbol. Si la mejora es menor que *cp* el proceso se detiene. El usuario decide si quiere profundizar mucho y hacer el árbol muy complejo (*cp* pequeño), o tener un árbol más simple (*cp* más grande).
`surrogatestyle`	CART. Controla el criterio para elegir la mejor variable suplente para la partición en cada nodo. Si es 0 se utiliza la variable con el mayor número de elementos clasificados correctamente, y si es 1 la que tiene el mayor porcentaje de elementos clasificados correctamente sobre el número de casos válidos de esa variable; la primera opción penaliza a las variables que tienen valores perdidos.
`type`	CART. Formato del gráfico: 0 muestra las variables que han sido seleccionadas en cada partición y al final las categorías predominantes en los nodos terminales, 1 es igual que la anterior pero además muestra la categoría predominante en cada partición, 2 igual que 1 pero pone encima el nodo y debajo la variable responsable de la partición, 3 la variable responsable de la partición se pone en

ambas ramas y, por último, 4 es igual que 3 pero se etiquetan los nodos, no las ramas.

`extra` CART. Permite mostrar información adicional en los nodos: 0 solamente muestra el grupo mayoritario en el nodo, 1 muestra el número de casos de todas las categorías, 2 muestra el número de casos correctos frente al total, 3 muestra el número de casos incorrectos frente al total, 4 muestra la probabilidad de todas las categorías, 5 es igual que 4 pero no muestra la categoría que predomina en el nodo, 6 muestra la probabilidad de la segunda clase que predomina en el nodo, 7 es igual que 6 pero no muestra la categoría que predomina en el nodo, 8 muestra la probabilidad de la clase que predomina en el nodo y, por último, 9 muestra la probabilidad relativa considerando todas las observaciones.

`varlen` CART. Define el número de caracteres de los textos: 0 utiliza el texto completo, >0 los abrevia teniendo en cuenta el número que se indique, y <0 recorta el texto de las variables teniendo en cuenta el número que se indique al máximo posible para ajustar el tamaño,

`ycompress` CART. Si es TRUE, cuando se solapan los nodos se desplazan verticalmente para evitar este solapamiento.

`main` CART. Título del gráfico.

`cex.main` CART. Tamaño de letra del título.

`cex` CART. Tamaño de letra de las etiquetas de los nodos.

`model` Nombre del archivo con el modelo.

`file1` ARCHIVOS CSV. Nombre del archivo con los coeficientes de la regresión.

`file2` ARCHIVOS CSV. Nombre del archivo con las predicciones del modelo logístico.

`na` ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.

`dec` ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".

`row.names` ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

`file` ARCHIVO TXT. Nombre del archivo de salida con los resultados de la regresión logística.

Details

XI. REGRESIONES

XI.2. MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUALITATIVAS

XI.2.1. Regresión logística

En el caso de que la variable dependiente sea cualitativa un modelo adecuado es el de regresión logística. Se utiliza ampliamente en investigación clínica ya que permite estimar la probabilidad de ocurrencia de un proceso en función de ciertas variables, permitiendo evaluar la influencia de las variables independientes sobre la variable dependiente, dando como resultado una probabilidad. La variable dependiente es siempre cualitativa pero las variables independientes pueden ser continuas, discretas, categóricas, dicotómicas o una mezcla de todas ellas. Al igual que ocurría con la regresión para variables dependientes cuantitativas es necesario que no exista multicolinealidad

entre las diferentes variables independientes y las observaciones de la muestra deben ser independientes entre sí. Sin embargo, no se requiere que los residuos presenten una distribución Normal ni la hipótesis de homocedasticidad (varianza constante de los residuos). La variable dependiente puede ser dicotómica (0 si el hecho no ocurre y 1 si ocurre) o politómica (hay varias categorías), dando lugar a dos tipos diferentes de regresión logística, la binomial o multinomial, respectivamente.

XI.2.1.1. Regresión logística binomial

XI.2.1.1.1. Estimación del modelo

La variable dependiente es:

$$Y_i = log\left(\frac{\Pi_i}{1-\Pi_i}\right)$$

donde Π_i es la probabilidad de que en el caso *i* se produzca el evento estudiado e $Y_i = \alpha + \beta X_i$ es el valor de la variable dependiente en el caso *i*.

Expresado en forma de regresión:

$$\Pi_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$$

donde α y β son los coeficientes de la ecuación.

En caso de existir varias variables predictoras (independientes) la regresión se transforma en:

$$\Pi_i = \frac{1}{1+e^{\left[-(\alpha+\sum_{j=1}^{k}\beta_j X_{ji})\right]}}$$

donde Π_i es la probabilidad de que se produzca el evento estudiado en el caso *i*, *k* es el número de variables predictoras, α es un coeficiente, β_j es el coeficiente de la variable predictora *j* y X_{ji} es el valor de la variable predictora *j* en el caso *i*.

En caso de que Π_i sea mayor de 0,5 se asume (a efectos de predicción) que el evento se produce y si es menor que 0,5 que no se produce.

Como complemento a la regresión logística, en esta función se usan los árboles de clasificación y regresión (CARTs). Son un procedimiento no paramétrico de predicción de una variable dependiente sobre la base de un conjunto de variables independientes (Breiman et al., 1984). Su carácter no paramétrico hace que no requiera ninguna hipótesis relativa a la distribución de las variables dependientes e independientes, ni a la relación entre ellas y sus posibles interacciones. La variable dependiente puede ser categórica (Árboles de Clasificación), como por ejemplo diferentes especies, rechazo a un tratamiento, presencia/ausencia de un cáncer, etc. La variable dependiente también puede ser continua (Árboles de Regresión), como por ejemplo el índice de masa corporal, el azúcar en sangre, etc. En ambos casos, Árboles de Clasificación y de Regresión (CARTs), el objetivo es identificar las variables que mejor permiten identificar la variable dependiente. En esta función se utiliza los CARTs para determinar las variables que mejor discriminan las dos categorías de la variable dependiente. Son por tanto, una herramienta excelente para elegir el mejor modelo al ayudar a definir cuales son las variables independientes con mayor poder de discriminación. Para más detalles véase el manual de ayuda de la función rpart y/o Guisande et al. (2012).

FUNCIONES

Para estimar la regresión logística binomial se usa la función glm del paquete stats. Para estimar el modelo por pasos se usó la función stepAIC del paquete MASS (Venables & Ripley, 2002; Ripley

et al., 2014). Para la estimación del CART la función rpart (Therneau et al., 2014) y para su representación la función rpart.plot (Milborrow, 2014).

EJEMPLO

El estudio consistió en examinar pacientes a los que se les ha diagnosticado un tipo de cáncer, así como una muestra de personas que no lo tienen. El objetivo es determinar si variables como el sexo, edad, estado civil (soltero, casado y separado) y presencia/ausencia de sangre en la orina, pueden ser indicadoras de la presencia del cáncer.

Se dejaron las opciones por defecto y, por tanto, se realiza la regresión por pasos. Se eligió un *minbucket=10* para hacer un CART más simplificado, ya que solamente interesa tener información sobre cuales son las variables que mejor discriminan entre las diferentes familias.

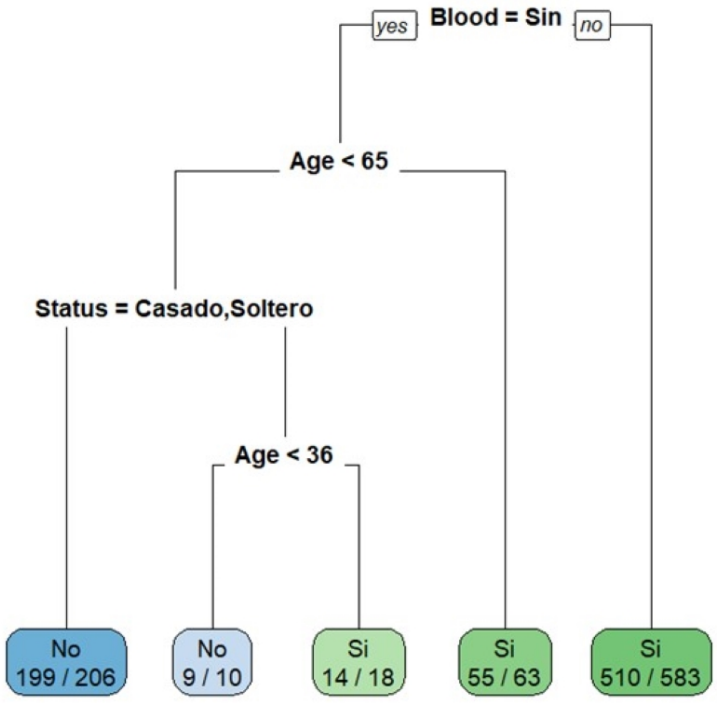
Los resultados muestran los coeficientes de la regresión, que también se obtienen en un archivo CSV. El modelo predice correctamente el 74,1% de las personas que no tienen cáncer. Sin embargo, lo importante es que es capaz de predecir correctamente el 95,1% de las personas que tienen cáncer. Por lo tanto, es un modelo relativamente bueno para poder predecir cuando una persona potencialmente puede tener ese tipo cáncer.

```
Call:
glm(formula = Cancer ~ Age + Gender + Status + Blood, family

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2989  -0.3378   0.2016   0.5672   1.6030

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.53024    0.31146  -1.702  0.08867 .
Age              0.06690    0.00661  10.121  < 2e-16 ***
GenderMujer     -0.33310    0.21244  -1.568  0.11689
StatusSeparado   1.15068    0.37779   3.046  0.00232 **
StatusSoltero   -0.92790    0.23059  -4.024 5.72e-05 ***
BloodSin        -3.64301    0.25749 -14.148  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1119.80  on 879  degrees of freedom
Residual deviance:  579.36  on 874  degrees of freedom
AIC: 591.36

Number of Fisher Scoring iterations: 6

[1] "Porcentaje correctamente clasificado total"
[2] "88.0681818181818"
[1] "Porcentaje correctamente clasificado Cancer Si"
[2] "95.0596252129472"
[1] "Porcentaje correctamente clasificado Cancer No"
[2] "74.061433447099"
```

El CART muestra que la presencia de la sangre en la orina es la variable más importante (Figura XI8.1). Como el coeficiente para la categoría es negativo (-3,64) y ese coeficiente hace referencia a la categoría de sin sangre en la orina, esto indica que las personas sin cáncer no suelen tener sangre en la orina. La segunda variable en importancia es la edad. El coeficiente positivo (0,067) indica que a mayor edad es más frecuente la aparición del cáncer. Por último, en el CART se incorpora

el estado civil de la persona (Figura XI8.1). En el caso de que las variables categóricas no sean dicotómicas (sólo el estado civil), se crea un coeficiente *B* para cada categoría distinta de la de referencia (que tiene valor 0). El valor negativo del coeficiente *B* nos indica que entre las personas solteras (-0,927) la presencia de cáncer es menos frecuente. El hecho de que sea hombre o mujer no influye de forma importante en la presencia del cáncer (p =0,11), y aunque se incluye en el modelo, posiblemente su exclusión no significaría cambios importantes en la capacidad predictiva del modelo.

Figura XI8.1. Arbol de clasificación y regresión que muestra la importancia de las variables analizadas en el estudio para predecir la presencia de cáncer.



El modelo vendría definido por la siguiente ecuación:

$$Pi_{Cancer} = \frac{1}{1 + e^{-(-0{,}53+0{,}067*Edad-0{,}333*Sexo.Mujer+1{,}151*Separado-0{,}928*Soltero-3{,}643*SinSangre)}}$$

Por ejemplo, supongamos el primer caso de la tabla de datos, la probabilidad de que un hombre casado, de 59 años, y con presencia de sangre en la orina tenga este tipo de cáncer sería de 0,96 y, por tanto, como es mayor que 0,5, se pronostica que la persona podría tener cáncer:

$$Pi_{Cancer} = \frac{1}{1 + e^{-(-0{,}53+0{,}067*59-0{,}333*0+1{,}151*0-0{,}928*0-3{,}643*0)}} = 0,96$$

La probabilidad de que un hombre soltero, de 23 años, y sin sangre en la orina, tenga este tipo de cáncer sería 0,027 y, por tanto, como es menor que 0,5, se pronostica que la persona no tendría

cáncer:

$$Pi_{Cancer} = \frac{1}{1 + e^{-(-0{,}53+0{,}067*23-0{,}333*0+1{,}151*0-0{,}928*1-3{,}643*1)}} = 0{,}027$$

Value

Se obtiene un archivo TXT con los coeficientes de la regresión logística y el porcentaje de casos correctamente identificados. Se obtienen dos archivos CSV: 1) Las predicciones del modelo logístico y 2) Los coeficientes de la regresión. Por último, se muestra el árbol de clasificación y regresión, si se ha seleccionado.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: http://CRAN.R-project.org/package=rpart.plot.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: http://CRAN.R-project.org/package=rpart.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI7)

XI6(data=ZXI7, cat="Cancer", var=c("Age","Gender", "Status","Blood"),
gruporesp="Si", minbucket=10)


## End(Not run)
```

XI7 *REGRESION LOGISTICA BINOMIAL-PREDICCION*

Description

Permite calcular la categoría a la que pertenece un caso, pudiendo realizar el cálculo para muchos casos a la vez, usando un modelo de regresión logística binomial estimado con la función XI6.

Usage

```
XI7(data, cat, gruporesp, model="Model.rda", file="Model predictions.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`cat`	Variable dependiente.
`gruporesp`	Categoría que será referencia en la variable dependiente.
`model`	Nombre del archivo con el modelo estimado con la función XI6
`file`	ARCHIVO CSV. Nombre del archivo con las predicciones del modelo logístico binomial, para cada caso.
`na`	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI.2.1. Regresión logística

XI.2.1.1. Regresión logística binomial

XI.2.1.1.2. Predicción del modelo

El porcentaje de casos correctamente identificados es un parámetro importante para evaluar la calidad del modelo de regresión logística. Sin embargo, un parámetro mejor para evaluar la calidad del modelo es comprobar el grado de acierto usando unos datos que no se habían incluido dentro de la matriz original que se usó para estimar el modelo.

Esta función permite de forma fácil pronosticar el grupo de pertenencia de un caso, lo cual además de servir para evaluar la calidad del modelo, también sirve para poder usar el modelo, es decir, pronosticar a que categoría pertenece un determinado caso, lo cual puede ser de gran interés por ejemplo en medicina.

Para poder usar esta función es necesario haber, en primer lugar, calculado el modelo logístico binomial con la función XI6.

EJEMPLO

Los datos deben tener el mismo formato, es decir, las mismas variables y en la misma posición, que la matriz de datos utilizada para estimar el modelo con la función XI6. Además, el directorio de trabajo debe ser el mismo que se usó cuando se estimó el modelo con la función XI6.

El estudio consistió en examinar pacientes a los que se les ha diagnosticado un tipo de cáncer, así como una muestra de personas que no lo tienen. Sin embargo, estos datos no fueron incluidos cuando se estimó el modelo original de regresión logística binomial (véase el ejemplo de la función XI6). El objetivo es predecir si la persona puede tener ese tipo de cáncer.

Como se mencionó anteriormente, para poder usar esta función es necesario haber, en primer lugar, calculado el modelo logístico binomial con la función XI6. También es muy importante que el grupo de referencia, en este caso *gruporesp="Si"*, sea igual al que se especificó cuando se calculó el modelo logístico binomial con la función XI6.

Los resultados muestran que el porcentaje de acierto es del 88,3%, en estos 762 casos de los que consta la base de datos.

Value

Se obtiene un archivo CSV con las predicciones del modelo logístico para cada caso.

Examples

```
## Not run:

data(ZXI8)

XI7(data=ZXI8, cat="Cancer", gruporesp="Si")


## End(Not run)
```

XI8 *REGRESION LOGISTICA MULTINOMIAL-ESTIMACION*

Description

Se aplican diferentes técnicas estadísticas para obtener el tipo de función que mejor se ajusta a la posible relación que existe entre una variable dependiente cualitativa con múltiples categorías y varias variables independientes cualitativas o cuantitativas.

Usage

```
XI8(data, cat, var, optimize=TRUE, methop="cart", stepwise=FALSE, order=TRUE,
cart=TRUE, minsplit=2, minbucket=2, cp=0.0001, surrogatestyle=0, type=0,
extra=2, varlen=0, ycompress= TRUE, main=NULL, cex.main=2, cex=1.2,
file1="Predictions.csv", file2="Coefficients.csv", file3="Table.csv",
na="NA", dec=",", row.names=FALSE, model= "Model.rda", file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`cat`	Variable dependiente.
`var`	Variables independientes.
`optimize`	Si es TRUE se pueden utilizar dos algoritmos diferentes, que se definen en el argumento *methop*, que tienen como finalidad buscar un modelo que identifique correctamente el 100% de los casos. Si es TRUE la primera variable del argumento *var* debe ser cuantitativa. Si es TRUE el modelo que se graba, el cual puede ser posteriormente utilizado para predicción en la función XI9, es el optimizado y no el logístico ni por pasos, aunque ambos se calculen.
`methop`	Define el algoritmo que se utiliza para optimizar el modelo. Si es "cart" se calcula con los Arboles de Regresión y Clasificación (CARTs), que se explicó en la función XIV1, las variables que más contribuyen a diferenciar las categorías, y luego se van introduciendo esas variables en el modelo de regresión por orden de importancia. Si es "stepwise" se aplica una regresión logística por pasos, para eliminar aquellas variables que no son significativas. Utiliza el criterio de información de Akaike (*Akaike Information Criterion, AIC*) para definir cuales son las variables que se excluyen (véase la sección *details* de la función XI5 para más detalles). Posterior a la regresión por pasos, se añaden más variables al modelo si contribuyen a aumentar el porcentaje de casos identificados correctamente. Por tanto, se saca un modelo con todas las variables, el modelo de regresión logística por pasos y el modelo optimizado, que es el último modelo que queda grabado.
`stepwise`	Si es TRUE se aplica una regresión logística por pasos, además del modelo sin pasos, para eliminar aquellas variables que no son significativas utilizando el criterio AIC mencionado anteriormente en el argumento *methop*. La diferencia con el método *optimize=TRUE* es que posteriormente no se añaden más variables al modelo utilizando el criterio de aumentar el porcentaje de casos identificados correctamente. Este argumento se pone de forma automática en modo TRUE si el argumento *optimize=TRUE* y el argumento *methop="stepwise"*.
`order`	Si es TRUE los coeficientes de regresión se ordenan en el archivo *Coefficients.CSV* de la A a la Z para caracteres y de menor a mayor para números.
`cart`	Si es TRUE se crea un gráfico de árbol de clasificación y regresión (*CART*).
`minsplit`	CART. Se especifica el número mínimo de observaciones que deben existir en un nodo para que se haga la partición, es decir, conforme sea más pequeño el árbol tendrá más nodos.
`minbucket`	CART. Se define el número mínimo de observaciones en un nodo terminal y, por tanto, al igual que el argumento anterior, conforme es más pequeño se profundiza más en el árbol, es decir, tiene más nodos y se hace más complejo.
`cp`	CART. Controla la complejidad y permite ahorrar tiempo en el proceso de poda. Es la fracción (por defecto 0,0001) en la que debe mejorar el indicador de ajuste para que continúe la construcción del árbol. Si la mejora es menor que *cp* el proceso se detiene. El usuario decide si quiere profundizar mucho y hacer el árbol muy complejo (*cp* pequeño), o tener un árbol más simple (*cp* más grande).

`surrogatestyle` CART. Controla el criterio para elegir la mejor variable suplente para la partición en cada nodo. Si es 0 se utiliza la variable con el mayor número de elementos clasificados correctamente, y si es 1 la que tiene el mayor porcentaje de elementos clasificados correctamente sobre el número de casos válidos de esa variable; la primera opción penaliza a las variables que tienen valores perdidos.

`type` CART. Formato del gráfico: 0 muestra las variables que han sido seleccionadas en cada partición y al final las categorías predominantes en los nodos terminales, 1 es igual que la anterior pero además muestra la categoría predominante en cada partición, 2 igual que 1 pero pone encima el nodo y debajo la variable responsable de la partición, 3 la variable responsable de la partición se pone en ambas ramas y 4 es igual que 3 pero se etiquetan los nodos, no las ramas.

`extra` CART. Permite mostrar información adicional en los nodos: 0 solamente muestra el grupo mayoritario en el nodo, 1 muestra el número de casos de todas las categorías, 2 muestra el número de casos correctos frente al total, 3 muestra el número de casos incorrectos frente al total, 4 muestra la probabilidad de todas las categorías, 5 es igual que 4 pero no muestra la categoría que predomina en el nodo, 6 muestra la probabilidad de la segunda clase que predomina en el nodo, 7 es igual que 6 pero no muestra la categoría que predomina en el nodo, 8 muestra la probabilidad de la clase que predomina en el nodo y, por último, 9 muestra la probabilidad relativa considerando todas las observaciones.

`varlen` CART. Define el número de caracteres de los textos: 0 utiliza el texto completo, >0 los abrevia teniendo en cuenta el número que se indique, y <0 recorta el texto de las variables teniendo en cuenta el número que se indique al máximo posible para ajustar el tamaño,

`ycompress` CART. Si es TRUE, cuando se solapan los nodos se desplazan verticalmente para evitar este solapamiento.

`main` CART. Título del gráfico.

`cex.main` CART. Tamaño de letra del título.

`cex` CART. Tamaño de letra de las etiquetas de los nodos.

`file1` ARCHIVOS CSV. Nombre del archivo con las predicciones para cada caso del modelo logístico, del modelo por pasos y del modelo optimizado, en el caso de que se seleccionara realizar los dos últimos.

`file2` ARCHIVOS CSV. Nombre del archivo con los coeficientes de la regresión. Si *optimize=TRUE* son los coeficientes del modelo optimizado, si *optimize=FALSE* y *stepwise=TRUE* son los coeficientes del modelo por pasos y, por último, si *optimize=FALSE* y *stepwise=FALSE* son los coeficientes del modelo en el que se incluyen todas las variables.

`file3` ARCHIVOS CSV. Nombre del archivo con el porcentaje de casos identificados para cada categoría de la variable dependiente.

`na` ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.

`dec` ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".

`row.names` ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

`model` Nombre del archivo con el modelo.

`file` ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XI.2.1. Regresión logística

XI.2.1.2. Regresión logística multinomial

XI.2.1.2.1. Estimación del modelo

La diferencia principal con respecto a la regresión logística binomial es que la variable dependiente cualitativa no es dicotómica, sino que puede tener más de 2 categorías (para más detalles véase Guisande et al., 2011).

Suponiendo *k* clases o categorías, y *j* variables independientes, este modelo se puede resumir en las siguientes funciones, que proporcionan las probabilidades de pertenencia a las primeras *k* - 1 clases:

$$Z_{in} = \beta_{n0} + \beta_{n1}X_{i1} + \beta_{n2}X_{i2} + ... + \beta_{nj}X_{ij}$$

$$\Pi_{in} = \frac{e^{Z_{in}}}{1 + e^{Z_{i1}} + e^{Z_{i2}} + e^{Z_{i3}} + ... + e^{Z_{ik-1}}}$$

donde Π_{in} es la probabilidad de pertenencia del caso *i* a la clase *n*; Z_{in} es el valor de la variable dependiente *Z* correspondiente a la clase *n* en el caso *i*; β_{nj} es el coeficiente de la variable independiente *j* para la clase *n*; x_{ij} es el valor del predictor o variable independiente *j* para el caso *i*. La probabilidad para la última clase *k* se obtiene por diferencia a 1.

Igual que en el caso de la regresión logística binomial, el árbol de clasificación y regresión (CART) es una excelente herramienta para identificar cuales son las variables que presentan mayores diferencias entre las categorías de la variable dependiente (Breiman et al., 1984) y, por tanto, permite elegir más fácilmente las variables con las que se va a conseguir un mayor número de casos identificados correctamente con el modelo de regresión. Por lo tanto, es de gran ayuda para obtener el modelo de regresión logística más predictivo. Para más detalles véase el manual de ayuda de la función rpart y/o Guisande et al. (2012).

FUNCIONES

Para la estimación de la regresión logística multinomial se utiliza la función multinom (Ripley, 2014). Para estimar el modelo por pasos se usó la función stepAIC del paquete MASS (Venables & Ripley 2002; Ripley et al., 2014). Para la estimación del CART se usó la función rpart (Therneau et al., 2014) y para su representación la función rpart.plot (Milborrow, 2014).

EJEMPLO

Los datos son medidas corporales de varias especies de peces (variables continuas), así como el tipo de boca que tienen (variable categórica). El objetivo es determinar si es posible identificar a que familia pertenece un individuo en función de la biometría.

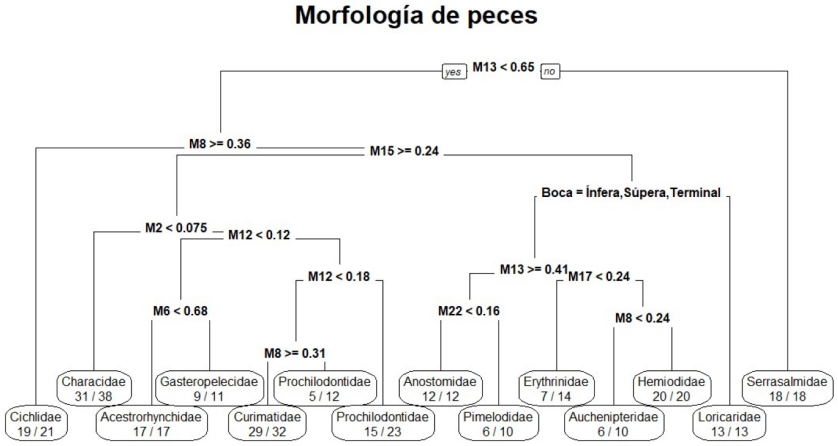
Paso 1.

En primer lugar se selecciona la variable dependiente *cat="Familia"* y todas las variables morfométricas como independientes. No se dejan las condiciones por defecto, ya que el argumento *optimize=FALSE* y, por tanto, no se realiza el modelo optimizado. Por el contrario el argumento *stepwise=TRUE* y, por tanto, se realiza la regresión por pasos. Se eligió un *minbucket=10* para hacer un CART más simplificado, ya que solamente interesa tener información sobre cuales son las variables que mejor discriminan entre las diferentes familias.

El CART que se obtiene (Figura XI8.2) muestra las variables más importantes que mejor discriminan a las familias y es incluso mejor método que la regresión por pasos, para seleccionar las variables. Las variables que están más arriba en el árbol y que son las responsables de los primeros nodos, son las que tienen una mayor capacidad de discriminación.

En el modelo logístico sin pasos, que incluye todas las variables, se consigue un 100% de casos identificados correctamente, mientras que en el modelo por pasos se consigue un 97,6% de acierto. En el archivo CSV que se genera para guardar los coeficientes estimados, cuyo nombre por defecto es *Coefficients.CSV*, se observa que las variables seleccionadas en el modelo por pasos son M12, M16, M17 y M22. No salen coeficientes para la primera categoría, que se calcula por defecto a partir de las otras categorías. En el archivo *Predictions.CSV* se muestra la categoría real y las predichas para caso, por los modelos sin y por pasos.

Figura XI8.2. Arbol de clasificación y regresión de medidas de peces de agua dulce.



Paso 2.

A continuación lo que se hace es realizar un modelo en el que se deja el modo por defecto, es decir, el argumento *optimize=TRUE*. Por otro lado, no se realiza el CART *cart=FALSE*. Los resultados muestran que el modelo identifica correctamente el 100% de los casos.

Una vez elaborado el modelo es posible utilizarlo para predecir a que categoría pertenece un individuo. La función XI9 permite calcular la categoría a la que pertene un individuo, pudiendo realizar el cálculo para muchos individuos a la vez, usando el modelo de regresión sin pasos, por pasos o el optimizado, que se estiman con esta función.

Value

Se obtiene un archivo TXT con los coeficientes de la regresión logística que no es por pasos, el porcentaje de casos identificados correctamente, y los coeficientes de la regresión logística por pasos y el porcentaje de casos correctamente identificados, en el caso de que se seleccionara la regresión por pasos. Se obtienen cinco archivos CSV: 1) Las predicciones del modelo logístico para cada caso y del modelo por pasos, en el caso de que se seleccionara realizarlo; 2) Los coeficientes de la regresión; 3) Los coeficientes de la regresión por pasos, si se seleccionó; 4) Una tabla con el porcentaje de casos identificados para cada grupo y 5) Una tabla con el porcentaje de casos identificados para cada grupo usando el modelo logístico por pasos, si fue seleccionado. Se genera también un archivo con el modelo logístico sin pasos y otro para el modelo por pasos, si fue seleccionado, que se usan posteriormente en la función XI9 para pronosticar a que categorías pertenecen los individuos. Por último, se muestra el árbol de clasificación y regresión, si se ha seleccionado.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: http://CRAN.R-project.org/package=rpart.plot.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Ripley, B. (2014) Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models. R package version 7.3-8. Available at: http://CRAN.R-project.org/package=nnet.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: http://CRAN.R-project.org/package=rpart.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI9)

#Paso 1
windows(26,14)
XI8(data = ZXI9, cat="Familia", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Boca"), optimize=FALSE, stepwise=TRUE,
minbucket=10, main="Morfología de peces", cex=1)

#Paso 2
XI8(data=ZXI9, cat="Familia", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Boca"), cart=FALSE)


## End(Not run)
```

XI9 *REGRESION LOGISTICA MULTINOMIAL-PREDICCION*

Description

Permite calcular la categoría a la que pertenece un caso, pudiendo realizar el cálculo para muchos casos a la vez, usando los modelos de regresión sin pasos y por pasos, que se estiman con la función XI8.

Usage

```
XI9(data, cat, model="Model.rda", file="Model predictions.csv", na="NA",
dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`cat`	Variable dependiente.
`model`	Nombre del archivo con el modelo estimado con la función XI8
`file`	ARCHIVO CSV. Nombre del archivo con las predicciones del modelo logístico o del modelo por pasos, para cada caso.
`na`	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XI.2.1. Regresión logística

XI.2.1.2. Regresión logística multinomial

XI.2.1.2.2. Predicción del modelo

El porcentaje de casos correctamente identificados es un parámetro importante para evaluar la calidad del modelo de regresión logística. Sin embargo, un parámetro mejor para evaluar la calidad del modelo es comprobar el grado de acierto usando unos datos que no se habían incluido dentro de la matriz original utilizada para estimar el modelo.

Esta función permite de forma fácil pronosticar el grupo de pertenencia de un caso, lo cual además de servir para evaluar la calidad del modelo, también sirve para poder usar el modelo, es decir, pronosticar a que categoría pertenece un determinado caso, lo cual puede ser de gran interés por ejemplo en medicina.

Para poder usar esta función es necesario haber, en primer lugar, calculado el modelo sin pasos o por pasos con la función XI8.

EJEMPLO

Los datos deben tener el mismo formato, es decir, las mismas variables y en la misma posición, que la matriz de datos que se usó cuando se estimo el modelo con la función XI8. Además, el directorio de trabajo debe ser el mismo que se usó cuando se estimó el modelo con la función XI8.

Los datos son medidas corporales de varias especies de peces (variables continuas), así como el tipo de boca que tienen (variable categórica), de la misma población de los datos que se usaron para el ejemplo de la función XI8. Sin embargo, estos datos no fueron incluidos cuando se estimó el modelo original. Ya que se conoce a que familia pertenece cada individuo, el objetivo es evaluar la calidad del modelo por medio del grado de acierto con estos individuos.

Como se mencionó anteriormente, para poder usar esta función es necesario haber, en primer lugar, calculado el modelo sin pasos o por pasos con la función XI8.

Los resultados muestran que el porcentaje de acierto es del 100%, en estos 29 individuos de los que consta la base de datos. Por tanto, parece ser que el modelo obtenido es bastante fiable.

Value

Se obtiene un archivo CSV con las predicciones del modelo logístico para cada caso.

Examples

```
## Not run:

data(ZXI10)

XI9(data=ZXI10, cat="Familia")


## End(Not run)
```

XII1 *META-ANALISIS/DIAGRAMA DE BOSQUE*

Description

Se aplica un meta-análisis y se muestra en un diagrama de bosque.

Usage

```
XII1(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
event.e	Variable con el número de eventos en el grupo experimental.
n.e	Variable con el número total de observaciones en el grupo experimental.
event.c	Variable con el número de eventos en el grupo control.
n.c	Variable con el número total de observaciones en el grupo control.
studlab	Variable con el código de los estudios.
META	Se accede a la función metabin que permite modificar los argumentos del meta-análisis.
FOREST	Se accede a la función forest.meta que permite modificar los argumentos del gráfico del meta-análisis.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XII. META-ANALISIS

XII.1. DIAGRAMA DE BOSQUE

El meta-análisis es una técnica estadística que permite combinar varios estudios estadísticos en uno solo. Con ello se consigue construir estadísticos de resumen que puedan sacar ventaja de la mayor potencia proporcionada por el tamaño de muestra, necesariamente más grande, resultante de la agregación de las diferentes muestras de todos los estudios. Este análisis es de especial interés en medicina, donde es importante determinar si un determinado medicamento, un compuesto, una terapia, etc., puede ser beneficioso o perjudicial para la salud y, sobre todo, en aquellos casos en los cuales existe discrepancia entre los diferentes estudios. Como se mencionó anteriormente, la ventaja del meta-análisis es que permite combinar la información de todos los trabajos realizados. Las representaciones que se utilizan más frecuentemente son el diagrama de bosque y el diagrama de embudo.

FUNCIONES

Para realizar el meta-análisis se usó la función metabin y para el diagrama del bosque forest.meta, ambas del paquete meta (Schwarzer, 2013). Para más detalles de como usar los argumentos de estas funciones, a las cuales se da acceso desde en los argumentos *META* y *FOREST*, consultar la ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO

Los datos corresponden a un estudio de Silagy & Ketteridge (1997) sobre el efecto que tiene el consejo médico a la hora de dejar de fumar. En el archivo de datos, la variable *event.e* hace referencia al número de personas que dejaron el tabaco por recomendación del médico, *n.e* el número total de personas del grupo a las que el médico recomendó dejar de fumar y, *event.c* y *n.c* son las personas que dejaron de fumar y el total de personas que estaban en el grupo control, a las que el médico no les recomendó dejar el tabaco, respectivamente.

En la función metabin, el argumento *method* especifica el método de agrupación de los casos: "MH" Mantel-Haenszel, "Inverse" inverso de la varianza, o "Peto" método Peto. Mantel-Haenszel (opción por defecto) calcula la OR agregada ponderando los datos de cada estudio por el número total de elementos que contiene considerando su distribución en clases, y es el más utilizado. Peto es un método alternativo para el cálculo de la OR que no requiere corrección para celdas con cero casos, por lo que se puede utilizar cuando se investigan eventos raros y el tamaño de muestra es pequeño.

Los primeros resultados que se observan en el archivo de salida TXT se muestran a continuación:

```
                    OR            95%-CI %W(fixed) %W(random)
Porter, 1972    1.1198 [0.2913;  4.3049]      1.83       2.29
Rusell, 1979    4.6848 [2.1585; 10.1681]      3.40       5.84
Wilson, 1982    2.1112 [0.9618;  4.6345]      4.04       5.72
Stewart, 1982   1.0208 [0.3210;  3.2461]      2.60       3.01
Rusell, 1983    1.0677 [0.6748;  1.6895]     16.12      11.62
Jamrozik, 1984 1.4985 [1.0409;  2.1573]     21.66      14.48
McDowell, 1985 1.0012 [0.4141;  2.4206]      4.49       4.76
Page, 1986      0.9509 [0.2981;  3.0336]      2.65       3.00
Janz, 1987      1.8908 [0.9122;  3.9193]      5.07       6.41
Slama, 1990     1.0194 [0.0629; 16.5161]      0.45       0.58
Vetter, 1990    1.7921 [0.9987;  3.2160]      7.85       8.72
Demens, 1990    3.1083 [1.1147;  8.6674]      2.16       3.71
Wilson, 1990    2.4394 [1.3735;  4.3326]      7.46       8.92
Haug, 1994      2.1748 [0.8856;  5.3409]      3.25       4.63
Higashi, 1995   1.6566 [1.0593;  2.5907]     13.83      11.95
Slama, 1995     3.5983 [1.4191;  9.1239]      3.14       4.37

Number of studies combined: k=16
```

El valor de *OR* (*Odds Ratio*), que mide el efecto del consejo del médico para el abandono del tabaco se calcula de la siguiente forma, por ejemplo, para el caso del segundo estudio de Rusell en 1979:

$$OR = \frac{34/(1031-34)}{8/(1107-8)} = 4,68$$

Ese valor indica que entre aquellos pacientes que recibieron el consejo médico, la relación entre los que abandonaron el tabaco y los que no lo hicieron (predominio o ventaja) es 4,68 veces más alta que entre los que no recibieron consejo médico. El consejo médico permitió, por lo tanto, multiplicar por algo más que cuatro el éxito alcanzado en el abandono del tabaco. Si la OR es significativamente menor que la unidad indica que el evento estudiado tiene un predominio menor en el grupo experimental que en el grupo de control: por ejemplo, si el evento estudiado es el fallecimiento del paciente, una OR menor que uno corresponde a un factor de protección, y una OR mayor que la unidad a un factor de riesgo. Se observa que no todos los estudios tienen una OR tan elevado como el del segundo estudio calculado anteriormente; algunos muestran valores próximos a la unidad (por ejemplo los de McDowell 1985 y Slama 1990), indicando un efecto nulo del consejo médico. El intervalo de confianza al 95% que sigue en las dos columnas siguientes nos ayuda a interpretar mejor la OR: si el intervalo contiene el valor 1 el efecto no es significativo (es decir indica que el consejo médico no parece ser eficaz), lo que ocurre en 10 de los 16 estudios.

Las resultados siguientes que aparecen en el archivo TXT, que se muestra a continuación, son el peso o ponderación de cada estudio, en función de los tamaños de muestra utilizados en cada uno. El modelo de efectos fijos (*fixed*) supone que los estudios analizados son todos los que se quiere considerar, y el de efectos aleatorios (*random*) que los datos utilizados son una muestra aleatoria de estudios obtenida a partir de un conjunto más amplio, lo que añade una fuente de variabilidad adicional. Se muestran los resultados agregados: OR = 1,76; aunque algunos estudios no parecen mostrar un efecto real del consejo médico, el conjunto sí que indica un efecto significativo (el intervalo no contiene el valor 1 y el valor p del contraste de nulidad (hipótesis nula: OR = 1) es muy pequeño, menor que 0,0001. En el test de Cochran-Mantel-Haenszel (véase Guisande et al. 2011, para más detalles sobre este test), el valor de p < 0,001 y, por tanto el número de pacientes que dejan de fumar, es diferente en el grupo experimental (aconsejados por el doctor) en comparación con el de control (no aconsejados por el doctor). El consejo médico multiplica por 1,76 el predominio de abandono del tabaco (el cociente entre los que abandonan y los que no).

```
                          OR           95%-CI      z  p.value
Fixed effect model    1.7614 [1.4922; 2.0792] 6.6905 < 0.0001
Random effects model  1.7621 [1.4225; 2.1828] 5.1865 < 0.0001

Cochran-Mantel-Haenszel (CMH) test for overall effect:
     Q d.f.  p.value
 45.73    1 < 0.0001
```

Los últimos resultados muestran algunas medidas de heterogeneidad. Cuando los estudios son muy heterogéneos, es decir muy diferentes entre sí, es discutible que puedan agregarse en uno solo. Por esa razón se utilizan diferentes medidas de heterogeneidad. Entre ellas están Tau cuadrado, variabilidad entre estudios, o el estadístico I cuadrado, que mide la proporción de la variabilidad total que es atribuible a la heterogeneidad entre estudios, es decir, la variabilidad entre estudios dividida por la variabilidad total = variabilidad entre estudios + variabilidad entre pacientes. Los autores del estadístico I cuadrado (Higgins et al., 2003) proponen los valores de corte 25 y 75 para interpretar que la heterogeneidad es baja (menos de 25%), moderada (25% - 75%) o alta (más de 75%). En nuestro ejemplo, 27,6% es moderada-baja y, por consiguiente, no supone ningún problema. Por último, se muestra el valor P del contraste de nulidad. Un valor, como p = 0,1451 observado en el ejemplo, mayor que el nivel de significación habitual 0,05 nos permite aceptar que no existe heterogeneidad importante.

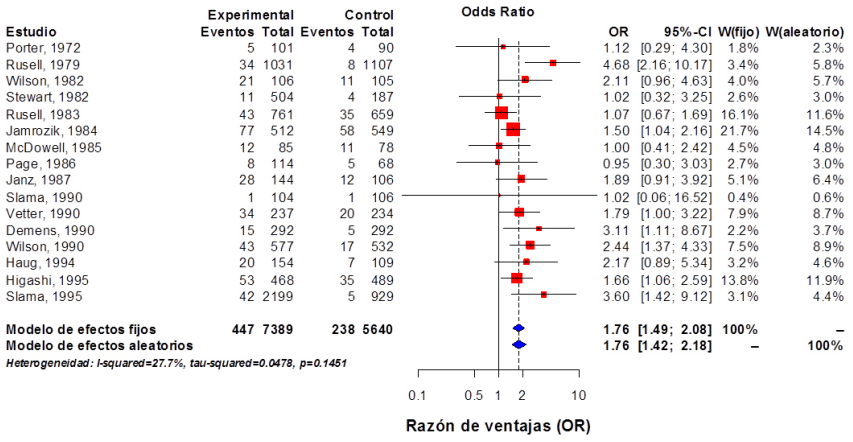
```
Quantifying heterogeneity:
tau^2 = 0.0478; H = 1.18 [1; 1.59]; I^2 = 27.7% [0%; 60.4%]

Test of heterogeneity:
     Q d.f.  p.value
 20.75   15   0.1451

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

En el diagrama de bosque (Figura XII.1), se muestran todos los resultados mencionados anteriormente. El valor estimado se representa mediante un cuadrado, y los límites de confianza son los extremos de cada línea horizontal. Vemos como muchas líneas sobrepasan el valor 1 por la derecha, indicando que la proporción de personas que dejan de fumar es mayor en el grupo experimental que fue aconsejado por el médico, que en el grupo control que no estaba influenciado por las recomendaciones del médico. El tamaño del cuadrado indica la muestra utilizada: cuanto más grande es el cuadrado más representativo es el estudio y, por lo tanto, más fiables sus conclusiones.

W es el peso o ponderación que corresponde a cada estudio en la obtención de la *OR* promedio o de resumen. Su valor depende del método de agregación utilizado, de los tamaños de muestra y del número de eventos en cada estudio. Si los estudios utilizados en el meta-análisis son todos los estudios existentes debemos utilizar el modelo de efectos fijos y, si se trata de una muestra de estudios elegida al azar, el aleatorio.

La *OR* de resumen calculada con el meta-análisis, se representa mediante un rombo en la parte inferior, siendo los extremos del rombo los límites de confianza correspondientes. Observamos como el intervalo de confianza es mucho más pequeño que cualquiera de los estudios individuales, lo que se debe al mayor tamaño de la muestra agregada, y el hecho de que sea mayor de 1, en concreto 1,76 tanto para los efectos fijos como para los aleatorios, significa que el médico influye positivamente a la hora de dejar el tabaco y, además, como mostró el test de Cochran-Mantel-Haenszel, esta diferencia es significativa. El gráfico permite, por lo tanto, apreciar los resultados y conclusiones de cada uno de los estudios individuales y el resultado agregado conseguido con el meta-análisis.

Figura XII.1. Diagrama de bosque de un estudio realizado sobre el efecto de la recomendación del médico en dejar de fumar.



Value

Se obtiene un archivo TXT con los resultados del meta-análisis, además del diagrama de bosque.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) Gráficos estadísticos y mapas con R. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influencia del consejo médico para dejar de fumar.

XII1(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
studlab="Research", FOREST=c("text.fixed ='Modelo de efectos fijos'",
"text.random='Modelo de efectos aleatorios'", "xlab='Razón de ventajas (OR)'",
"leftlabs=c('Estudio', 'Eventos', 'Total', 'Eventos', 'Total')",
"rightlabs=c('OR', '95
```

```
"hetlab='Heterogeneidad: '", "col.square='red'","col.diamond='blue'"))


## End(Not run)
```

XII2 *META-ANALISIS/DIAGRAMA DE BOSQUE ACUMULATIVO*

Description

Se aplica un meta-análisis y se muestra en un diagrama de bosque acumulativo.

Usage

```
XII2(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`event.e`	Variable con el número de eventos en el grupo experimental.
`n.e`	Variable con el número total de observaciones en el grupo experimental.
`event.c`	Variable con el número de eventos en el grupo control.
`n.c`	Variable con el número total de observaciones en el grupo control.
`studlab`	Variable con el código de los estudios.
`META`	Se accede a la función metabin que permite modificar los argumentos del meta-análisis.
`FOREST`	Se accede a la función forest.meta que permite modificar los argumentos del gráfico del meta-análisis.
`file`	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XII. META-ANÁLISIS

XII.2. DIAGRAMA DE BOSQUE ACUMULATIVO

Otra forma de representar el diagrama de bosque(ver sección de *details* de la función XII1) es por medio del diagrama de bosque acumulativo. La diferencia con el diagrama de bosque es que se van añadiendo los estudios conforme fueron apareciendo, lo cual permite ver si existe solidez en la información y, determinar si se llegó a un equilibrio en el resultado.

FUNCIONES

Para el meta-análisis se usó la función metabin y para el diagrama del bosque acumulativo metacum y forest.meta, todas del paquete meta (Schwarzer, 2013). Para más detalles de como usar estas funciones, a las cuales se da acceso desde en los argumentos META y FOREST, consultar la ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO

El ejemplo es el mismo que se mostró en la sección de *details* de la función XII1.

```
                                        I^2
Adding Porter, 1972 (k=1)
Adding Rusell, 1979 (k=2)           69.5%
Adding Wilson, 1982 (k=3)           49.9%
Adding Stewart, 1982 (k=4)          52.6%
Adding Rusell, 1983 (k=5)           66.1%
Adding Jamrozik, 1984 (k=6)         57.6%
Adding McDowell, 1985 (k=7)         52.4%
Adding Page, 1986 (k=8)             46.8%
Adding Janz, 1987 (k=9)             41.2%
Adding Slama, 1990 (k=10)           34.1%
Adding Vetter, 1990 (k=11)          28.5%
Adding Demens, 1990 (k=12)          30.2%
Adding Wilson, 1990 (k=13)          32.5%
Adding Haug, 1994 (k=14)            28.3%
Adding Higashi, 1995 (k=15)         22.8%
Adding Slama, 1995 (k=16)           27.7%

Pooled estimate                     27.7%

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

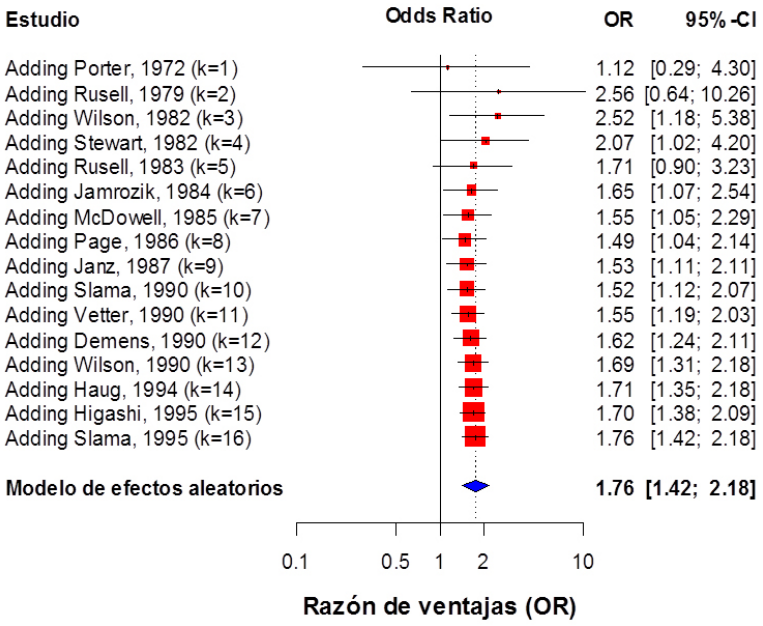
Los primero resultados que muestra el archivo TXT ya se explicaron en la sección *details* de la función XII1. La siguiente tabla muestra los valores de I^2, el cual ya se explicó en la sección *details* de la función XII1. Los autores del estadístico I cuadrado (Higgins et al., 2003) proponen una heterogeneidad moderada si los valores están entre el 25% -75%. En nuestro ejemplo se estabiliza al final con un valor del 27,7%, el cual es moderado-bajo y, por consiguiente, no supone ningún problema.

En el diagrama de bosque acumulativo que se obtiene (Figura XII.2), se muestran las OR e intervalos de confianza de los distintos estudios. El valor estimado se representa mediante un cuadrado, y los límites de confianza son los extremos de cada línea horizontal. Vemos como muchas líneas sobrepasan el valor 1 por la derecha, indicando que la proporción de personas que dejan de fumar es mayor en el grupo experimental que fue aconsejado por el médico, que en el grupo control que no estaba influenciado por las recomendaciones del médico. El tamaño del cuadrado indica la muestra utilizada: cuanto más grande es el cuadrado más representativo es el estudio y, por lo tanto, más fiables sus conclusiones.

En esta representación se van añadiendo los estudios conforme fueron apareciendo, lo cual permite ver si existe solidez en la información y, si se llegó a un equilibrio en el resultado, cuando se alcanzó este equilibrio. Aunque el orden cronológico es el método más usado a la hora de ir añadiendo estudios al meta-análisis, es posible usar cualquier otro orden.

El gráfico muestra como la estimación agregada se estabiliza, después de un período inicial con notable fluctuación, de modo que a partir del año 1990, los sucesivos estudios realizados no aportan cambios importantes, y el resultado del meta-análisis es prácticamente el mismo.

Figura XII.2. Diagrama de bosque acumulativo de un estudio realizado sobre el efecto de la recomendación del médico en dejar de fumar.



Value

Se obtiene un archivo TXT con los resultados del meta-análisis, además del diagrama de bosque acumulativo.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influencia del consejo médico para dejar de fumar.

XII2(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
```

```
studlab="Research", FOREST=c("text.random='Modelo de efectos aleatorios'",
"leftlabs='Estudio'", "xlab='Razón de ventajas (OR)'", "col.square='red'",
"col.diamond='blue'"))


## End(Not run)
```

XII3 *META-ANALISIS/GRAFICO DE L'ABBE*

Description

Se aplica un meta-análisis y se muestra en un gráfico de L'Abbé.

Usage

```
XII3(data, event.e, n.e, event.c, n.c, studlab, META=NULL, LABBE=NULL,
file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`event.e`	Variable con el número de eventos en el grupo experimental.
`n.e`	Variable con el número total de observaciones en el grupo experimental.
`event.c`	Variable con el número de eventos en el grupo control.
`n.c`	Variable con el número total de observaciones en el grupo control.
`studlab`	Variable con el código de los estudios.
`META`	Se accede a la función metabin que permite modificar los argumentos del meta-análisis.
`LABBE`	Se accede a la función labbe.metabin que permite modificar los argumentos del gráfico de L'Abbé.
`file`	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XII. META-ANÁLISIS

XII.3. GRÁFICO DE L'ABBÉ

En el caso de que exista heterogeneidad en los estudios de un meta-análisis (ver sección de *details* de la función XII1), el gráfico de L'Abbé permite identificar cuales son los estudios responsables de esta heterogeneidad.

El gráfico de L'Abbé ayuda a profundizar en el análisis de la presencia y causas de heterogeneidad en el meta-análisis. Se representa para cada estudio el resultado en el grupo de tratamiento (*eje y*) frente al resultado en el grupo de control (*eje x*), con una línea diagonal a 45º que divide al gráfico en dos partes: a un lado quedan los estudios en los que fue favorable el grupo experimental, y en el otro los estudios en los que fue favorable el grupo de control.

FUNCIONES

Para el meta-análisis se usó la función metabin y para el gráfico de L'Abbé labbe.metabin, ambas del paquete meta (Schwarzer, 2013). Para más detalles de como usar estas funciones, a las cuales se da acceso desde en los argumentos META y LABBE, consultar la ayuda de la función y/o Guisande & Vaamonde (2012).

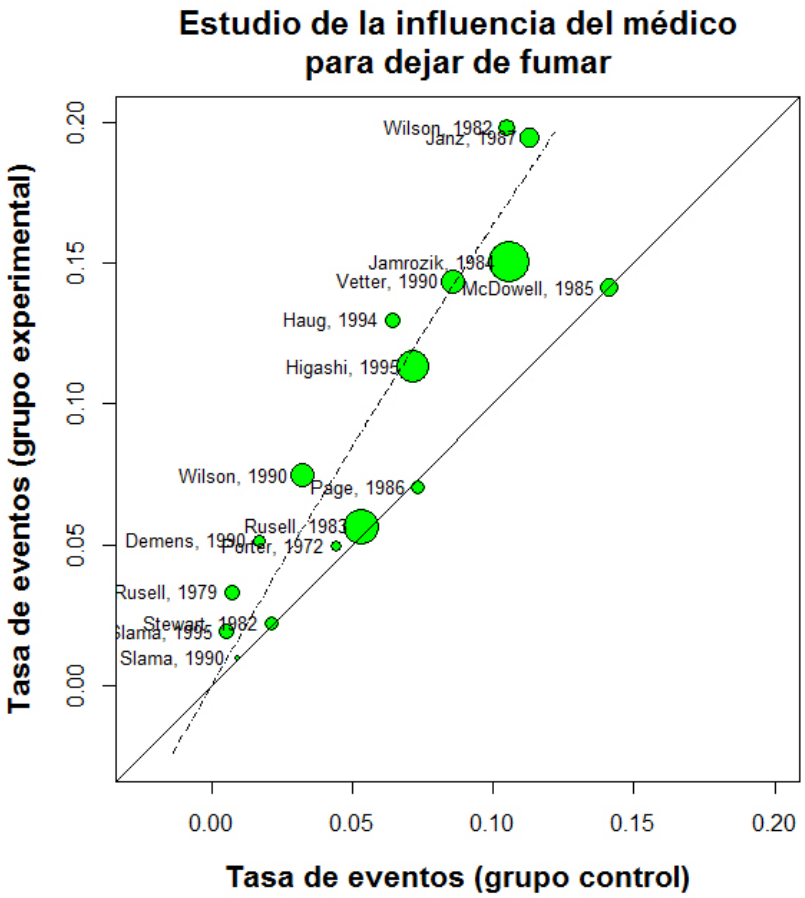
EJEMPLO

El ejemplo es el mismo que se explicó detalladamente en la sección de *details* de la función XII1.

Todos los resultados del archivo TXT también se explicaron en la sección de *details* de la función XII1.

En el gráfico de L'Abbé del ejemplo (Figura XII.3), por encima de la diagonal están los estudios en los que la proporción de pacientes que dejaron de fumar gracias al consejo médico es mayor que en el grupo de control. Los puntos se representan en tamaño variable en función del tamaño de muestra y de la precisión de cada estimación, de modo que los puntos más grandes son los estudios más representativos. Si los puntos se agrupan en una zona estrecha, próxima a una línea recta, significa que los resultados son homogéneos, y si están dispersos muestran heterogeneidad. Se representa también la recta de ajuste a la nube de puntos (línea punteada), cuya pendiente está relacionada con la *OR* conjunta. Es fácil identificar los estudios más alejados del patrón mayoritario, es decir, de la recta de ajuste, que serían los responsables de la heterogeneidad observada, cuando existe.

Figura XII.3. Gráfico de L'Abbé de un estudio realizado sobre el efecto de la recomendación del médico en dejar de fumar.



Value

Se obtiene un archivo TXT con los resultados del meta-análisis, además del gráfico de L'Abbé.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 1.1-12. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influencia del consejo médico para dejar de fumar.

XII3(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
studlab="Research", LABBE=c("xlim=c(-0.025,0.2)", "xlab='Tasa de eventos (grupo control)'",
"ylab='Tasa de eventos (grupo experimental)'", "main='Estudio de la influencia del médico\
para dejar de fumar'", "cex.main=1.4","bg='green'", "cex.lab=1.3", "font.lab=2", "studlab=TRUE"))


## End(Not run)
```

XII4 *META-ANALISIS/DIAGRAMA DE EMBUDO*

Description

Se aplica un meta-análisis y se muestra en un diagrama de embudo.

Usage

```
XII4(data, event.e, n.e, event.c, n.c, studlab, META=NULL, FUNNEL=NULL,
LEGEND=NULL, COLOR=c("#FFF68FFF", "#FFA54FFF", "#FF6347FF"),
file="Output.txt")
```

Arguments

data	Archivo de datos.
event.e	Variable con el número de eventos en el grupo experimental.
n.e	Variable con el número total de observaciones en el grupo experimental.
event.c	Variable con el número de eventos en el grupo control.
n.c	Variable con el número total de observaciones en el grupo control.
studlab	Variable con el código de los estudios.
META	Se accede a la función metabin que permite modificar los argumentos del meta-análisis.
FUNNEL	Se accede a la función funnel.meta que permite modificar los argumentos del diagrama de embudo.
LEGEND	Permite modificar la leyenda del gráfico.
COLOR	Permite modificar los colores del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XII. META-ANÁLISIS

XII.4. DIAGRAMA DE EMBUDO

Otro problema que puede tener el meta-análisis (ver sección de *details* de la función XII1) es el sesgo en la publicación, es decir, la publicación selectiva de los estudios en base a sus resultados. En otras palabras, que se publiquen más los resultados que tienen una determinada conclusión y menos los estudios en los cuales los resultados dan lugar a conclusiones diferentes.

En algunos estudios se ha observado que si los resultados son negativos, es decir, que no se observa un efecto, no se suelen publicar, ya que se basa en la idea equivocada de que un no efecto no es noticia. También puede ocurrir que no se publiquen los resultados que son desfavorables a la terapia usada con un determinado fármaco. El diagrama de embudo permite visualizar y evaluar el sesgo en la publicación sobre un determinado tema.

El análisis para realizar el gráfico de embudo parte de la idea de que los estudios negativos, que no muestran efecto, tienen menor probabilidad de ser publicados. Si existe sesgo de publicación, que puede ocurrir preferentemente en los estudios pequeños en los que hay mayor probabilidad de que se alteren los resultados por azar, se tendería a publicar los que mostraran diferencias.

En el diagrama de embudo se representa el estimador de diferencias calculado en el meta-análisis (razón de ventajas, riesgo relativo, etc.) para cada estudio, junto con su error estándar. En el caso de no existir sesgo, los puntos se deberían agrupar alrededor de un estimador central, y mostrarían tanta mayor dispersión alrededor de este valor cuanto más pequeño fuera su tamaño, de modo que la nube de puntos se distribuirá en forma de embudo invertido.

Si hubiera sesgo de publicación en el sentido descrito anteriormente, la nube de puntos se deformaría y el embudo perdería su simetría. En la gráfica anterior se observa que la distribución tiene forma de embudo, corroborando la ausencia de sesgo que también mostró el test de asimetría.

FUNCIONES

Para el meta-análisis se usó la función metabin y para el diagrama de embudo funnel.meta, ambas del paquete meta (Schwarzer, 2013). Para más detalles de como usar estas funciones, a las cuales se da acceso desde en los argumentos META y FUNNEL, consultar la ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO

El ejemplo es el mismo que se explicó detalladamente en la sección de *details* de la función XII1.

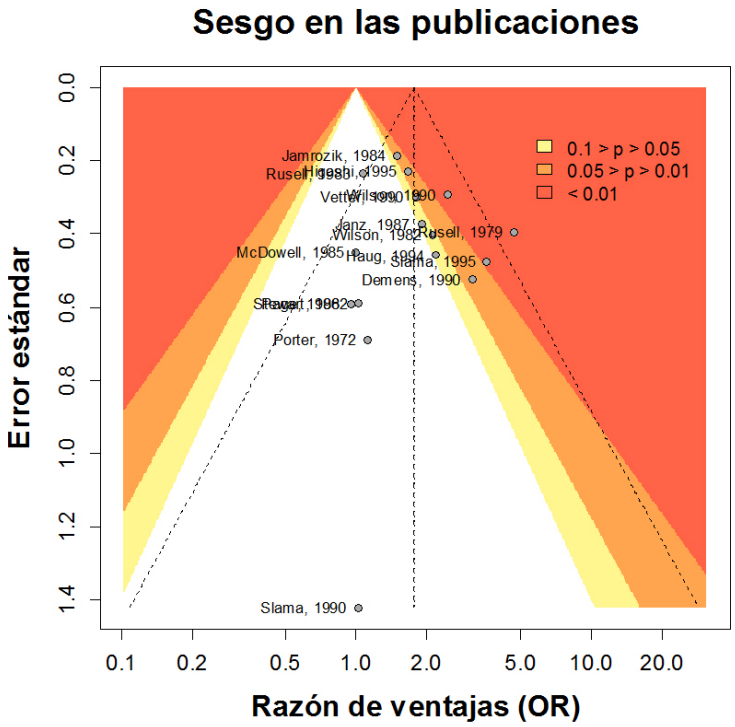
Todos los resultados del archivo TXT también se explicaron en la sección de *details* de la función XII1.

En el gráfico (Figura XII.4) se muestra, con líneas punteadas, el embudo al que se ajustan los datos, centrado en la *OR* estimada. Si los puntos encajan razonablemente en el embudo, sin asimetrías notables, significa que no existe evidencia de sesgo (así ocurre en el ejemplo).

También se muestran, con distintos tonos de colores, los límites de confianza (en este ejemplo 90%, 95% y 99%), basados en la hipótesis nula de que no existe efecto ($OR = 1$), es decir centrados los embudos en el valor 1.

Los puntos que están fuera de los límites de confianza muestran los estudios con efectos significativos para el nivel correspondiente

Figura XII.4. Diagrama de embudo de un estudio realizado sobre el efecto de la recomendación del médico en dejar de fumar.



Value

Se obtiene un archivo TXT con los resultados del meta-análisis, además el diagrama de embudo

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: `http://CRAN.R-project.org/package=meta`.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influencia del consejo médico para dejar de fumar.

XII4(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
studlab="Research", LEGEND= c("x=5", "y=0.1", "legend=c('0.1 > p > 0.05',
'0.05 > p > 0.01', '0 > p > 0.01')", "fill=COLOR", "bty='n'"), FUNNEL=c("level=0.95",
"contour.levels=c(0.9, 0.95, 0.99)", "studlab=TRUE", "col.contour = COLOR",
"xlab='Razón de ventajas (OR)'", "ylab='Error estándar'",
"font.lab=2",
"cex.lab=1.6", "cex.axis=1.2", "cex.main=1.8", "main='Sesgo en las publicaciones'"))


## End(Not run)
```

XII5 *META-ANALISIS/DIAGRAMA DE ANALISIS DE SENSIBILIDAD*

Description

Se aplica un meta-análisis y se muestra en un diagrama de análisis de sensibilidad.

Usage

```
XII5(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
event.e	Variable con el número de eventos en el grupo experimental.
n.e	Variable con el número total de observaciones en el grupo experimental.
event.c	Variable con el número de eventos en el grupo control.
n.c	Variable con el número total de observaciones en el grupo control.
studlab	Variable con el código de los estudios.
META	Se accede a la función metabin que permite modificar los argumentos del meta-análisis.
FOREST	Se accede a la función forest.meta que permite modificar los argumentos del gráfico del meta-análisis.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XII. META-ANÁLISIS

XII.5. DIAGRAMA DE ANALISIS DE SENSIBILIDAD

En este diagrama se realiza un análisis de sensibilidad para determinar la influencia de cada uno de los estudios en la estimación global del efecto y, por lo tanto, la robustez o estabilidad de la medida final obtenida en el meta-análisis. Por tanto, es un complemento importante del diagrama de bosque explicado en la sección de *details* de la función XII1, dentro de un estudio de meta-análisis. El método consiste en la repetición del meta-análisis tantas veces como estudios seleccionados, de forma que cada vez se omite un estudio. Si los resultados de los distintos meta-análisis son similares, es decir, el efecto tiene una misma dirección, magnitud y significación estadística, se puede concluir que el resultado del meta-análisis tiene solidez. El análisis de sensibilidad también se puede utilizar para ver que ocurre cuando se eliminan estudios que no han sido publicados, hay dudas sobre la calidad del estudio, etc.

FUNCIONES

Para el meta-análisis se usó la función metabin y para el diagrama de análisis de sensibilidad metainf y forest.meta, todas del paquete meta (Schwarzer, 2013). Para más detalles de como usar estas funciones, a las cuales se da acceso desde en los argumentos META y FOREST, consultar la ayuda de la función y/o Guisande & Vaamonde (2012).

EJEMPLO

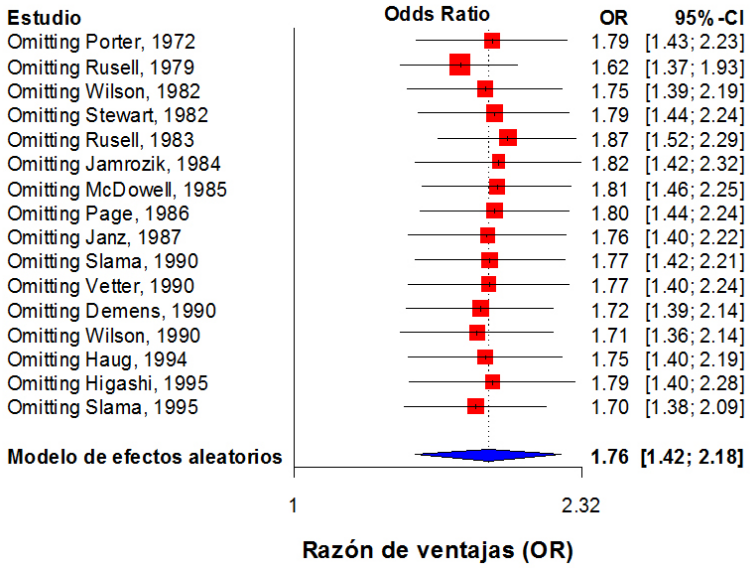
El ejemplo es el mismo que se explicó detalladamente en la sección de *details* de la función XII1.

Los resultados -similares a los explicados en las funciones anteriores- muestran la OR conjunta con su intervalo de confianza, obtenida omitiendo sucesivamente cada uno de los estudios, el valor *P* del contraste del efecto agregado, la varianza entre estudios *tau*^2 , y el estadístico de heterogeneidad *I*^2. En este ejemplo, la OR agregada cambia levemente (los cambios mayores se producen con la eliminación de los estudios Russel (1979) y Slama (1995) y en todos los casos se mantiene la heterogeneidad en valores reducidos, por lo que ninguno de los estudios tiene un efecto distorsionador grande sobre el resultado global.

```
Influential analysis (Random effects model)

                            OR           95%-CI  p.value    tau^2     I^2
Omitting Porter, 1972     1.7852 [1.4307; 2.2275] < 0.0001   0.054   31.2%
Omitting Rusell, 1979     1.6241 [1.3665; 1.9302] < 0.0001   0         0.0%
Omitting Wilson, 1982     1.7478 [1.3920; 2.1944] < 0.0001   0.0569  31.6%
Omitting Stewart, 1982    1.7946 [1.4404; 2.2359] < 0.0001   0.051   29.9%
Omitting Rusell, 1983     1.8656 [1.5214; 2.2878] < 0.0001   0.0203  12.9%
Omitting Jamrozik, 1984   1.8171 [1.4250; 2.3172] < 0.0001   0.0645  30.4%
Omitting McDowell, 1985   1.8111 [1.4565; 2.2520] < 0.0001   0.046   27.4%
Omitting Page, 1986       1.7970 [1.4445; 2.2355] < 0.0001   0.0491  29.1%
Omitting Janz, 1987       1.7600 [1.3981; 2.2157] < 0.0001   0.0593  32.3%
Omitting Slama, 1990      1.7722 [1.4218; 2.2090] < 0.0001   0.0553  32.1%
Omitting Vetter, 1990     1.7674 [1.3974; 2.2353] < 0.0001   0.062   32.5%
Omitting Demens, 1990     1.7233 [1.3871; 2.1411] < 0.0001   0.0464  27.8%
Omitting Wilson, 1990     1.7066 [1.3641; 2.1351] < 0.0001   0.0474  26.8%
Omitting Haug, 1994       1.7485 [1.3953; 2.1913] < 0.0001   0.0562  31.6%
Omitting Higashi, 1995    1.7867 [1.4032; 2.2750] < 0.0001   0.0663  32.5%
Omitting Slama, 1995      1.6998 [1.3793; 2.0947] < 0.0001   0.0358  22.8%

Pooled estimate           1.7621 [1.4225; 2.1828] < 0.0001   0.0478  27.7%

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

El diagrama de análisis de sensibilidad (Figura XII.5) muestra que solo el trabajo de Russel (1979) tiene una cierta influencia en el análisis, ya que cuando se elimina, la estimación global del efecto cambia más que con los otros estudios.

Figura XII.5. Diagrama de análisis de sensibilidad de un estudio realizado sobre el efecto de la recomendación del médico en dejar de fumar.



Value

Se obtiene un archivo TXT con los resultados del meta-análisis, además del diagrama de análisis de sensibilidad.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:
data(ZXII1)
```

```
#Influencia del consejo médico para dejar de fumar.

XII5(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
studlab="Research", FOREST=c("text.random='Modelo de efectos aleatorios'",
"leftlabs=c('Estudio')", "xlab='Razón de ventajas (OR)'", "col.square='red'",
"col.diamond='blue'"))

## End(Not run)
```

XIII1 *ANALISIS DE CORRESPONDENCIAS*

Description

Se aplica un Análisis de Correspondencias.

Usage

```
XIII1(data, var, cat=NULL, plot3d=TRUE, ResetPAR=TRUE, PAR=NULL,  xlim=NULL,
ylim=NULL, main=NULL, cex.main=1.7, font.main=2, dim1 = c(1,2), dim2 = c(1,2,3),
map = "symmetric", what = c("all", "all"), mass = c("FALSE", "FALSE"),
contrib=c("none", "none"), col=c("#0000FF","#FF0000"), labcol=c("#0000FF","#FF0000"),
pch = c(16, 1, 17, 24), labels = c(2, 2), arrows = c("FALSE", "FALSE"), LEGEND=NULL,
AXIS=NULL, MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis. Si son factores o caracteres se transforman a numéricas y se muestra, en los resultados, una tabla con el valor numérico asignado a cada categoría.
cat	Columna con las etiquetas de los casos.
plot3d	Si es TRUE se muestra el gráfico 3D con los tres ejes seleccionadas por el usuario. Este gráfico no admite acentos en las categorías de la variable *cat*
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
xlim	Límites del eje X.
ylim	Límites del eje Y.
main	Título del gráfico.
cex.main	Tamaño de letra del título del gráfico.
font.main	Tipo de fuente del título.

dim1	Vector con dos valores indicando los ejes que se representan en el gráfico bidimensional.
dim2	Vector con tres valores indicando los ejes que se representan en el gráfico 3D.
map	Define el tipo de mapa y puede ser alguna de las siguientes opciones: "symmetric", "rowprincipal", "colprincipal", "symbiplot", "rowgab", "colgab", "rowgreen" o "colgreen"
what	Vector con dos caracteres, que puede ser "all" o "none", que indican si se muestran los códigos de los casos (primer carácter del vector) y/o las columnas (segundo carácter del vector).
mass	Vector con dos valores lógicos TRUE o FALSE, indicando si la masa se representa por el área de los símbolos, para los casos (primer valor lógico) y/o las columnas (segundo valor lógico).
contrib	Vector de dos cadenas de caracteres que especifican si las contribuciones (relativa o absoluta) deben estar representadas por diferentes intensidades de color. Las opciones disponibles son "none" (las contribuciones no se indican en el diagrama), "absolute" (las contribuciones absolutas se indican con intensidades de color) o "relative" (las contribuciones relativas se indican con intensidades de color). Si se establece "absolute" o "relative", las contribuciones cero se muestran en blanco. Cuanto mayor sea la contribución de un punto, más cercana será al color definido en el argumento *col*.
col	Vector con colores para los casos (primer color) y las columnas (segundo color).
labcol	Vector con dos colores para los casos y las columnas en el gráfico 3D.
pch	Vector con 4 números que indican los símbolos, aunque se usan solamente el primero (para los casos) y el tercero (para las columnas).
labels	Vector con dos valores indicando si el gráfico debe incluir sólo símbolos (0), etiquetas solamente (1) o ambos, símbolos y etiquetas (2).
arrows	Vector de dos valores lógicos que especifican si el gráfico debe contener puntos (FALSO, por defecto) o flechas (VERDADERO). En primer valor establece las filas y el segundo valor establece el columnas.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XIII. ANÁLISIS MULTIVARIANTE FACTORIAL

XIII.1. ANÁLISIS DE CORRESPONDENCIAS

Es una técnica utilizada para representar las filas y las columnas de una tabla de contingencia de dos variables cualitativas como puntos en un espacio de dos dimensiones, de forma que se superponen las filas y las columnas para obtener una representación conjunta. En general es útil en grandes tablas de contingencia. Por ejemplo, si tenemos veinte especies de árboles y 15 bosques y hemos

hecho un recuento por muestreo, tendremos una tabla de contingencia con las especies en filas y los bosques en columnas. Cada casilla de la tabla nos dará la frecuencia relativa bidimensional para cada especie y bosque. El análisis de correspondencias nos permitirá representar las especies y los bosques en un mismo gráfico, de modo que podremos asociar especies con perfil de distribución similar, bosques afines, e identificar las especies características de cada bosque.

1. Perfil de filas y de columnas

Cada fila de la tabla establece un perfil de fila, definido por sus frecuencias.

1.1. Perfil de la fila *i*-ésima es la proporción de cada categoría de la variable *j* en la fila *i*:

$$r_{ij} = \frac{n_{ij}}{n_i}$$

El perfil de los totales de columna, es decir las frecuencias marginales de todas las columnas, define el perfil medio de las filas.

1.2. Perfil de la columna *j*-ésima es la proporción de cada categoría de la variable *i* en la columna *j*:

$$c_{ij} = \frac{n_{ij}}{n_j}$$

El perfil de los totales de fila, frecuencias marginales de fila, es el perfil medio de las columnas.

2. Masa de cada elemento o perfil

La masa asociada a cada elemento fila (r_i) o columna (c_j) mide la importancia relativa de ese elemento en el conjunto, y es igual a su frecuencia marginal dividida por el total general o tamaño de muestra *n*, es decir, el correspondiente elemento del perfil medio de columnas, mientras que la masa de cada perfil columna será el correspondiente elemento del perfil medio de las filas.

$$r_i = \tfrac{n_i}{n} = f_{i.} \quad c_j = \tfrac{n_j}{n} = f_{.j}$$

Al representar los perfiles fila/columna deberemos dar más importancia a los que tengan mayor masa.

La mejor medida de distancia entre perfiles (que vienen definidos por las frecuencias) no es la habitual distancia euclídea, sino la distancia Chi-cuadrado, ya que esta distancia tiene una propiedad importante de equivalencia distribucional, según la cual si agregamos dos perfiles iguales, el perfil agregado que sustituye a ambos conserva la distancia de los anteriores con cualquier otro perfil, y no cambia ninguna otra distancia. Como la agregación de filas y columnas es habitual en la construcción de tablas de contingencia, y a menudo se utilizan criterios subjetivos para decidir el número y los límites de las categorías o perfiles, esta propiedad adquiere un interés especial. Por esta razón se realizan algunas transformaciones simples en los datos para obtener esta distancia Chi cuadrado.

Distancia Chi cuadrado entre perfiles fila:

$$d^2_{\chi^2}(r_i, r_{i'}) = \sum_{j=1}^{J} \frac{(r_{ij} - r_{i'j})^2}{c_j}$$

Distancia Chi cuadrado entre perfiles columna:

$$d^2_{\chi^2}(c_j, c_{j'}) = \sum_{i=1}^{I} \frac{(c_{ij} - c_{ij'})^2}{r_i}$$

Para construir el gráfico se utiliza el método de Componentes Principales, con el cual se obtienen generalmente dos dimensiones.

3. Contribuciones absolutas y relativas

3.1. Contribuciones absolutas: Parte de la variabilidad de cada eje o dimensión (componente principal) que es explicada por cada perfil. Sirven para interpretar los ejes en función de los perfiles, observando cuales son los perfiles que tienen más peso -o contribución absoluta- en la construcción de cada dimensión.

3.2. Contribuciones relativas: Parte de la variabilidad de cada perfil explicado por cada eje o componente. Indican hasta qué punto cada perfil está adecuadamente explicado o representado por los dos ejes o dimensiones. Los perfiles fila o columna que tienen pequeñas contribuciones relativas en ambos ejes no responden al modelo, y están deficientemente representados en el gráfico, lo que debe tenerse en cuenta en la interpretación del mismo.

FUNCIONES

El análisis de correspondencias se realiza con la función ca y el gráfico con la función plot.ca, ambas del paquete ca (Greenacre & Pardo, 2006; Greenacre, 2007; Nenadic & Greenacre, 2007; Greenacre, 2013).

EJEMPLO

Datos de criterios de valoración, en una escala de 0 a 10, de 48 candidatos que optan a un puesto de trabajo. El objetivo es determinar si, con un Análisis de Correspondencias, es posible identificar cuales son las cualidades que permiten diferenciar mejor a los candidatos, para facilitar el proceso de selección.

Paso 1.

Las Figuras XIII.1 y XIII.2 muestran que en el extremo derecho del primer eje está la experiencia junto con la adecuación al puesto y carta de motivación. En el extremo izquierdo del primer eje están asociadas la motivación y confianza. El segundo eje viene asociado al sentido comercial en el extremo superior, mientras que en el inferior están a la honestidad y nivel de estudios.

Figura XIII.1. Análisis de Correspondencias 2D mostrando la variabilidad observada en las cualidades de candidatos a un puesto de trabajo.

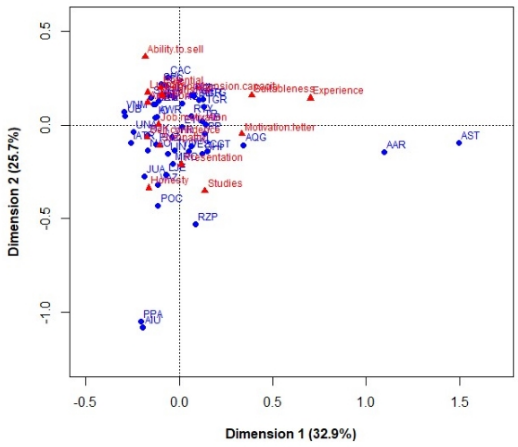
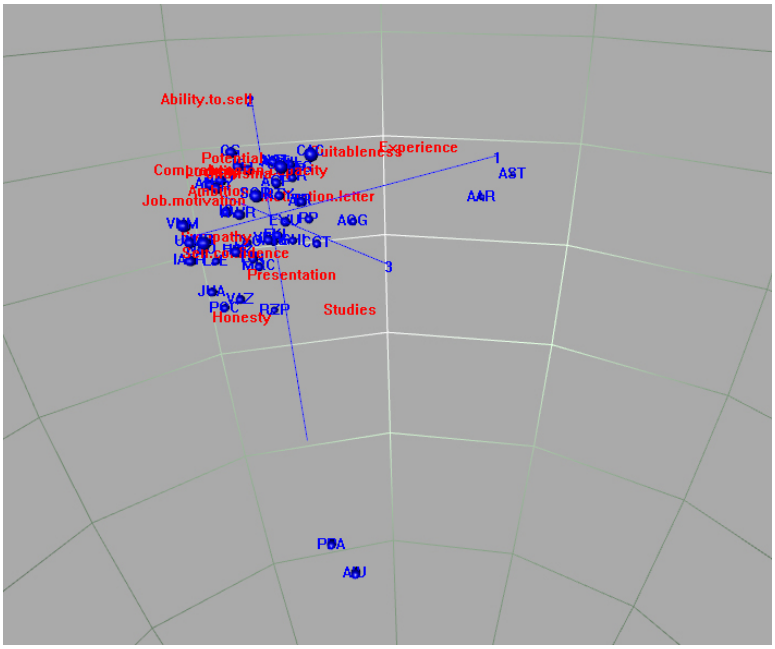


Figura XIII.2. Análisis de Correspondencias 3D mostrando la variabilidad observada en las cualidades de candidatos a un puesto de trabajo.



Los autovalores muestran el porcentaje de varianza de los ejes: 32,9% primer eje, 25,7% el segundo, etc.

```
Principal inertias (eigenvalues):

 dim     value       %    cum%   scree plot
 1       0.054749  32.9   32.9   ************************
 2       0.042711  25.7   58.6   ******************
 3       0.018796  11.3   69.9   ********
 4       0.013157   7.9   77.9   ******
 5       0.008201   4.9   82.8   ****
 6       0.006548   3.9   86.7   ***
 7       0.006122   3.7   90.4   ***
 8       0.004844   2.9   93.3   **
 9       0.003894   2.3   95.7   **
 10      0.002278   1.4   97.0   *
 11      0.002017   1.2   98.3   *
 12      0.001380   0.8   99.1
 13      0.000908   0.5   99.6
 14      0.000619   0.4  100.0
         -------- -----
 Total:  0.166223 100.0
```

La inercia (*Inertia*) indica el peso relativo o la distinta importancia de cada variable. Los mayores valores son los de la experiencia (0,0271), nivel de estudios (0,0168), honestidad (0,0155), carta

de motivación (0,015), adecuación (0,0149), sentido comercial (0,012, *Ability.to.sell*), etc. Los resultados también muestran las coordenadas de las variables en los ejes I (Dim. 1) y II (Dim. 2), lo cual permite ubicar cada variable en la gráfica mostrada anteriormente y poder disponer de las coordenadas para realizar cualquier otro tipo de representación.

```
 Columns:
        Motivation.letter Presentation   Studies  Sympathy Self.confidence
Mass             0.065619     0.077466  0.077466  0.067213        0.075871
ChiDist          0.478458     0.301583  0.465896  0.366283        0.267700
Inertia          0.015022     0.007046  0.016815  0.009018        0.005437
Dim. 1           1.435655     0.048741  0.589168 -0.449309       -0.730234
Dim. 2          -0.235968    -1.024413 -1.706546 -0.533872       -0.313194
         Lucidity   Honesty Ability.to.sell Experience  Charisma  Ambition
Mass     0.069036  0.087947        0.053087   0.046252  0.058100  0.065391
ChiDist  0.336224  0.420710        0.475928   0.766059  0.344512  0.309461
Inertia  0.007804  0.015566        0.012025   0.027143  0.006896  0.006262
Dim. 1  -0.722219 -0.692624       -0.771407   3.002571 -0.389261 -0.725739
Dim. 2   0.843376 -1.630336        1.760155   0.688959  0.738588  0.581699
        Comprehension.capacity Potential Job.motivation Suitableness
Mass                  0.068353  0.062201       0.060834     0.065163
ChiDist               0.289309  0.340237       0.391699     0.478734
Inertia               0.005721  0.007200       0.009334     0.014934
Dim. 1               -0.392606 -0.429611      -0.478928     1.651207
Dim. 2                0.800778  0.967160       0.018634     0.760602
```

El instrumento más importante en el análisis de correspondencia es un gráfico, en el cual los objetos fila y columna se relacionan mediante distancias y proximidades que pueden ser interpretadas. Entre los resultados se muestran las distancias (*ChiDist*) al perfil medio.

La masa (*mass*) asociada a cada elemento fila o columna -concepto similar y alternativo al de inercia- mide la importancia de ese elemento en el conjunto, y es igual a su frecuencia relativa marginal. Al representar los perfiles fila-columna deberemos dar más importancia en la interpretación a los que tengan mayor masa.

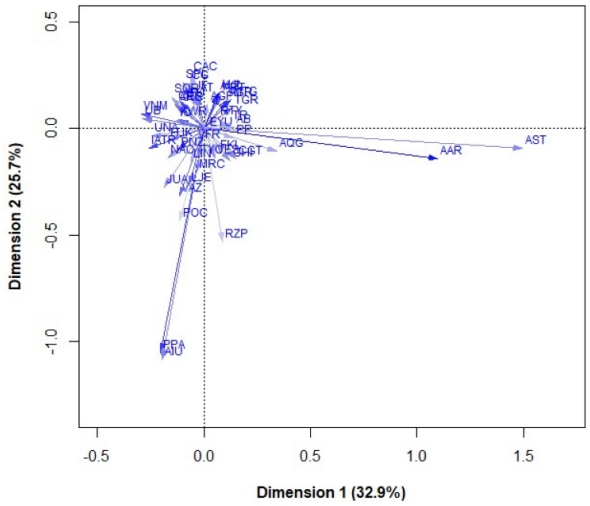
También se muestra la inercia de los candidatos, que permitirá escoger al mejor candidato en función de la cualidad a la que se le de prioridad, así como las coordenadas en los ejes I (Dim. 1) y II (Dim. 2).

```
 Rows:
                CG         AT         JF         PP         AB         TR         MTG
Mass      0.022784   0.028936   0.025290   0.020278   0.022784   0.023468   0.029392
ChiDist   0.329400   0.207207   0.258433   0.296582   0.265844   0.219732   0.210780
Inertia   0.002472   0.001242   0.001689   0.001784   0.001610   0.001133   0.001306
Dim. 1   -0.303331  -0.131564  -0.213101   0.587658   0.610982   0.519180   0.545789
Dim. 2    1.023682   0.740720   0.785052  -0.217328   0.020578   0.110352   0.656200
             AGR        TGR        SPL        CAC        SQR        FKL        GHI
Mass     0.030759   0.027797   0.026202   0.022784   0.024607   0.021189   0.020506
ChiDist  0.191141   0.199670   0.398671   0.499780   0.347427   0.272508   0.313858
Inertia  0.001124   0.001108   0.004164   0.005691   0.002970   0.001574   0.002020
Dim. 1   0.454288   0.572825  -0.403952  -0.256931  -0.648240   0.275567   0.521933
Dim. 2   0.640714   0.463348   1.047138   1.219957   0.701377  -0.554786  -0.741527
```

Paso 2.

Se realiza el Análisis de Correspondencias mostrando solamente los candidatos (los casos), el cual se muestra en la Figura XIII.3.

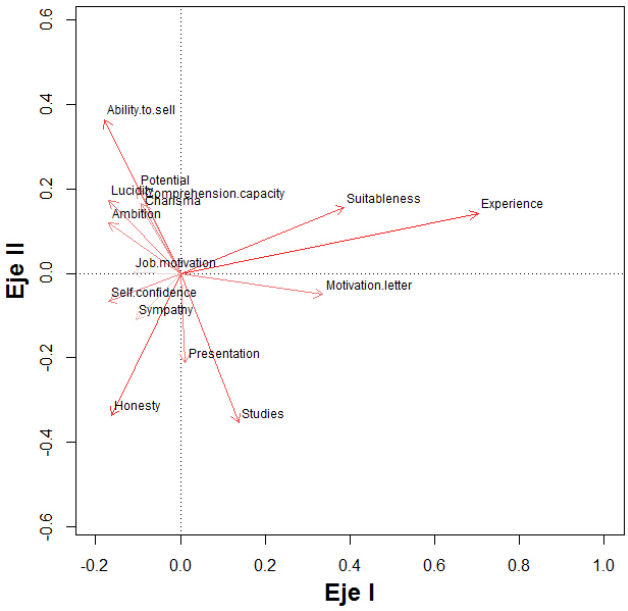
Figura XIII.3. Análisis de Correspondencias mostrando la variabilidad observada en los candidatos a un puesto de trabajo.



Paso 3.

Se realiza el Análisis de Correspondencias mostrando solamente las cualidades de los candidatos (las columnas), el cual se muestra en la Figura XIII.4.

Figura XIII.4. Análisis de Correspondencias mostrando la variabilidad observada en las cualidades.



Si para seleccionar al candidato nos basamos en las variables que más influían en el primer eje en su extremo derecho (experiencia y adecuación), los candidatos que tienen las coordenadas más positivas en este primer eje son AST y AAR. Si por el contrario, nos basamos para la selección en las variables que mejor discriminaban el segundo eje, como por ejemplo sentido comercial que estaba en el extremo positivo superior del eje, entonces el candidato seleccionado debería ser CAC. Si elegimos la honestidad, el candidato elegido debería ser PPA o AIU.

Value

Se obtiene un archivo TXT con los resultados del Análisis de Correspondencias, un gráfico con las puntuaciones y el gráfico 3D si es seleccionado.

References

Greenacre, M. (2013). Simple, Multiple and Joint Correspondence Analysis. R package version 0.53. Available at: http://CRAN.R-project.org/package=ca.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC.

Greenacre, M.J. & Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, 35: 193-218.

Nenadic, O. & Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20: 1-13.

Examples

```
## Not run:

data(ZXIII1)

#Paso 1
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy","Self.confidence", "Lucidity","Honesty","Ability.to.sell","Experience",
"Charisma","Ambition","Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), xlim=c(-0.5,1.7))

#Paso 2
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy","Self.confidence", "Lucidity","Honesty","Ability.to.sell","Experience",
"Charisma","Ambition", "Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), what = c("all", "none"), mass = c("TRUE","FALSE"), contrib = "relative",
arrows = c("TRUE", "FALSE") , xlim=c(-0.5,1.7))

#Paso 3
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy", "Self.confidence", "Lucidity","Honesty","Ability.to.sell", "Experience",
"Charisma","Ambition","Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), xlim=c(-0.2, 1), what = c("none", "all"), mass = c("FALSE","TRUE"),
contrib = "relative",  arrows = c("FALSE", "TRUE"))
```

```
## End(Not run)
```

XIII2 *ANALISIS DE COMPONENTES PRINCIPALES*

Description

Se aplica un Análisis de Componentes Principales.

Usage

```
XIII2(data, var, cat=NULL, labels=NULL, ellipse=FALSE, convex=FALSE, dim=c(1,2),
VIF=FALSE, threshold=10,  ResetPAR=TRUE, PAR=NULL, PCA=NULL, SCATTERPLOT=NULL,
COLOR=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL, LEGEND=NULL,
AXIS=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL, arrows=TRUE, larrow=0.7,
colArrows="black", file1="Output.txt", file2="Var loadings.csv",
file3="Cat loadings.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis.
cat	Categorías de los casos.
labels	Variable que permite mostrar una etiqueta para cada caso.
ellipse	Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *cat*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.
convex	Si es TRUE, la envolvente convexa se calcula para cada categoría.
dim	Vector con dos valores indicando los ejes que se representan en el gráfico.
VIF	Si es TRUE se utiliza el factor de inflación de la varianza (VIF) para seleccionar las variables altamente correlacionadas y, por tanto, las variables no correlacionadas son excluidas del análisis.
threshold	Valor de corte para el VIF.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
PCA	Se accede a la función prcomp del paquete stats.
SCATTERPLOT	Se accede a la función scatterplot del paquete car.
COLOR	Permite modificar los colores del gráfico, pero deben ser tantos como grupos diferentes tenga la variable *cat*.

PCH	Vector con los símbolos del gráfico, que deben ser tantos como grupos diferentes tenga la variable *cat*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
XLIM, YLIM	Vectores con los límites de los ejes *X* e *Y*.
XLAB, YLAB	Leyendas de los ejes *X* e *Y*.
LEGEND	Permite incorporar o modificar la leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXTvar	Se accede a la función que permite modificar las etiquetas de las variables del gráfico *biplot*.
TEXTlabels	Se accede a la función que permite modificar las etiquetas de los casos en el gráfico *biplot*.
arrows	Si es TRUE se muestran las flechas en el scatterplot.
larrow	Permite modificar la longitud de las flechas.
colArrows	Colores de las flechas.
file1	ARCHIVO TXT. Nombre del archivo de salida con los resultados.
file2	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las variables en el gráfico.
file3	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de los casos en el gráfico.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XIII. ANÁLISIS MULTIVARIANTE FACTORIAL

XIII.2. COMPONENTES PRINCIPALES

Consiste en la transformación del conjunto de variables originales en otro conjunto de variables -Componentes Principales- obtenidas como combinación lineal de aquellas. Las nuevas variables, en igual número que las antiguas, conservan toda la información -en lo que se refiere a la variabilidad del conjunto de objetos- de las variables primitivas, pero la mayoría de las componentes principales tienen una variabilidad tan pequeña que pueden ser ignoradas, de tal manera que unas pocas componentes -generalmente tres o menos- permiten representar y explicar razonablemente el conjunto de objetos de la muestra sin pérdida importante de información.

La principal virtud del Análisis de Componentes Principales (ACP) consiste precisamente en la reducción de la complejidad de los datos, al pasar de muchas variables a pocas (frecuentemente una o dos), que pueden ser representadas gráficamente. Frente a esta ventaja, el inconveniente de las nuevas variables está en la pérdida parcial de información (a menudo pequeña) y en el hecho de que las variables originales tienen un significado real del que carecen en general las componentes, que son mezcla o combinación de variables.

La primera componente principal es la combinación lineal de las variables originales que tiene máxima varianza. La segunda componente es la combinación lineal de las variables originales con máxima varianza con la condición añadida de que sea independiente de la primera (ortogonal), y así sucesivamente se pueden obtener todas las componentes principales, que siendo independientes entre sí contienen información distinta, ya que la independencia o ausencia de correlación significa que las nuevas variables o componentes no comparten información común. Cada componente principal explica por lo tanto la máxima variabilidad residual posible (la que no han explicado ya las anteriores).

Para encontrar las combinaciones óptimas de las variables originales deben ser calculados los autovalores y autovectores de la matriz de covarianzas (o de la matriz de correlaciones). Los autovalores, ordenados de mayor a menor, son las varianzas de las nuevas componentes, y los autovectores expresan las combinaciones lineales que definen las componentes principales. Debemos decidir cuantas componentes deben ser conservadas: un número alto permite explicar una mayor proporción de la variabilidad total, y un número bajo permite obtener una mayor simplicidad en la representación gráfica. Existen varios criterios que pueden ayudarnos a la hora de pronunciarnos:

- Criterio de la varianza. Se seleccionan las componentes que explican un porcentaje suficiente (por ejemplo 80% o 90%) de la varianza.

- Criterio de Kaiser. Se conservan las componentes cuyas varianzas (autovalores) son mayores que la varianza media.

- Gráfico de sedimentación (Scree plot). Representamos los autovalores, que en general decrecen rápidamente al principio y lentamente después, y conservamos únicamente los que se sitúan por encima del codo de la curva.

- Dos componentes. Mantener solamente las dos primeras permite construir un gráfico en dos dimensiones, generalmente el más adecuado para describir el conjunto de datos.

La única condición para que pueda aplicarse el Análisis de Componentes Principales es que las variables sean cuantitativas. Sin embargo, existe otra condición o hipótesis necesaria para que sea útil su aplicación: las variables originales deben estar correlacionadas. Cuanto más alta sea la correlación entre las variables, mayor es la proporción de varianza explicada por las primeras componentes y, por lo tanto, más pequeña será la pérdida de información al conservarlas prescindiendo de las restantes. Para comprobar si nuestras variables están correlacionadas existen varios procedimientos:

- Inspeccionar la matriz de correlaciones. Podremos observar si la mayoría de los coeficientes de correlación lineal entre las variables originales son próximos a 1 o a -1.

- Inspeccionar la matriz de valores *p* del contraste de nulidad de los coeficientes de correlación. Se contrasta -para cada uno de los coeficientes de correlación- la hipótesis de que es cero. El valor *p* debe ser menor que el nivel de significación (habitualmente 0,05) para rechazar esa hipótesis. De este modo, la matriz de valores *p* cuando existe correlación debe estar formada mayoritariamente por ceros o valores muy pequeños.

- Determinante de la matriz de correlaciones. Toma un valor próximo a 1 cuando no existe correlación y próximo a cero cuando existe.

- Test de esfericidad de Bartlett. Está basado en el determinante de la matriz de correlaciones y contrasta la hipótesis de "esfericidad" o ausencia de correlación con un único test. Un valor *p* del contraste menor que el nivel de significación elegido permite rechazar la hipótesis y concluir que existe correlación.

Cuando las variables originales son heterogéneas, es decir, vienen expresadas en unidades de medida diferentes, ocurre que la variabilidad viene determinada por la elección de la unidad de medida

de cada una de ellas, lo que influye en los resultados de la aplicación del ACP al asignar una importancia arbitraria a cada variable. Para evitar esta influencia negativa, la solución consiste en utilizar la matriz de correlaciones en lugar de la matriz de covarianzas, lo que equivale a tipificar las variables originales. De este modo todas las variables pasan a tener media 0 y desviación típica 1. Como este proceso implica una cierta pérdida de información, no debemos utilizar la matriz de correlaciones en sustitución de la matriz de covarianzas en el caso de que nuestras variables sean homogéneas.

Una vez obtenidas las combinaciones lineales que definen a las componentes principales, podemos interpretar los coeficientes (saturaciones o cargas): cuanto más alto sea su valor absoluto, mayor influencia tiene esa variable en la construcción de la componente. Es posible interpretar las componentes, y asignarles significados genéricos, a partir del conjunto de variables con mayor peso en su construcción.

Para facilitar esta interpretación de las componentes y atribuirles algún significado puede ser útil el gráfico de saturaciones, que muestra -en el plano de las dos primeras componentes- los coeficientes o saturaciones para cada variable, de modo que cada variable original se representa como un punto en ese plano. Las variables que están más alejadas con respecto al cero del eje horizontal son las que dan significado a la primera componente (que a menudo expresa una idea genérica de "tamaño"), y las que están más alejadas respecto al cero del eje vertical son las que dan sentido a la segunda componente (a menudo refleja una idea genérica de "forma"). Las variables próximas al centro de coordenadas no tienen interés, y las que se alejan de ambos ejes son comunes a ambas componentes y tampoco ayudan a su interpretación. El gráfico puede hacerse también para otras componentes distintas de las dos primeras.

El principal resultado de esta técnica es en general el gráfico de puntuaciones. Consiste en la representación, en el plano de las dos primeras componentes, de las coordenadas de los puntos que representan a los elementos del conjunto objeto (las filas o casos en nuestros datos). Esta representación permite describir el conjunto de datos multidimensionales de un modo relativamente sencillo. Puede ser útil añadir al gráfico etiquetas de caso o marcas.

Las variables originales tienen que estar correlacionadas para que el método funcione bien. La utilización del factor de inflación de la varianza (VIF) que se propone opcionalmente como criterio para elegir las variables que debemos utilizar en el ACP permite excluir aquellas que están escasamente correlacionadas con las restantes, que en general aparecerán asociadas a componentes principales poco importantes (distintas de las primeras), aunque no necesariamente es siempre así, por lo que esta selección automática de variables debe utilizarse con reservas, únicamente como una ayuda en la selección de las variables a utilizar.

FUNCIONES

El Análisis de Componentes Principales se realizó con la función prcomp del paquete stats. Para el cáculo del VIF se usó la función vif del paquete usdm (Naimi, 2013; Naimi et al., 2014). Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2014). Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014). La envolvente convexa se estima con la función chull del paquete grDevices. El test KMO se realiza con la función KMO del paquete psych (Revelle, 2018). El test de esfericidad de Bartlett se realizó con la función bart_spher del paquete REdaS (Maier, 2015).

EJEMPLO

El estudio consistió en analizar los parámetros demográficos de 57 países de Europa, África y América. Las variables usadas fueron la esperanza de vida masculina y femenina al nacer (en

años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el producto interior bruto per cápita (en miles de dólares anuales) y la tasa de alfabetización de hombres y mujeres (en porcentaje) en el año 2000. Los datos fueron obtenidos del Banco Mundial (http://datos.bancomundial.org/). El objetivo es determinar cuales son los parámetros demográficos que presentan una mayor variablidad entre los países y, por tanto, son responsables de las diferencias que existen entre ellos.

Paso 1.

En primer lugar se utilizan todas las variables *var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF")* y como categorías o variable de clasificación se utiliza el continente *cat="Continent"*. Se realiza el análisis eliminando las variables que no están correlacionadas, para lo que debemos especificar *VIF=TRUE*. Eso significa que se van a eliminar aquellas con VIF por debajo del umbral seleccionado, que por defecto es *threshold=10*. Por ello, la mortalidad y el PIB serán eliminados de forma automática.

```
"VIF values"

  Variables        VIF
1  LifeExpF 150.016016
2  LifeExpM  85.308609
3   Mortinf  18.205941
4   PIB_cap   5.646459
5 Birthrate  51.549152
6 Mortality   9.478087
7 Fertility  36.521744
8 LiteracyM  26.794925
9 LiteracyF  36.397156

Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf Birthrate Fertility LiteracyM LiteracyF
     0.83      0.84      0.98      0.86      0.86      0.81      0.83

Bartlett's Test of Sphericity
Call: REdaS::bart_spher(x = datos1)

     X2 = 909.413
     df = 21
p-value < 2.22e-16

"Correlation matrix"

          LifeExpF LifeExpM Mortinf Birthrate Fertility LiteracyM LiteracyF
LifeExpF    *****    0.993   -0.962   -0.921    -0.926     0.802     0.859
LifeExpM   <0.001    *****   -0.958   -0.900    -0.904     0.794     0.849
Mortinf    <0.001   <0.001    *****    0.912     0.915    -0.850    -0.888
Birthrate  <0.001   <0.001   <0.001    *****     0.978    -0.836    -0.867
Fertility  <0.001   <0.001   <0.001   <0.001     *****    -0.846    -0.887
LiteracyM  <0.001   <0.001   <0.001   <0.001    <0.001     *****     0.975
LiteracyF  <0.001   <0.001   <0.001   <0.001    <0.001    <0.001     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values

"Summary PCA"

Importance of components:
                          PC1     PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5259 0.60357 0.4073 0.21010 0.15900 0.12330 0.07151
Proportion of Variance 0.9114 0.05204 0.0237 0.00631 0.00361 0.00217 0.00073
Cumulative Proportion  0.9114 0.96348 0.9872 0.99349 0.99710 0.99927 1.00000
```

En los resultados lo primero que se muestra son los valores de VIF. Como ya se mencionó anteriormente, las variables originales deben estar correlacionadas, para conseguir una mayor proporción de varianza explicada por las primeras componentes.

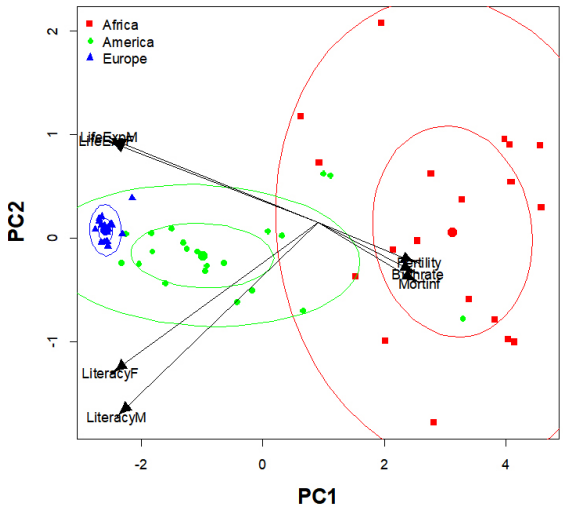
El siguiente resultado es el test KMO, que nos indica si las variables son adecuadas para el Componentes Principales. El valor debe ser mayor de 0,5. Por tanto, todas las variables que no tengan un valor mayor a 0,5, podrían ser eliminadas del análisis. En el caso de que el valor sea exactamente 0,5 es que no fue posible estimar el KMO.

Otro estadístico que parece es el test de esfericidad de Bartlett, el cual está basado en el determinante de la matriz de correlaciones y contrasta la hipótesis de "esfericidad" o ausencia de correlación con un único test. Un valor *p* del contraste menor que el nivel de significación permite rechazar la hipótesis y concluir que existe correlación. Por tanto, para que el Componentes Principales sea válido, la probabilidad deber ser menor de 0,05, como es en este caso.

Como podemos observar en la matriz de valores *p*, todos son menores que 0,05 lo que indica que ninguno de los coeficientes de correlación parece ser nulo. Por tanto, todas las variables están correlacionadas. Los resultados también muestran la contribución relativa de cada eje. El primer eje explica el 91,1%, el segundo el 5,2% y el tercero el 0,02% de la variación observada. Los dos primeros ejes explican el 96,3% de la varianza.

En la Figura XIII.5 se observa que en el primer eje se diferencian perfectamente los países de Europa, América y África. Los de Europa se caracterizan por alta esperanza de vida y alta alfabetización, tanto en hombres como en mujeres. Los de Africa por su alta mortalidad infantil y alta tasa de natalidad. Con el argumento *ellipse=TRUE* se representan las elipses, las cuales muestran los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías. Es un indicador del grado de solapamiento entre categorías. Estos niveles se pueden modificar entrando en la función scatterplot usando el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.

Figura XIII.5. Análisis de Componentes Principales mostrando la variabilidad observada en los parámetros demográficos de los 57 países analizados.



Paso 2.

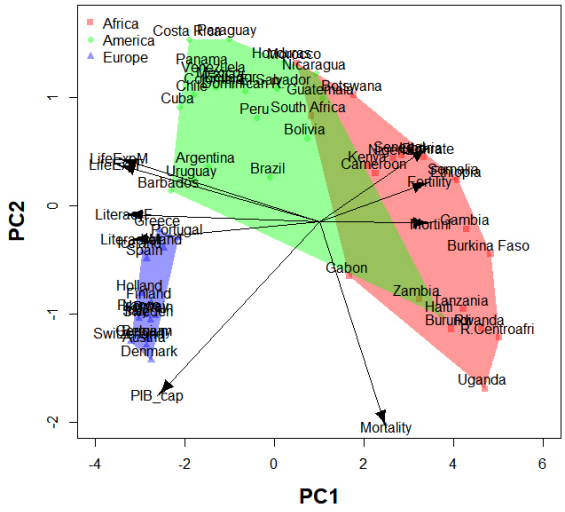
A continuación se realiza el análisis pero sin eliminar ninguna variable y, para ello, simplemente dejamos la opción por defecto *VIF=FALSE*, es decir, no introducimos el argumento ya que está por defecto esa opción. Además, con *labels="Country"* se especifica que las etiquetas de los casos sean

los nombres de los países. En la figura se utiliza la envolvente convexa en vez de las elipses con los argumentos *convex=TRUE* y *ellipse=FALSE*.

```
"VIF values"

"No estimated"

Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf   PIB_cap Birthrate Mortality Fertility LiteracyM
     0.83      0.89      0.99      0.85      0.85      0.77      0.88      0.83
LiteracyF
     0.84

Bartlett's Test of Sphericity

Call: REdaS::bart_spher(x = datos1)

     X2 = 1084.808
     df = 36
p-value < 2.22e-16

[Correlation matrix"

          LifeExpF LifeExpM Mortinf PIB_cap Birthrate Mortality Fertility LiteracyM LiteracyF
LifeExpF   *****    0.993   -0.962   0.673  -0.921     -0.774    -0.926     0.802     0.859
LifeExpM  <0.001    *****   -0.958   0.659  -0.900     -0.778    -0.904     0.794     0.849
Mortinf   <0.001   <0.001    *****  -0.685   0.912      0.704     0.915    -0.850    -0.888
PIB_cap   <0.001   <0.001   <0.001   *****  -0.795     -0.152    -0.716     0.620     0.606
Birthrate <0.001   <0.001   <0.001  <0.001   *****      0.538     0.978    -0.836    -0.867
Mortality <0.001   <0.001   <0.001   0.260  <0.001      *****     0.611    -0.508    -0.587
Fertility <0.001   <0.001   <0.001  <0.001  <0.001     <0.001     *****    -0.846    -0.887
LiteracyM <0.001   <0.001   <0.001  <0.001  <0.001     <0.001    <0.001     *****     0.975
LiteracyF <0.001   <0.001   <0.001  <0.001  <0.001     <0.001    <0.001    <0.001     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values

"Summary PCA"

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.7144 0.9615 0.6552 0.37512 0.26068 0.18871 0.13331 0.10906 0.06650
Proportion of Variance 0.8186 0.1027 0.0477 0.01563 0.00755 0.00396 0.00197 0.00132 0.00049
Cumulative Proportion  0.8186 0.9214 0.9691 0.98471 0.99226 0.99621 0.99819 0.99951 1.00000
```

Como era de esperar, en la matriz de correlación hay valores de p > 0,05, lo cual significa que no todas las variables están correlacionadas. Además se observa que el primer eje explica el 81,86% de la variabilidad observada, el segundo el 10,2% y el tercero el 4,77%. Los dos primeros ejes explican el 92,1% de la varianza, lo cual es inferior, como era de esperar, que en el caso anterior donde se usaron solamente las variables correlacionadas (96,3%). De nuevo se separan bien los continentes (Figura XIII.7), donde Europa se diferencia por un mayor PIB.

Figura XIII.7. Análisis de Componentes Principales 2D mostrando la variabilidad observada en los parámetros demográficos, sin eliminar los no correlacionados, de los 57 países analizados.



Value

Se obtiene un archivo TXT con el VIF, el factor de adecuación Kaiser-Meyer-Olkin (KMO), el test de esfericidad de Bartlett, las correlaciones entre variables, y los resultados del Análisis de Componentes Principales. Además se obtiene un gráfico 2D representando las variables y categorías.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Maier, M.J. (2015) Companion Package to the Book 'R: Einführung durch angewandte Statistik. R package version 0.9.3. Available at: http://CRAN.R-project.org/package=REdaS.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Revelle,W. (2018) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.8.4. Available at: http://CRAN.R-project.org/package=psych.

Examples

```
## Not run:
data(ZXIII2)

#Paso 1
XIII2(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent",
```

```
VIF=TRUE, ellipse=TRUE)

#Paso 2
XIII2(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf",
"PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent",
labels="Country", convex=TRUE, XLIM=c(-4,6), YLIM=c(-2,1.7))

## End(Not run)
```

XIII3 *ESCALAMIENTO MULTIDIMENSIONAL*

Description

Se aplica un Escalamiento Multidimensional.

Usage

```
XIII3(data, var, cat=NULL, ellipse=FALSE, convex=FALSE, dim=c(1,2), ResetPAR=TRUE,
PAR=NULL, METAMDS=NULL, SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL, XLIM=NULL,
YLIM=NULL,  XLAB=NULL, YLAB=NULL, TEXT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
arrows=TRUE, larrow=0.95, colArrows="black", file1="Var loadings.csv",
file2="Cat loadings.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis.
cat	Categorías de los casos.
ellipse	Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *cat*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.
convex	Si es TRUE, la envolvente convexa se calcula para cada categoría.
dim	Vector con dos valores indicando los ejes que se representan en el gráfico.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
METAMDS	Se accede a la función metaMDS del paquete vegan.
SCATTERPLOT	Se accede a la función scatterplot del paquete car.
COLOR	Permite modificar los colores del gráfico, pero deben ser tantos como grupos diferentes tenga la variable *cat*.

PCH	Vector con los símbolos del gráfico, que deben ser tantos como grupos diferentes tenga la variable *cat*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
XLIM, YLIM	Vectores con los límites de los ejes *X* e *Y*.
XLAB, YLAB	Leyendas de los ejes *X* e *Y*.
TEXT	Se accede a la función que permite modificar las etiquetas de los casos en el gráfico *biplot*.
LEGEND	Permite incorporar o modificar la leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
arrows	Si es TRUE se muestran las flechas en el scatterplot.
larrow	Permite modificar la longitud de las flechas.
colArrows	Color de las flechas.
file1	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las variables en el gráfico.
file2	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las filas o casos, añadiendo las categorías, en el gráfico.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XIII. ANÁLISIS MULTIVARIANTE FACTORIAL

XIII.3. ESCALAMIENTO MULTIDIMENSIONAL

Trata de encontrar la estructura subyacente en un conjunto de medidas de proximidad/distancia entre objetos. Los objetos pueden ser de cualquier tipo: Países, marcas comerciales, relaciones de parentesco, empresas, descriptores de calidad, etc. Las medidas que permiten establecer distancias o proximidades entre objetos también pueden ser de distinto tipo, cuantitativas o cualitativas, reales o supuestas, objetivas o subjetivas. Se suele aplicar esta técnica al estudio de percepciones, de gran interés en psicología, marketing, control de calidad, y otras áreas.

Si tenemos un mapa, es relativamente sencillo construir una tabla de distancias entre ciudades. El escalamiento multidimensional trata de resolver el problema inverso: a partir de una tabla de distancias se construye un mapa, que permite situar los objetos e interpretar gráficamente las relaciones existentes entre ellos. Se asigna cada observación a una posición específica en un espacio de reducidas dimensiones, generalmente dos, de tal modo que las distancias entre los puntos se correspondan del mejor modo posible con las distancias entre los objetos. El resultado principal de esta técnica es por lo tanto un gráfico.

El criterio o medida de ajuste más importante utilizado en el escalamiento multidimensional es el Stress (Kruskal), que es un promedio de desviaciones entre las distancias finales en el mapa y las

distancias o disimilaridades iniciales, normalizado para que tome valores entre 0 y 1.

$$S = \frac{\sum \sum \left(d_{ij} - \bar{d}_{ij}\right)^2}{\sum \sum d_{ij}^2}$$

Cuanto más grande es el stress, peor es el ajuste, considerándose generalmente que un valor superior a 0,2 aconseja no utilizar esta técnica. A menudo se expresa en porcentaje: el ajuste es bueno cuando el stress es menor que el 5% (0,05). Para su interpretación se puede utilizar la siguiente escala orientativa:

Stress	**Ajuste**
0,1-0,2	Pobre
0,05-0,1	Regular
0,025-0,05	Bueno
<0,025	Excelente
0	Perfecto

También se emplea como medida de ajuste el coeficiente de correlación múltiple al cuadrado RSQ entre las distancias y las disparidades, que se puede interpretar como la proporción de varianza de los datos transformados -o porcentaje de variabilidad- recogida por la representación obtenida. Varía entre 0 y 1, y cuanto más próximo a 1 mejor es el ajuste; se suele considerar que 0,6 es un valor mínimo aceptable. Existen numerosas variantes y algoritmos para la realización del Escalamiento Multidimensional.

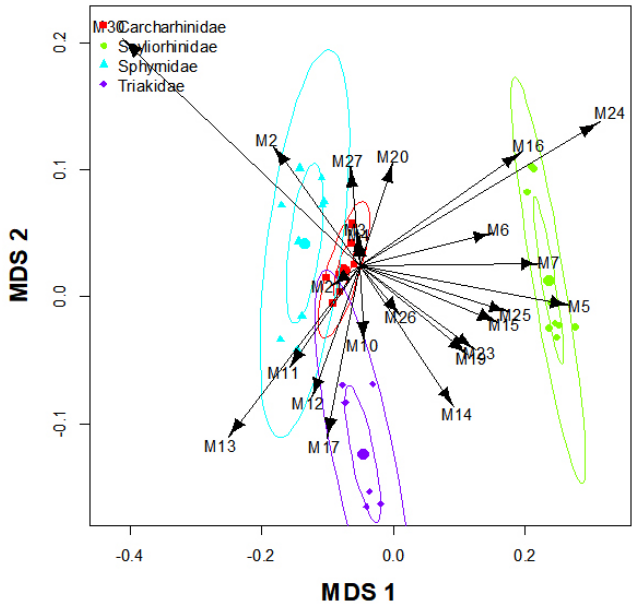
FUNCIONES

El Escalamiento Multidimensional se realizó con la función metaMDS del paquete vegan (Oksanen et al., 2013). Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2014). Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014). La envolvente convexa se estima con la función chull del paquete grDevices.

EJEMPLO

El estudio consistió en realizar diferentes medidas morfométricas de especies de tiburones, que pertenecen a distintas familias del orden Carcharhiniformes. El objetivo es determinar si las especies de la misma familia tienen morfometrías parecidas y, por tanto, el Escalamiento Multidimensional muestra que existe un patrón morfométrico asociado a cada una de las familias.

Figura XIII.8. Escalamiento multidimensional usando medidas morfométricas de diferentes familias de tiburones.



En la ventana de resultados se muestra el valor de la medida de ajuste, con un Stress de 0,009 (0,9%), que de acuerdo con la clasificación mostrada en la introducción se considera excelente. Se observa que existe un agrupamiento relativamente bueno entre familias (Figura XIII.8). Por tanto, se concluye que es posible diferenciar las distintas familias de tiburones en función de su morfometría. El gráfico *biplot* también permite ver cuáles son las variables morfométricas responsables de esa diferenciación entre familias. Las elipses muestran los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las familias. Es un indicador del grado de solapamiento entre las familias de tiburones. Estos niveles se pueden modificar entrando en la función scatterplot usando el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.

Es importante mencionar que el argumento *cat* solamente se utiliza para el gráfico *biplot* y no en el Escalamiento Multidimensional. Es decir, no se está dando ninguna información previa para realizar el análisis, de los potenciales grupos que hay en los datos, como se hace por ejemplo en las regresiones logísticas, Análisis Discriminante, CARTs, etc. Por tanto, en el caso de obtener un valor bajo de stress y un gráfico que claramente muestre una separación de grupos, como en este ejemplo, esto indicaría que existe una estructura subyacente -unos grupos claramente diferenciados- en base al conjunto de medidas utilizadas.

Si se accede a la función metaMDS con el argumento *METAMDS*, es posible cambiar el índice de disimilitud que se utiliza en el Escalamiento Multidimensional, con el argumento *distance*. Probar con distintos tipos de índice de disimilitud es aconsejable, ya que se puede obtener un valor de stress menor y un gráfico donde los grupos queden más claramente separados.

En caso de obtener este error *some dissimilarities are negative*, ello puede deberse a que hay filas o columnas en las que todos los valores son cero.

Value

Se obtiene un gráfico *biplot* con las dos primeras dimensiones del Escalamiento Multidimensional.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H. & Wagner, H. (2013) Community Ecology Package. R package version 2.0-10. Available at: http://CRAN.R-project.org/package=vegan.

Examples

```
## Not run:
data(ZXIII3)
XIII3(data=ZXIII3, var=c("M2", "M3", "M4", "M5", "M6", "M7",
"M10", "M11",
"M12", "M13", "M14", "M15", "M16", "M17", "M19", "M20", "M21",
"M23", "M24",
"M25", "M26", "M27", "M30"), cat="Familia", ellipse=TRUE)

## End(Not run)
```

XIII4 *CORRELACION CANONICA*

Description

Se aplica una Correlación Canónica.

Usage

```
XIII4(data, varX,  varY, labels=NULL, labelsvar="varX", log=FALSE, dim=c(1,2),
expand=TRUE, ellipse=FALSE, convex=FALSE, ResetPAR=TRUE, PAR=NULL, CANCOR=NULL,
SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL,
TEXTX=NULL, TEXTY=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, arrows=TRUE, larrow=0.95,
colAX="black", colAY="blue", file1="Output.txt", file2="xcoef.csv", file3="ycoef.csv",
file4="xscores.csv", file5="yscores.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
varX	Variables independientes.
varY	Variables dependientes.
labels	Variable que permite etiquetar los casos.

`labelsvar`	Los casos se representan usando las puntuaciones (*scores*) de la variables independientes (X) o dependientes (Y).
`log`	Si es TRUE se aplica el logaritmo a las variables dependientes e independientes.
`dim`	Dimensiones que se representan en el gráfico.
`expand`	Si es TRUE los coeficientes de las variables X e Y y las puntuaciones (*scores*) de los casos de la variables independientes (X) o dependientes (Y) se ajustan a una misma escala, cuando se representan en el gráfico. Esto permite ver mejor el ajuste de los coeficientes y puntuaciones en el espacio canónico, es decir, ver que variables y puntuaciones coinciden en el espacio generado por las variables canónicas, las cuales se han creado como combinación lineal de las variables dependientes e independientes.
`ellipse`	Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *labels*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.
`convex`	Si es TRUE, la envolvente convexa se calcula para cada categoría.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`CANCOR`	Se accede a la función cancor del paquete candisc.
`SCATTERPLOT`	Se accede a la función scatterplot del paquete car.
`COLOR`	Permite modificar los colores del gráfico, pero deben ser tantos como grupos diferentes tenga la variable *labels*.
`PCH`	Vector con los símbolos del gráfico, que deben ser tantos como grupos diferentes tenga la variable *labels*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
`XLIM, YLIM`	Vectores con los límites de los ejes X e Y.
`XLAB, YLAB`	Leyendas de los ejes X e Y.
`TEXTX`	Se accede a la función que permite modificar las etiquetas de los casos en el gráfico *biplot*.
`TEXTY`	Se accede a la función que permite modificar las etiquetas de los coeficientes de las variables dependientes (Y) en el gráfico *biplot*.
`LEGEND`	Permite incorporar una leyenda.
`AXIS`	Permite añadir ejes.
`MTEXT`	Permite añadir textos en los márgenes del gráfico.
`arrows`	Si es TRUE se muestran las flechas en el scatterplot.
`larrow`	Permite modificar la longitud de las flechas.
`colAX, colAY`	Colores de las flechas de los ejes X e X.
`file1`	ARCHIVO TXT. Nombre del archivo de salida con resultados del análisis.

file2	ARCHIVO CSV. Nombre del archivo de salida con los coeficientes de las variables *X*, que se muestran en el gráfico, pero sin la corrección aplicada en el caso de que el argumento *expand=TRUE*.
file3	ARCHIVO CSV. Nombre del archivo de salida con los coeficientes de las variables *Y*, que se muestran en el gráfico, pero sin la corrección aplicada en el caso de que el argumento *expand=TRUE*.
file4	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de los casos considerando las variables *X*, que se muestran en el gráfico, pero sin la corrección aplicada en el caso de que el argumento *expand=TRUE*.
file5	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de los casos considerando las variables *Y*, que se muestran en el gráfico, pero sin la corrección aplicada en el caso de que el argumento *expand=TRUE*.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XIII. ANÁLISIS MULTIVARIANTE FACTORIAL

XIII.4. CORRELACIÓN CANÓNICA

El análisis de correlación canónica es una generalización de las regresiones. Una regresión simple correlaciona una variable independiente con otra dependiente. Una regresión múltiple hace lo mismo con más de una variable independiente y una dependiente. La correlación canónica correlaciona un grupo de variables $x_1, ..., x_m$ con otro grupo $y_1, ..., y_n$. El método consiste en encontrar dos variables $A = a_1x_1 + ... + a_mx_m$, $B = b_1y_1 + ... + b_ny_n$ entre las cuales la correlación sea máxima. Estas variables compuestas se llamarán variables canónicas. El análisis calculará varias correlaciones canónicas, cada una de ellas con un par de variables canónicas *A* y *B*, en concreto, calculará tantas variables canónicas como el número de variables que tenga el grupo más pequeño (el de las *x* o el de las *y*), aunque sólo las primeras tienen interés práctico. A menudo unos pocos pares de variables canónicas permitirán analizar las relaciones entre ambos grupos, facilitando el estudio de nuestros datos al reducir la dimensión.

Este tipo de análisis multivariante comparte aspectos con el análisis de componentes principales y el análisis factorial. Pero a diferencia de estos, que buscan relaciones internas entre las variables de un grupo, la correlación canónica busca una relación entre dos grupos de variables.

Los requerimientos generales de este tipo de análisis son los mismos que para un análisis factorial:

1. Se excluyen variables cualitativas.

2. Las relaciones entre variables deben ser lineales.

No obstante, conviene apuntar algunas características que, aunque no son hipótesis previas, es conveniente que se cumplan particularmente en este análisis:

1. El método funciona mejor cuando la correlación entre variables dentro de cada grupo es pequeña, y es grande entre variables de distinto grupo.

2. Las variables, dentro de un grupo, aunque distintas, deben ser homogéneas respecto al tipo de información, o el tema, al que se refieren (recuérdese que en el análisis factorial también existe este requisito, pero no hay dos grupos); de lo contrario, si la información que contienen no está relacionada, pierde significado la construcción de variables nuevas (canónicas) como combinación lineal de las variables antiguas.

En síntesis, este tipo de análisis será especialmente útil cuando tenemos dos grupos de variables, cada uno de los cuales nos daría información sobre un tema, medido por diferentes variables. Entonces, nosotros pretendemos buscar una correlación entre esos dos grupos, o temas de información. Este método puede ser más adecuado que comparar las variables dos a dos, y además, la correlación encontrada entre las variables canónicas calculadas siempre será superior a la existente en cualquiera de esas comparaciones dos a dos.

FUNCIONES

La Correlación Canónica se realizó con la función cancor del paquete candisc (Friendly, 2007; Firendly y Fox, 2013). Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2014). Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014). La envolvente convexa se estima con la función chull del paquete grDevices.

EJEMPLO

Vamos a utilizar este análisis para comprobar si una serie de variables, que globalmente se refieren a la calidad de un suelo para el crecimiento de vegetales, nos predicen la mejor o peor condición en la que pueden estar los vegetales que han crecido sobre ese suelo. Tenemos un grupo de variables que nos da información sobre calidad del suelo que son: contenido en agua (en % del peso del suelo), nitrógeno, fosfato y una serie de oligoelementos minerales (en micro moles por gramo de suelo). El otro grupo de variables nos da una idea sobre condición de las plantas, y son una serie de parámetros independientes relacionados con el crecimiento y la reproducción: cobertura foliar (en m^2), la longitud promedio de entrenudos (en cm) y el % de semillas viables del total producidas. En nuestro caso tenemos datos de un conjunto de árboles de la misma edad que han crecido en diferentes suelos. Queremos ver si ambos grupos de variables se correlacionan.

Es muy importante tener en cuenta que se debe valorar la pertinencia incluir variables que tengan poca variabilidad, es decir, que todos los valores sean iguales salvo unos pocos casos. Estas variables con poca variabilidad pueden tener un gran peso en los ejes canónicos y se debe valorar si su inclusión es necesaria. Por ejemplo, en el caso de trabajar con abundancia de especies, puede ocurrir que aquellas que no están presente en la mayoría de las muestras y aparecen una o dos veces tengan un peso importante en los ejes canónicos y es necesario valorar si es correcto incluirlas.

En el script se seleccionan el grupo de variables independientes con el argumento *varX* y dependientes con *varY*. Con el argumento *labels="Area"* se especifica que se muestren las puntuaciones de los casos diferenciando por las áreas de muestreo.

La siguiente tabla muestra los resultados del análisis. Se calculan tres pares de variables canónicas, porque el grupo de las variables dependientes tiene el menor número de variables, que son tres. El primer par de variables canónicas explica la mayor parte de la varianza (94,72%), y los restantes pares explican, sucesivamente, lo máximo posible de la varianza restante. Al explicar diferentes aspectos de la varianza, los pares de variables canónicas nunca están correlacionados entre sí.

```
[1] "CANONICAL CORRELATIONS"

[[2]]

Canonical correlation analysis of:
      5   X  variables:  H2O, N, P, Fe, Mn
  with     3   Y  variables:  Coverage, Internodes, Seeds

    CanR CanRSQ  Eigen percent    cum
1 0.9403 0.8841 7.6265  94.719  94.72 ******************
2 0.4810 0.2313 0.3009   3.738  98.46 *
3 0.3325 0.1106 0.1243   1.544 100.00

Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR  WilksL      F df1    df2 p.value
1 0.94025 0.07925 8.3501  15 83.218 0.00000
2 0.48096 0.68369 1.6229   8 62.000 0.13659
3 0.33251 0.88944 1.3259   3 32.000 0.28312
```

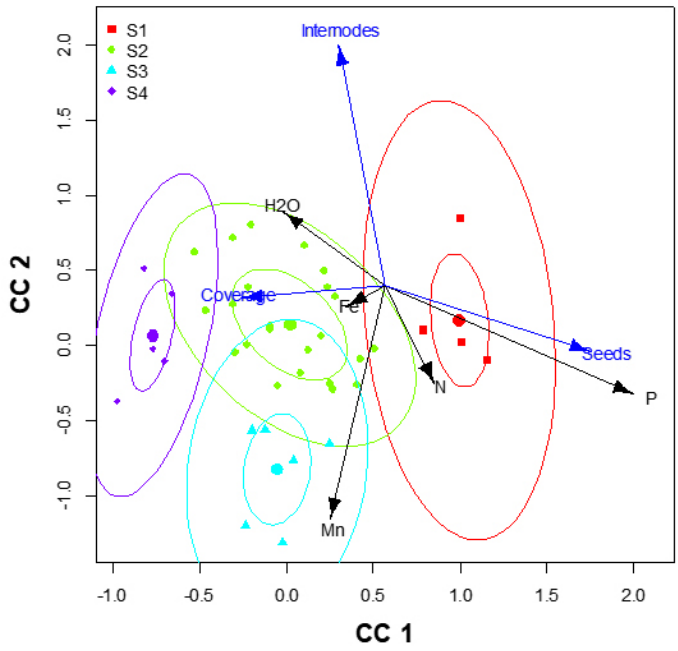
El primer R canónico (*CanR 1= 0,94*) es equivalente al R global del análisis. El segundo R canónico (*CanR 2 = 0,48*) es independiente del primero, el tercer R canónico (*CanR 3 = 0,33*) es independiente del primero y del segundo, y así sucesivamente en el caso de existir más pares canónicos. Vemos que el valor de R baja, y de hecho, con eliminar el primer par de variables canónicas calculado, R deja de ser significativo (p = 0,136). Por lo tanto, la correlación sólo es significativa en el primer par de variables canónicas calculado, es decir, para el primer eje canónico (p < 0,001). Globalmente se puede decir que la calidad del suelo está relacionada significativamente con la condición de las plantas con las variables utilizadas.

En la Figura III.9 se muestran los coeficientes de las variables dependientes e independientes, así como las puntuaciones de los casos diferenciadas en función del área de muestreo. Cuanto mayor sea el valor absoluto del coeficiente, mayor es la contribución de esa variable a la variable canónica. En este ejemplo interesa especialmente cuáles son las variables de más peso en la primera correlación canónica (*CC1*), porque como ya vimos, es la única significativa. Vemos que las variables independientes de más peso son el fosfato y el nitrógeno. En lo que se refiere a las variables dependientes observaremos, igualmente para la primera correlación canónica, que la variable de más peso es el porcentaje de semillas viables, que es mayor en suelos con más concentración de fosfato y nitrógeno.

Las elipses muestran los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada uno de los sitios de muestreo. Es un indicador del grado de solapamiento entre los sitios de muestreo. Estos niveles se pueden modificar entrando en la función scatterplot usando el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*. Con algunos ejemplos se ha observado que no es posible estimar las elipses y se obtiene este error *Error in svd(sqrt(w/sum(w)) * X, nu = 0):* o este otro *Error in cov.trob(cbind(x[use], y[use]), wt = weights[use]): no positive weights*. En caso de obtener esos errores o simplemente de no querer mostrar las elipses, de nuevo usando el argumento *SCATTERPLOT* hay que seleccionar *ellipse=FALSE*. La figura muestra que se pueden diferenciar bien las áreas de muestreo en función de las características del suelo y la condición de las plantas. El sitio 1 es donde hay mayores concentraciones de fosfato y nitrógeno en el suelo y, debido a ello, mayor porcentaje de semillas viables. El sitio 4, por el contrario, es donde el porcentaje de semillas viables es menor, en principio provocado por un suelo de una menor calidad, es decir, concentraciones de fosfato y nitrógeno más bajas.

Figura XIII.9. Ejes canónicos I y II de la Correlación Canónica aplicada

a los datos de crecimiento y reproducción de plantas y de variables que miden la calidad del suelo.



Es posible obtener un mayor porcentaje de explicación en los primeros ejes, e incluso que sean más significativos, si no se incluyen en el análisis aquellas variables que contribuyen poco a los ejes canónicos, es decir, que quedan en el centro del gráfico. Por tanto, se aconseja realizar diferentes pruebas eliminando variables que tienen poco peso en el análisis y observar si con ello se aumenta el porcentaje de explicación y/o grado de significación de los ejes canónicos.

Value

Se obtiene un gráfico *biplot* con la Correlación Canónica, un archivo TXT con los resultados del análisis y cuatro archivos CSV: los coeficientes de las variables *X* e *Y* y las coordenadas de los casos considerando las variables *X* e *Y*.

References

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2013) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.6-5. Available at: http://CRAN.R-project.org/package=candisc.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Examples

```
## Not run:
data(ZXIII4)
XIII4(data=ZXIII4, varX=c("H2O","N","P","Fe","Mn"), varY=c("Coverage",
"Internodes","Seeds"), labels="Area", ellipse=TRUE)

## End(Not run)
```

XIV1 *ARBOLES DE CLASIFICACION Y REGRESION (CARTs)*

Description

Se realiza un Árbol de Clasificación y Regresión (CART).

Usage

```
XIV1(data, cat, var, ResetPAR=TRUE, PAR=NULL, CART=NULL, RPART=NULL,
file="Output.txt")
```

Arguments

data	Archivo de datos.
cat	Variable dependiente.
var	Variables independientes.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
CART	Se accede a la función rpart.
RPART	Se accede a la función rpart.plot.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados.

Details

XIV. ANÁLISIS MULTIVARIANTE DE CLASIFICACIÓN

Es el conjunto de métodos de análisis multivariante cuyo objetivo principal es agrupar o clasificar a los elementos de una muestra de objetos descritos mediante varias variables. Frente a los métodos factoriales, las técnicas de clasificación tienen en general un planteamiento algorítmico (no algebraico), y la mayoría de ellas no están basadas en el modelo lineal. Históricamente surgen con el computador y la informática, y se benefician de la extraordinaria potencia de cálculo que proporcionan estas herramientas. Existen centenares de métodos de clasificación o análisis clúster, de los cuales veremos solamente los más importantes.

En general se trata de obtener grupos (clases, clústeres) homogéneos, es decir construidos de tal forma que los elementos de cada grupo sean similares entre sí; las clases deben ser diferentes

o separadas, de modo que sean muy distintos los elementos de grupos diferentes. Se busca una estructura en los datos que ayude a comprender mejor la realidad de la población investigada.

Muchos métodos de clasificación permiten utilizar variables de tipo diverso, cuantitativas o cualitativas, ordinales o nominales, y en general no se necesitan las hipótesis restrictivas, como Normalidad y homogeneidad de varianzas, habituales en otras técnicas estadísticas. Tampoco es necesario en general un conocimiento previo de los datos.

Por estas razones los métodos de clasificación se suelen emplear con fines exploratorios, para establecer una primera aproximación a los datos, y para generar ideas, plantear hipótesis que puedan ser contrastadas con otras técnicas, o para iniciar la construcción de modelos explicativos.

XIV.1 ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN (CARTs)

Los árboles de clasificación y regresión son un procedimiento no paramétrico de predicción de una variable dependiente sobre la base de un conjunto de variables independientes (Breiman et al., 1984). Su carácter no paramétrico hace que no requiera ninguna hipótesis relativa a la distribución de las variables dependientes e independientes, ni a la relación entre ellas y sus posibles interacciones. La variable dependiente puede ser categórica (Árboles de Clasificación), como por ejemplo diferentes especies, rechazo a un tratamiento, presencia/ausencia de un cáncer, etc. La variable dependiente también puede ser continua (Árboles de Regresión), como por ejemplo el índice de masa corporal, el azúcar en sangre, etc. En ambos casos, Árboles de Clasificación y de Regresión (CARTs), el objetivo es identificar las variables que mejor permiten identificar la variable dependiente. Por ejemplo, determinar cuales son las características clínicas, genéticas y/o ambientales que permiten predecir si una persona puede tener un determinado cáncer, o ser propenso a un infarto, etc. Amigo et al. (2007) aplicaron CARTs al índice de masa corporal (IMC) de niños y determinaron que la madre, el consumo de lípidos y las horas de televisión el domingo, eran las variables que mejor clasificaban a los niños en función de su IMC. Un análisis CARTs consiste generalmente en tres pasos (Guisande & Vaamonde, 2012):

1. **Construcción del árbol**. En el árbol, el nodo raíz representa a toda la población. Este nodo raíz se divide en dos subgrupos en base a la partición de una variable independiente o predictora. Los nodos hijos se dividen por medio de la partición de una variable, que puede ser la misma de antes o una nueva. El proceso se repite sucesivamente hasta que se cumpla alguna condición de parada. Las divisiones se seleccionan de modo que la "impureza" o heterogeneidad de los nodos hijos sea menor que la del nodo padre. La función de impureza o criterio de partición es una medida que permite determinar la calidad del nodo hijo, y decidir como se realiza la partición de un nodo en sus dos nodos hijos. Existen diferentes tipos de índices para cuantificar la impureza de las particiones (Breiman et al., 1984).

2. **Poda del árbol**. El árbol que se ha construido generalmente está sobreajustado, es decir, contiene gran cantidad de niveles y no necesariamente una mayor complejidad significa una mejor clasificación. Imaginemos que estamos trabajando con diferentes especies y se quiere determinar cuales son las variables que permiten diferenciarlas mejor. El árbol puede seguir dividiéndose en nodos hasta que el último individuo sea clasificado, utilizando en cada paso un número mayor de variables y un modelo más complejo hasta alcanzar un porcentaje cero de individuos mal clasificados. Al aumentar la complejidad de un modelo siempre se obtiene un mejor ajuste, pero a veces simplemente recogemos peculiaridades muy específicas de los datos, que no lo hacen útil para describir el comportamiento de nuevos datos (datos de validación). La poda consiste en, una vez construido el árbol, eliminar hacia atrás las particiones que no representen un aumento considerable de la capacidad de predicción total.

3. **Validación del árbol**. Conforme se va podando se reduce la complejidad del árbol pero el número de individuos mal clasificados puede aumentar. El mejor árbol será aquel que tenga la pro-

porción óptima entre la tasa de mala clasificación (cociente entre las observaciones mal clasificadas y el número total de observaciones) y la complejidad del árbol. La selección del mejor árbol se realiza por validación. El procedimiento consiste en separar una parte de la muestra, que no es utilizada para la construcción del árbol, sino únicamente para la predicción, midiendose en ella el porcentaje de acierto. También se suele utilizar un método de validación cruzada, que consiste en dividir la muestra en varios grupos excluyentes (normalmente 10 grupos) de aproximadamente igual tamaño, y hacer el árbol dejando cada vez un grupo fuera, que es utilizado para la predicción y medida del grado de acierto. El proceso se repite hasta usar todos los grupos y se selecciona finalmente el árbol con la menor tasa agregada de mala clasificación (media de los diez grupos).

FUNCIONES

Para la estimación del CART se usa la función rpart del paquete rpart (Therneau et al., 2014) y para su representación la función rpart.plot del paquete rpart.plot (Milborrow, 2014).

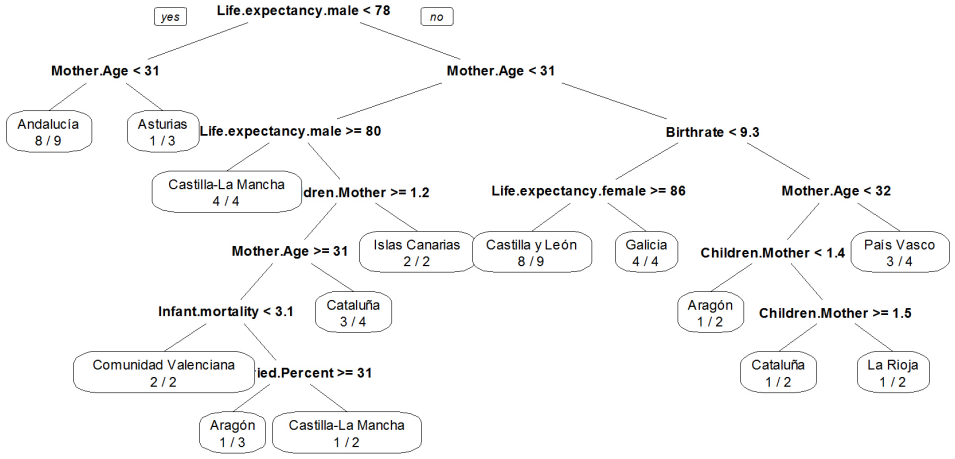
EJEMPLO

El ejemplo utiliza datos demográficos del año 2010 de las 19 regiones o comunidades autónomas de España publicados por el Instituto Nacional de Estadística (`http://www.ine.es`). El objetivo es determinar si es posible clasificar las diferentes regiones en función de las siguientes variables demográficas: número de hijos por madre, edad media de la madre cuando tiene su primer hijo, porcentaje medio de mujeres no casadas con hijos, la esperanza de vida de hombres y mujeres, y el número de nacimientos, mortalidad de la población y mortalidad infantil (menores de 1 año), los tres últimos parámetros demográficos por cada 1000 habitantes.

La variable dependiente son las regiones y las variables independientes o predictoras son todos los indicadores demográficos. Con *method* se define el método de partición; si la variable dependiente es categórica, como en nuestro ejemplo, pondremos *method='class'*. Se puede omitir el método, que es lo que se ha hecho en el ejemplo, en cuyo caso la propia función decide el método más adecuado en función del tipo de variable dependiente. Con el argumento *control*, y mediante la función *rpart.control()*, se pueden modificar aspectos importantes del árbol. Con el argumento *minsplit* se especifica el número mínimo de observaciones que deben existir en un nodo para que se haga la partición, es decir, conforme sea más pequeño el árbol tendrá más nodos. Con *minbucket* se define el número mínimo de observaciones en un nodo terminal y, por tanto, al igual que el argumento anterior, conforme es más pequeño se profundiza más en el árbol, es decir, tiene más nodos y se hace más complejo. Con *cp* se controla la complejidad y permite ahorrar tiempo en el proceso de poda, que es la fracción (por defecto 0,01) en la que debe mejorar el indicador de ajuste para que continúe la construcción del árbol. Si la mejora es menor que cp el proceso se detiene. El usuario decide si quiere profundizar mucho y hacer el árbol muy complejo (cp pequeño), o tener un árbol más simple (cp más grande). Con *surrogatestyle* se controla el criterio para elegir la mejor variable suplente para la partición en cada nodo. Si es 0 se utiliza la variable con el mayor número de elementos clasificados correctamente, y si es 1 la que tiene el mayor porcentaje de elementos clasificados correctamente sobre el número de casos válidos de esa variable; la primera opción penaliza a las variables que tienen valores perdidos.

En la función rpart.plot con el argumento *type* se selecciona el formato del gráfico, es decir, el modo en que sale la información de las variables seleccionadas en cada partición y las categorías predominantes en cada nodo. Con el argumento *extra* se puede mostrar información adicional en los nodos, como en este ejemplo, en el que le pedimos que muestre en cada nodo el número de casos clasificados correctamente frente al total. Con *varlen=0* se define que los textos de los nodos sean completos y no se corten. Con *ycompress=TRUE* se especifica que cuando se solapan las etiquetas se desplacen verticalmente para evitar este solapamiento.

Figura XIV1.1. Parámetros demográficos que mejor discriminan las regiones de España.



El CART se muestra en la Figura XIV.1. Los números indican las provincias que han sido clasificadas correctamente y el número total de provincias de cada región. Por ejemplo, en el grupo que pone Castilla y León, hay 8 provincias de esta región de un total de 9 que hay en ese grupo. Hay un grupo que muestra 8 provincias de Andalucía de un total de 9. En general se observa que muchas regiones se identifican bien. Por ejemplo, en Islas Canarias, Galicia, Castilla-La Mancha y la Comunidad Valenciana, todas las provincias de la región están dentro de un mismo grupo. En otras como Andalucía, Cataluña, País Vasco, etc., casi todas las provincias están en el mismo grupo. Esto quiere decir que hay diferencias demográficas entre regiones en España, ya que es posible identificar relativamente bien cada una de las regiones en base a estos indicadores demográficos.

La primera variable es la esperanza de vida de los hombres, que separa un grupo de regiones con hombres que tienen una esperanza de vida menor de 78 años a la izquierda y mayor de 78 a la derecha. El número medio de hijos por madre y la edad de la madre cuando tiene el primer hijo también son variables importantes ya que aparecen en muchos nodos. El número de nacidos por cada 1000 defunciones también es una variable importante indicando que las regiones crecen a ritmos diferentes. En el árbol se ha optado por no identificar de forma correcta todos los casos, es decir, hay grupos terminales donde hay casos de diferentes categorías, para no hacer el árbol muy complejo. Lo importante en este ejemplo era ver si los indicadores demográficos diferencian a las regiones, y cuales de estos indicadores son los más importantes para ello.

Value

Se obtiene un gráfico con el Árbol de Clasificación y Regresión y un archivo TXT con los resultados del CART.

References

Amigo, H., Bustos, P., Erazo, M., Cumsille, P. & Silva, C. (2007) Factores determinantes del exceso de peso en escolares. Un estudio multinivel. *Revista Médica de Chile*, 57: 353-357.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: http://CRAN.R-project.org/package=rpart.plot.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: http://CRAN.R-project.org/package=rpart.

Examples

```
## Not run:
data(ZXIV1)
XIV1(data=ZXIV1, cat="Region", var= c("Children.Mother", "Mother.Age",
"Unmarried.Percent", "Life.expectancy.male", "Life.expectancy.female",
"Birthrate","Mortality", "Infant.mortality"))

## End(Not run)
```

XIV2 *ANALISIS DISCRIMINANTE*

Description

Se aplican dos tipos de Análisis Discriminante: Lineal y Cuadrático.

Usage

```
XIV2(data, cat, var, ellipse=TRUE, convex=FALSE, quadratic=FALSE, expand=TRUE,
dimS=c(1,2), ResetPAR=TRUE, PAR=NULL, CANDISC1=NULL, CANDISC2=NULL, CANPLOT=NULL,
SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL,TEXT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
arrows=TRUE, larrow=0.95, colArrows="black", file1="Var loadings-Linear.csv",
file2="Cat loadings-Linear.csv", file3="Table cross-validation-Linear.csv",
file4="Cases cross-validation-Linear.csv",
file5="Table cross-validation-Quadratic.csv",
file6="Cases cross-validation-Quadratic.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
cat	Categorías de los casos.
var	Variables que se incluyen en el análisis.
ellipse	Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *cat*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.
convex	Si es TRUE, la envolvente convexa se calcula para cada categoría.
quadratic	Si es TRUE se realiza un Análisis Discriminante Cuadrático, además del Análisis Discriminante Lineal.

expand	Si es TRUE las coordenadas de las categorías, en el gráfico *biplot*, se ajustan a la escala de las coordenadas de los casos.
dimS	Dimensiones que se representan en el gráfico *biplot*.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
CANDISC1	Se accede a la función candisc del paquete candisc, que realiza el discriminante de una dimensión.
CANDISC2	Se accede a la función candisc del paquete candisc, que realiza el discriminante con dos dimensiones.
CANPLOT	Se accede a la función plot.cancor del paquete candisc, que realiza el gráfico con el discriminante de una dimensión.
SCATTERPLOT	Se accede a la función scatterplot del paquete car, que realiza el gráfico *biplot* con el discriminante de dos dimensiones.
COLOR	Permite modificar los colores del gráfico *biplot*, pero deben ser tantos como grupos diferentes tenga la variable *cat*.
PCH	Vector con los símbolos del gráfico *biplot*, que deben ser tantos como grupos diferentes tenga la variable *cat*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
TEXT	Se accede a la función que permite modificar las etiquetas de los casos en el gráfico *biplot*.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
arrows	Si es TRUE se muestran las flechas en el scatterplot.
larrow	Permite modificar la longitud de las flechas.
colArrows	Color de las flechas.
file1	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las variables en el gráfico del Análisis Discriminante Lineal.
file2	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las categorías en el gráfico del Análisis Discriminante Lineal.
file3	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Lineal.
file4	ARCHIVO CSV. Nombre del archivo de salida con el grupo a que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Lineal.
file5	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Cuadrático.
file6	ARCHIVO CSV. Nombre del archivo de salida con el grupo a que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Cuadrático.

na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

XIV. ANÁLISIS MULTIVARIANTE DE CLASIFICACIÓN

XIV.2. ANÁLISIS DISCRIMINANTE

Una técnica estadística multivariante ampliamente utilizada para caracterizar clases de elementos cuando se dispone de un cierto número de variables es el Análisis Discriminante, que permite -a partir de una clasificación inicial- obtener una nueva clasificación en función de los valores observados de un conjunto de variables.

El método consiste en la determinación de unas funciones de las variables originales, llamadas funciones discriminantes, que permiten decidir en qué clase debe estar cada elemento, utilizando como criterio de asignación la proximidad (o similitud) de cada elemento a las distintas clases o grupos existentes.

Puede utilizarse, por ejemplo, para identificar el origen geográfico de una muestra de miel desconocida, cuando disponemos de varias muestras de cada zona, utilizando los contenidos o concentraciones de distintos tipos de azúcares que han sido determinados analíticamente en todas las muestras.

La construcción de las funciones discriminantes, combinaciones lineales de las variables originales, se realiza buscando aquellas que hacen máxima la separación entre grupos, lo que equivale a minimizar la variabilidad dentro de las clases.

Las funciones obtenidas permiten discriminar de la mejor forma posible entre los elementos pertenecientes a las distintas clases o grupos. La asignación de los elementos a clases se realiza mediante el criterio de Bayes: cada elemento se asigna a la clase para la que es mayor la probabilidad de pertenencia condicionada por los valores que toman las funciones discriminantes.

XIV.2.1. Hipótesis

El Análisis Discriminante Lineal requiere, para que pueda ser utilizado correctamente, la hipótesis de homogeneidad de varianzas, es decir, que las matrices de covarianzas de cada clase sean iguales, o que la dispersión sea similar en cada clase para todas las variables. Se emplea la prueba M de Box para contrastar esta hipótesis de homogeneidad. También es necesario que las funciones discriminantes tengan distribución Normal. Ambas hipótesis de normalidad de residuos y homogeneidad de varianzas se utilizan para el cálculo de las probabilidades de asignación de los elementos a las distintas clases. Sin embargo, no suelen comprobarse en la práctica ya que, las pruebas M de Box y contrastes de normalidad son muy sensibles a pequeñas desviaciones y con datos reales prácticamente siempre nos dirán que no se cumplen. Si las comprobamos nunca podremos aplicar Análisis Discriminante Lineal.

El Análisis Discriminante Lineal es razonablemente robusto frente a un incumplimiento parcial de estas hipótesis (basta con que se cumplan de modo aproximado), y ese incumplimiento no es crítico si el porcentaje de acierto en la clasificación (especialmente en la validación cruzada) es elevado. Es necesario que el porcentaje de acierto sea alto no solamente en el conjunto de los elementos, sino también en cada uno de los grupos. En la práctica lo que se comprueba es, por lo tanto,

ese porcentaje de acierto en la validación cruzada, y no es necesario ni conveniente verificar las hipótesis, salvo que tengamos una clara evidencia de que se incumplen de forma manifiesta (por el tipo de datos, por ejemplo cuando las variables explicativas son binarias), en cuyo caso puede no ser correcto utilizar el Análisis Discriminante Lineal. Si el porcentaje de acierto es bajo, ello podría deberse al incumplimiento de las hipótesis, pero tampoco en ese caso es necesario comprobarlo, ya que lo prudente es renunciar al Análisis Discriminante cuando clasifica mal.

Si el argumento *quadratic=TRUE* se realiza un Análisis Discriminante Cuadrático, el cual no tiene como requisito la homogeneidad de varianzas (Venables & Ripley, 2002), además del método Lineal. La comparación del porcentaje de aciertos, por validación cruzada, obtenido con el Análisis Discriminante Cuadrático y el Lineal podría ser utilizado como un indicador de lo robusto que es el método Lineal, ya que si ambos porcentajes son parecidos, esto indicaría que el Análisis Discriminante Lineal es correcto. Sin embargo, el Análisis Discriminante Cuadrático no se puede aplicar cuando alguna categoría tenga pocos casos, generalmente menos de 5. Además, es frecuente obtener este error *rank deficiency in group...* al aplicar el método Cuadrático, en cuyo caso tendríamos que especificar el argumento *quadratic=FALSE*, para que no se realice.

XIV.2.2. Método paso a paso

Se puede utilizar para el análisis discriminante el método paso a paso, con el cual se seleccionan sucesivamente de forma automática las variables con mayor capacidad de discriminación de acuerdo con un determinado criterio estadístico (por ejemplo el mínimo valor del coeficiente lambda de Wilks), las cuales se incorporan al conjunto de variables que intervienen en el análisis. En algunos pasos se puede descartar una variable que se introdujo con anterioridad, pero que al considerar otras variables nuevas deja de aportar al conjunto una capacidad de discriminación significativa. En general es preferible sin embargo seleccionar las variables que deben ser utilizadas no de forma automática, sino con criterios técnicos o científicos -no estadísticos- empleando aquellas que tienen a priori una relación directa (por ejemplo de tipo causa-efecto) con la clasificación.

XIV.2.3. Influencia de las distintas variables independientes

Generalmente se utiliza la solución obtenida para clasificar casos cuyo grupo es desconocido a priori, es decir con fines de identificación o predicción. También se utiliza para reclasificar los elementos o reasignar los elementos a las distintas clases cuando esa clasificación no es fiable o es solamente orientativa, y para estudiar el efecto de las diferentes variables explicativas sobre la clasificación. Con este último objetivo se utilizan los coeficientes estándar de las primeras funciones discriminantes: cuanto mayor es el coeficiente, mayor es el efecto de esa variable.

También se pueden emplear los llamados "coeficientes de estructura", o coeficientes de correlación entre las funciones discriminantes y las variables independientes (las cuales han sido utilizadas para su construcción). Generalmente los coeficientes de estructura mayores en valor absoluto se corresponden con los mayores coeficientes estándar, aunque en ocasiones el coeficiente de estructura alto corresponde a un coeficiente estándar bajo, lo que permite detectar variables redundantes, o al revés, un coeficiente de estructura bajo con un coeficiente estándar alto, lo que identifica una variable confusora, que interactúa con las restantes distorsionando los resultados. Generalmente ambos tipos de variables deben ser excluidos del análisis.

Se suelen utilizar ANOVAs univariados, para contrastar la hipótesis de que las medias de cada variable independiente son iguales para todos los grupos. Permiten medir también la importancia o capacidad de discriminación de las distintas variables. Si las medias son iguales la variable tendrá en general escasa o nula capacidad de discriminación.

XIV.2.4. Validación

Cuando el número de variables es grande en relación con el tamaño de la muestra, puede aparecer un problema de sobreajuste. Las funciones discriminantes permiten clasificar con un elevado grado de acierto los elementos de la muestra, pero si aplicamos esas mismas funciones a la clasificación de elementos nuevos, no utilizados para la construcción de las funciones discriminantes, el grado de acierto disminuye notablemente. En esos supuestos se están utilizando las peculiaridades de la muestra para asignar correctamente un número alto de casos, pero las funciones no son válidas -no son representativas- para la población de la que procede la muestra. Es necesario realizar algún tipo de validación para asegurar que el procedimiento será válido para la población y no solamente para la muestra. Una opción es reservar una parte de la muestra (por ejemplo 10 elementos), que no se utilizan para construir las funciones discriminantes, pero son asignados con ellas a las distintas clases. Si la proporción de elementos correctamente clasificados con esta muestra de validación es también elevada, el análisis discriminante será adecuado.

Otra posibilidad es la llamada clasificación cruzada dejando uno fuera. Se realiza un análisis discriminante excluyendo el primer elemento de la muestra, y a continuación este elemento es asignado mediante las funciones discriminantes obtenidas. Posteriormente hacemos lo mismo con cada uno de los elementos de la muestra. De este modo asignamos todos los elementos de la muestra sin que el elemento correspondiente haya sido utilizado en los cálculos de las funciones discriminantes y, por lo tanto, hemos utilizado todos los datos como muestra de validación. Si el porcentaje de acierto con esta validación cruzada es alto y similar, o solo ligeramente inferior, al porcentaje de acierto del análisis discriminante con toda la muestra, la validación es positiva, y el análisis discriminante adecuado para esos datos.

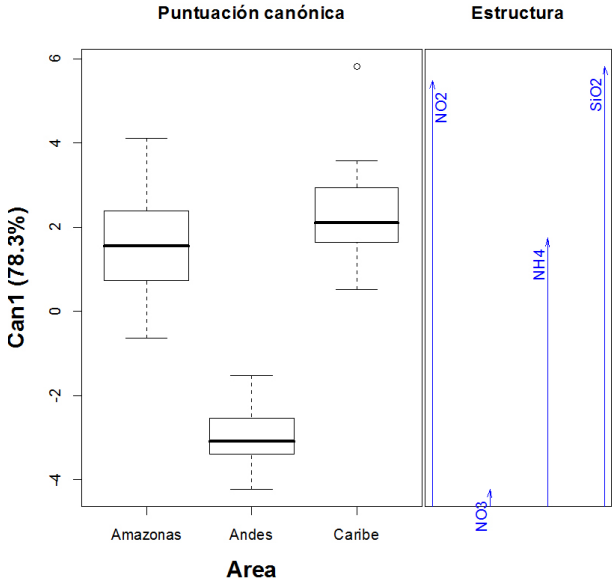
FUNCIONES

El Análisis Discriminante Lineal se realizó con las funciones candisc del paquete candisc (Friendly, 2007; Friendly & Fox, 2013) y lda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2014). El Análisis Discriminante Cuadrático se realizó con la función qda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2014). El gráfico con una dimensión se realizó con la función plot.cancor del paquete candisc (Friendly, 2007; Firendly & Fox, 2013). Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2014). Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014). La envolvente convexa se estima con la función chull del paquete grDevices.

EJEMPLO 1

Los datos corresponden a un estudio realizado en lagos de Colombia donde se analizó la abundancia de diferentes grupos del fitoplancton y también la concentración de algunos nutrientes. El objetivo es determinar si es posible discriminar entre regiones en base a la concentración de nutrientes. En el script se ha definido el argumento *quadratic=TRUE*, para que se realice el Análisis Discriminante Cuadrático.

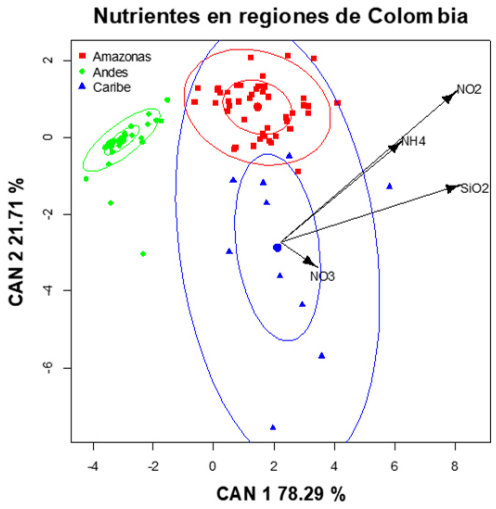
Figura XIV.2. Eje I del Análisis Discriminante aplicado a las regiones en función de la concentración de nutrientes.



Con el método Lineal, el 97,67% de los casos son asignados correctamente. Sin embargo, la validación cruzada muestra que el porcentaje de casos asignados correctamente es menor (94,19%), aunque todavía alto. El método Cuadrático muestra un porcentaje por validación cruzada del 95,3%, el cual es muy similar al obtenido con el método Lineal. Por tanto, las regiones se diferencian bien en función de la concentración de nutrientes. Los lagos de los Andes se diferencian por su baja concentración de nutrientes (Figura XIV.2) y el primer eje discriminante explica el 78,3% de la variabilidad.

El segundo eje, que absorbe un 21,7% de la varianza, discrimina entre los lagos del Amazonas y del Caribe (Figura XIV.3). Las elipses muestran los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las regiones. Es un indicador del grado de solapamiento entre regiones. Estos niveles se pueden modificar entrando en la función scatterplot usando el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*.

Figura XIV.3. Ejes I y II del Análisis Discriminante aplicado a las regiones en función de la concentración de nutrientes.



EJEMPLO 2

En el segundo ejemplo se usan los mismo datos de antes pero la diferenciación se realiza a nivel de lago en vez de región con el argumento *cat="Lago"* y que se añaden como variables las abundancias del fitolancton. En este caso se identifican correctamente el 94,19%, y con validación cruzada es el 90,7% con el método Lineal. Lógicamente, es más difícil diferenciar los lagos que las regiones. Aunque se ha definido el argumento *quadratic=TRUE*, de forma automática la función no aplica el método Cuadrático, ya que hay categorías con pocos casos.

El primer eje separa bien el lago de Tota por sus altas concentraciones de cianofíceas y clorofíceas, así como bajas concentraciones de nutrientes (Figura XIV.4). El segundo eje, que absorbe un 16,99% de la varianza, viene determinado principalmente por las diferencias en los silicatos (S_iO_2) y la abundancia de crisofíceas y dinofíceas (Figura XIV.5).

Figura XIV.4. Eje I del Análisis Discriminante aplicado a especies de fitoplancton discriminando los lagos en función de la composición del fitoplancton y concentración de nutrientes.

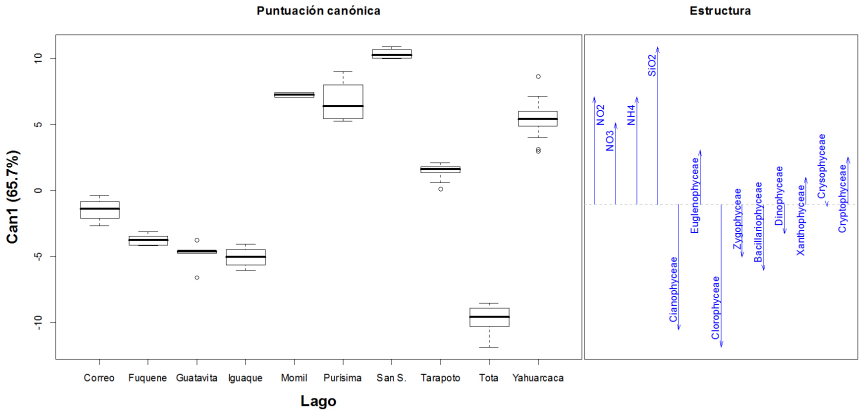
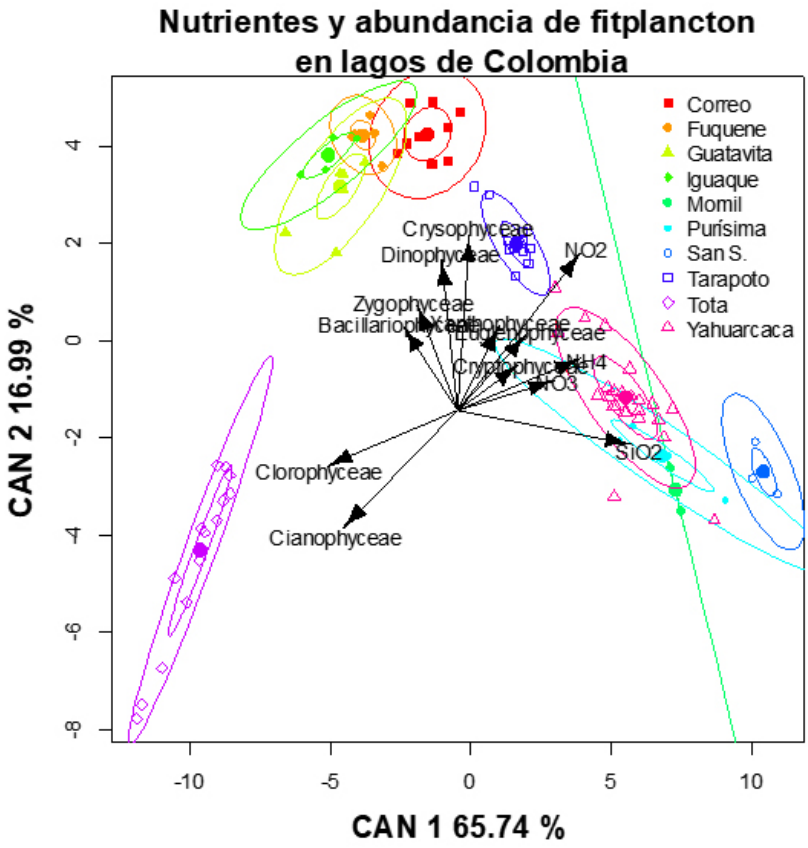


Figura XIV.5. Ejes I y II del Análisis Discriminante aplicado a especies de fitoplancton discriminando los lagos en función de la composición del fitoplancton y concentración de nutrientes.



EJEMPLO 3

Datos obtenidos en un citómetro de flujo del tamaño, rugosidad y emisión de radiación a diferentes longitudes de onda, de células que pertenecen a diferentes especies del fitoplancton: *Rhodomonas baltica*, *Alexandrium tamarense*, y dos cepas de *Alexandrium minutum*.

El objetivo es determinar si con las características de las células obtenidas en el citómetro de flujo es posible discriminar entre las diferentes especies, y las cepas dentro de la misma especie. Con los arumentos *ellipse=FALSE* y *convex=TRUE* se hace que se calculen las envolventes convexas en vez de las elipses.

Los resultados muestran que el 100% de los casos son asignados correctamente. La validación cruzada con los métodos Lineal y Cuadrático muestran un valor identico, siendo alto el grado de casos asignados correctamente (99,29%), de tal forma que solo un caso de *A. minutum-1* se identifica como *A. tamarense*.

El primer eje (Figura XIV.6) explica un 78,7% de la varianza y diferencia claramente la especie *A. tamarense* de las otras especies, por la radiación emitida entre 650 y 700 nm.

A veces, como se observa en la Figura XIV.7, dos ejes que no explican una alta proporción de la varianza, como en este caso los ejes II (18,49%) y III (2,81%), pueden diferenciar claramente los grupos.

Figura XIV.6. Eje I del Análisis Discriminante aplicado a especies de fitoplancton usando características de las células.

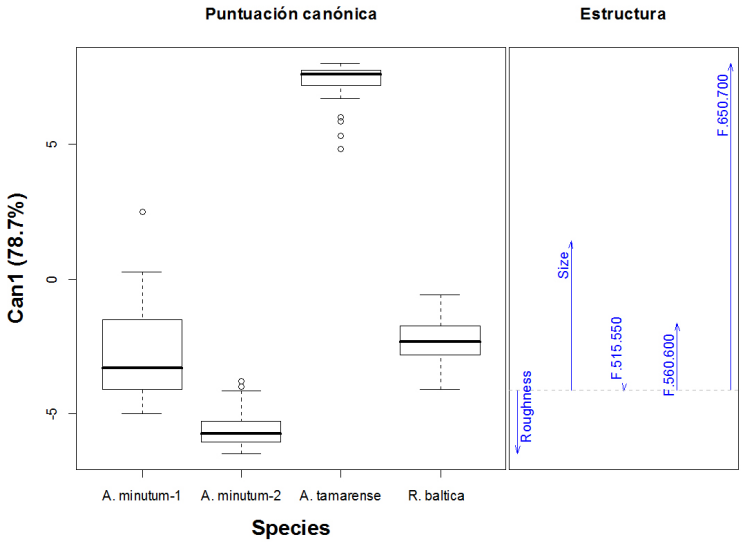
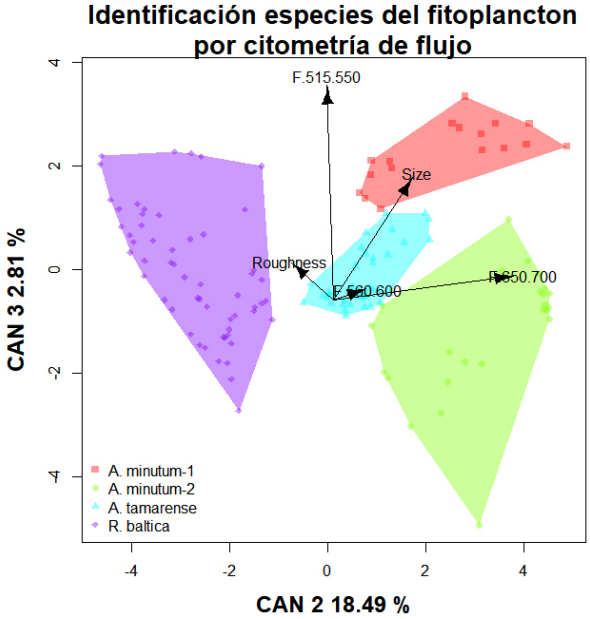


Figura XIV.7. Ejes II y III del Análisis Discriminante aplicado a especies de fitoplancton usando características de las células.



Value

Para el Análisis Discriminante Lineal se obtiene un gráfico con la primera dimensión y un *biplot* con las dos primeras dimensiones, que pueden ser otras que seleccione el usuario, en el caso de que sean más de dos categorías. En la ventana de *Replies* se obtiene la varianza explicada por cada función discriminante (solamente para el Análisis Discriminante Lineal) y el porcentaje de casos identificados correctamente, también por validación cruzada, para los dos tipos de discriminante. Los archivos CSV que se obtienen para el del Análisis Discriminante Lineal son: las coordenadas de los casos y las variables en los ejes, una tabla con el número de individuos clasificados en cada grupo por validación cruzada y un archivo con el grupo real al que pertenece cada caso y la predicción que hace el modelo. Para el Análisis Discriminante Cuadrático se obtiene una tabla con el número de individuos clasificados en cada grupo por validación cruzada y un archivo con el grupo real al que pertenece cada caso y la predicción que hace el modelo.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2013) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.6-5. Available at: http://CRAN.R-project.org/package=candisc.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:
data(ZXIV2)

#Ejemplo 1. Diferencias en la físico-química entre regiones

XIV2(data=ZXIV2, cat="Area", var=c("NO2","NO3","NH4","SiO2"), quadratic=TRUE,
CANPLOT=c("titles.1d=c('Puntuación canónica','Estructura')"),
MTEXT=c("text='Nutrientes en regiones de Colombia'", "side=3",
"line=1", "cex=1.7", "font=2"))

#Ejemplo 2. Diferencias en la abundancia de fitoplancton entre lagos

XIV2(data=ZXIV2, cat="Lago", var=c("NO2","NO3","NH4","SiO2", "Cianophyceae",
"Euglenophyceae", "Clorophyceae", "Zygophyceae", "Bacillariophyceae",
```

```
"Dinophyceae", "Xanthophyceae", "Crysophyceae", "Cryptophyceae"),
CANPLOT=c("titles.1d=c('Puntuación canónica','Estructura')"), quadratic=TRUE,
LEGEND=c("x='topright'", "legend=unique(datos3[,1])", "bty='n'","col=color1","pch=pcht"),
MTEXT=c("text='Nutrientes y abundancia de fitplancton\n en lagos de Colombia'",
"side=3", "line=0", "cex=1.6", "font=2"))

#Ejemplo 3

data(ZXIV3)

XIV2(data=ZXIV3, cat=("Species"), var=c("Roughness","Size","F.515.550","F.560.600","F.650.700"),
quadratic=TRUE, dimS=c(2,3), CANPLOT=c("titles.1d=c('Puntuación canónica','Estructura')"),
LEGEND=c("x='bottomleft'", "legend=unique(datos3[,1])","bty='n'","col=color1","pch=pcht"),
MTEXT=c("text='Identificación especies del fitoplancton\npor citometría de flujo'",
"side=3", "line=0", "cex=1.7", "font=2"), ellipse=FALSE, convex=TRUE)

## End(Not run)
```

XIV4 *DENDROGRAMA*

Description

Se aplica una clasificación jerárquica (Dendrograma).

Usage

```
XIV4(data, var, cat=NULL, ResetPAR=TRUE, PAR=NULL, HCLUST=NULL, PLOT1=NULL,
PLOT2=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL)
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis.
cat	Categorías de los casos.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
HCLUST	Se accede a la función hclust del paquete stats.
PLOT1	Se accede a la función que permite modificar la representación gráfica del dendrograma en el caso de que el argumento *cat* no es *NULL*.
PLOT2	Se accede a la función que permite modificar la representación gráfica del dendrograma en el caso de que el argumento *cat=NULL*.
LEGEND	Permite incorporar una leyenda.
AXIS	Permite añadir ejes.
MTEXT	Permite añadir textos en los márgenes del gráfico.
TEXT	Permite añadir texto en cualquier zona de la parte interna del gráfico.

Details

XIV. ANÁLISIS MULTIVARIANTE DE CLASIFICACIÓN

XIV.4. DENDROGRAMA

Es uno de los métodos más utilizados. Se obtiene una secuencia de particiones (clasificaciones) organizada jerárquicamente, desde una única clase que contiene a todos los elementos, hasta n clases, cada una de ellas con un solo elemento. La secuencia puede ser representada gráficamente mediante un árbol.

El algoritmo consiste en unir en primer lugar los dos elementos más próximos entre sí; la clase resultante se considera un nuevo elemento que sustituye a los dos anteriores. A continuación se buscan de nuevo los dos elementos más próximos entre sí, que se unirán en una clase nueva. El proceso continúa, y las diferentes clases se van uniendo sucesivamente en clases más grandes, hasta que todos los elementos forman finalmente una sola clase.

XIV.4.1. Criterio de distancia

Para decidir cuáles son los elementos de la muestra más próximos entre sí es necesario disponer de un criterio de distancia. Existen diferentes definiciones o tipos de distancia y es preciso elegir la más adecuada para los datos utilizados.

Euclídea: Es la más utilizada, y se corresponde con la distancia geométrica habitual en un espacio multidimensional. Se define como la raíz cuadrada de la suma de los cuadrados de las diferencias en los valores de cada variable para los dos elementos considerados. La mayoría de los programas estadísticos consideran como opción por defecto la distancia euclídea. Distancia Euclídea al cuadrado: Tiene como efecto asignar un mayor peso a los objetos más alejados.

Distancia Manhattan: Es el promedio de la diferencia entre dimensiones. El nombre proviene de su relación con la distancia real entre dos puntos en una ciudad con calles que se cortan perpendicularmente, donde en general la línea recta no es un camino posible.

Distancia de Mahalanobis: Cada dimensión está ponderada por la inversa de su desviación típica. Esta distancia tiene una importante propiedad: es invariante ante cambios en la unidad de medida, lo que la hace especialmente útil cuando las variables son heterogéneas.

XIV.4.2. Criterio de aglomeración

Cuando comenzamos el proceso de clasificación, al unir en una sola clase los dos elementos más próximos entre sí, esta clase pasa a considerarse un nuevo elemento que sustituye a los dos antiguos. Para continuar el algoritmo es necesario ahora calcular la distancia entre ese nuevo elemento y todos los anteriores, y para ello existen distintas opciones, que determinarán el criterio de aglomeración o conglomeración.

Existen diferentes criterios de distancia entre clases, o entre clases y elementos, que se corresponden con el concepto matemático de ultramétrica, más restrictivo que el concepto clásico de distancia o métrica:

Enlace simple (vecino más próximo): La distancia entre dos grupos es la distancia entre los dos objetos más próximos entre los diferentes grupos. Emplear este procedimiento de conglomeración puede dar lugar en ocasiones a un "efecto de encadenamiento", por el cual en cada fase se une una clase con un elemento aislado en lugar de unirse entre sí dos clases con varios elementos cada una, dando lugar a una secuencia jerárquica de clasificaciones atípicas de escasa utilidad práctica.

Enlace completo (vecino más lejano): La distancia entre dos grupos es la mayor distancia entre dos objetos de ambos grupos.

Enlace medio o vinculación inter-grupo: distancia media entre todos los pares de objetos de los dos grupos.

Vinculación intra-grupo: distancia media entre todos los pares de objetos del grupo unión de ambos grupos. Considera en el promedio no solo las distancias entre elementos de los dos grupos, sino también las distancias entre elementos del mismo grupo.

Método de McQuitty: la distancia a cualquier objeto desde un grupo que es unión de otros dos es la media de las distancias de aquellos dos ponderada por el número de unidades de cada grupo.

Enlace de centroides: distancia entre las medias o centroides de los dos grupos.

Método de Ward: En cada paso se unen los dos elementos que dan lugar a una menor pérdida de información, considerada como la suma de cuadrados de distancias de cada objeto al centro de su clase (mínima varianza o inercia).

XIV.4.3. Dendrograma

El principal resultado de esta técnica es un gráfico, denominado dendrograma o árbol de clasificación, que permite ver el proceso completo de clasificación desde su inicio con cada elemento en una clase hasta el final con una única clase en la que están todos los elementos de la muestra. En el gráfico se ven todas las particiones o clasificaciones obtenidas y su relación jerárquica. Para su mejor interpretación se pueden etiquetar los casos o elementos clasificados.

A partir del dendrograma es posible elegir la partición de mayor interés, en función de las distancias entre clases, que se miden en la escala lateral. La distancia entre dos clases viene dada por la longitud de la rama que las une en el árbol; en general se busca en el árbol la mayor separación entre particiones, que permite visualizar la clasificación más natural del conjunto de objetos.

El algoritmo exige en cada paso el cálculo de la matriz de distancias entre todos los elementos; cuando la muestra de datos contiene un número grande de objetos (varios cientos, o más de un millar) el procedimiento se hace excesivamente lento, de tal modo que puede requerir muchos minutos o incluso horas antes de alcanzar el resultado final. Por otra parte el gráfico obtenido, dendrograma, no es fácilmente interpretable cuando es demasiado grande. El método de clasificación jerárquica no es útil para muestras grandes, aunque en ese caso puede todavía ser aplicado a una parte de los datos (a una muestra aleatoria de los datos).

FUNCIONES

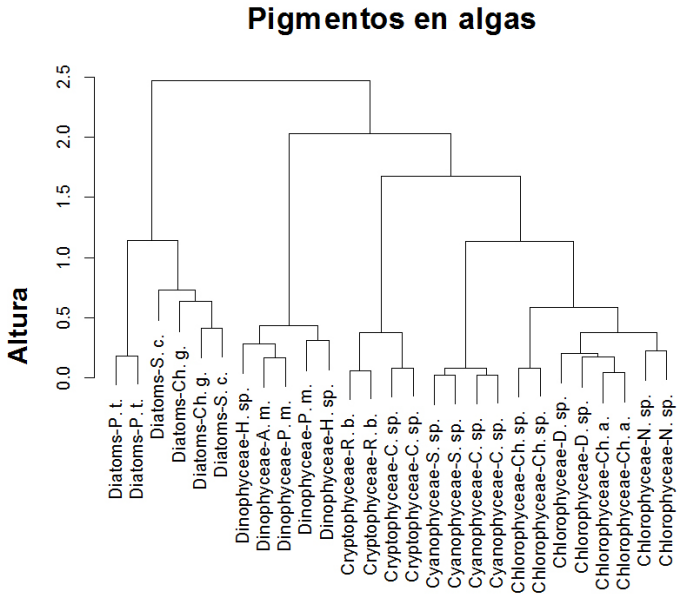
Se utiliza la función hclust del paquete stats.

EJEMPLO

Datos de la concentración relativa de 19 pigmentos, en relación con la concentración de clorofila a, de especies del fitoplancton que pertenecen a diferentes clases de algas: Diatomeas, Clorofíceas, Dinofíceas, Cianofíceas y Criptofíceas. El objetivo es determinar si es posible diferenciar las clases de algas en función de su composición relativa de pigmentos.

En el gráfico que se obtiene (Figura XIV.9), se observa que las diferentes clases de algas se agrupan perfectamente en función del tipo de pigmentos que tienen. Las Diatomeas forman un grupo muy diferente del resto, así como Criptofíceas y Dinofíceas.

Figura XIV.9. Clasificación jerárquica de especies del fitoplancton a partir de su composición de pigmentos.



Value

Se obtiene un gráfico denominado dendrograma.

Examples

```
## Not run:

data(ZXIV4)

XIV4(data = ZXIV4 , cat = "Species" , var = c("Chlorophyll.c2", "Chlorophyll.c1",
"Methyl.chlorophyllide.a", "Peridin", "Fucoxanthin", "cis.neoxanthin",
"Violaxanthin", "Diadinoxanthin", "Dinoxanthin", "Alloxanthin", "Zeaxanthin",
"Lutein", "Chlorophyll.b", "Chl.b.epir", "Chlorophyll.a.alllomer",
"Chlorophyll.a.epimer", "beta.u.carotene", "beta.e.carotene", "beta.beta.carotene"),
PLOT1 = c("main='Pigmentos en algas'", "sub=''", "xlab=''", "ylab='Altura'"))


## End(Not run)
```

XIV5 *MAPA DE CALOR*

Description

Se aplica un mapa de calor.

Usage

```
XIV5(data, cat, var, ResetPAR=TRUE, PAR=NULL, HEATMAP=NULL, mean=TRUE,
LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL)
```

Arguments

`data`	Archivo de datos.
`cat`	Categorías de los casos.
`var`	Variables que se incluyen en el análisis.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`HEATMAP`	Se accede a la función heatmap del paquete stats.
`mean`	Si es TRUE el mapa de calor se realiza utilizando el valor medio cada uno de los grupos de la variable *cat*, en vez de utilizar todos los casos.
`LEGEND`	Permite incorporar una leyenda.
`AXIS`	Permite añadir ejes.
`MTEXT`	Permite añadir textos en los márgenes del gráfico.
`TEXT`	Permite añadir texto en cualquier zona de la parte interna del gráfico.

Details

XIV. ANÁLISIS MULTIVARIANTE DE CLASIFICACIÓN

XIV.5. MAPA DE CALOR

El mapa de calor muestra, con distintos colores y tonos, la intensidad de la relación entre los grupos definidos por el usuario. Aplicado a una matriz de datos, con casos y variables (filas y columnas), clasifica ambos simultáneamente elaborando un dendrograma marginal para las filas y otro para las columnas, ordenados conjuntamente. Los colores indican la intensidad de la relación o los valores de la variable representada, desde el blanco (máxima relación o valor más alto), pasando por el amarillo y naranja hasta el rojo intenso (mínimo valor), aunque estos colores por defecto pueden ser modificados libremente por el usuario.

FUNCIONES

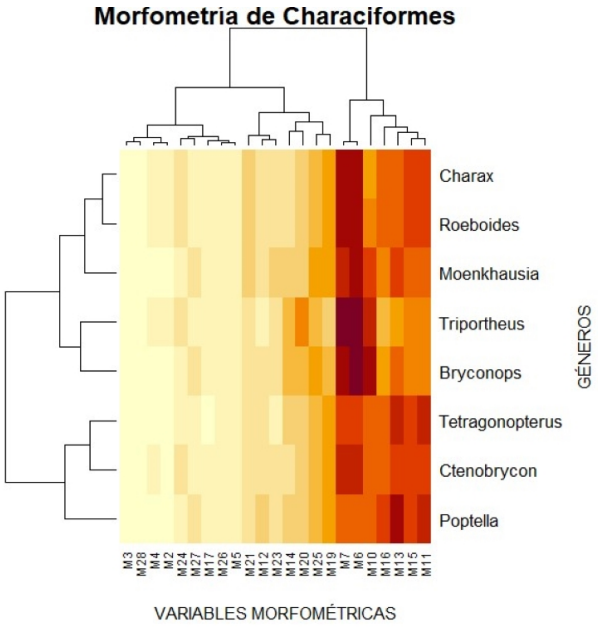
Se utiliza la función heatmap del paquete stats.

EJEMPLO

Datos morfométricos de varias especies del orden Characiformes, como la longitud de la base de la aleta dorsal (M12), altura del cuerpo (M11), etc. Para más detalles ver Guisande et al. (2010). El objetivo es determinar cuales son las variables que mejor discriminan entre géneros.

En el script, el argumento *margin* es un vector con 2 números que especifica los márgenes que se dejan para las filas y las columnas.

Figura XIV.10. Mapa de calor diferenciando géneros del orden Characiformes a partir de medidas morfométricas.



En un mapa de calor las variables que tienen las misma tonalidad, es decir, que tienen el mismo color a lo largo de todo el gradiente, son las que menos influyen en la clasificación. En la morfología de los peces, las variables que menos influyen en la clasificación son el grupo de variables de la izquierda, de la M3 a la M5, que forman un grupo diferenciado en el dendrograma de la parte superior (de columnas). Este grupo de variables contribuyen menos a distinguir los géneros de peces, ya que todas ellas tienen casi la misma tonalidad y, por ello, forman un único rectángulo, casi en su totalidad de color rojo, en la parte izquierda del gráfico. El grupo de variables de la M21 a la M19 permiten una mayor discriminación que las anteriores, ya que presentan una mayor variedad de totalidades. Las variables M7 y M6 también contribuyen a diferenciar los géneros, y el hecho de que estén diferenciadas en un bloque independiente parece indicar que discriminan bien a algún género en particular. También existen diferencias entre géneros, en el grupo de variables de la derecha, de la M10 a la M11.

Value

Se obtiene un gráfico denominado mapa de calor.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Examples

```
## Not run:
```

```
data(ZX3)

XIV5(data=ZX3,  cat="Genus", var=c("M2", "M3", "M4", "M5", "M6", "M7" ,"M10",
"M11", "M12" ,"M13", "M14", "M15", "M16", "M17", "M19", "M20", "M21", "M23",
"M24", "M25", "M26", "M27", "M28"), HEATMAP=c("margins=c(5,10)",
"xlab ='VARIABLES MORFOMÉTRICAS'", "ylab= 'GÉNEROS'", "main = 'Morfometría de Characiformes'"))


## End(Not run)
```

XV1 *MODELO ARIMA*

Description

Se estiman modelos ARIMA manual y automatizado

Usage

```
XV1(data, varY, varX=NULL, frequency, start, order=NULL, seasonal=NULL,
ResetPAR=TRUE, PAR=NULL, TS=NULL, PLOTFORECAST1=NULL, PLOTFORECAST2=NULL,
TSDISPLAY=NULL, NDIFFS=NULL, ARIMA=NULL, AUTOARIMA=NULL, FORECAST=NULL,
LINES=NULL, SEASONPLOT=NULL, file="Output.txt")
```

Arguments

data	Archivo de datos.
varY	Variable dependiente.
varX	Variables independientes.
frequency	Número de observaciones por unidad de tiempo.
start	La fecha de la primera observación. Puede ser un único número o un vector de dos números enteros, que especifican una unidad de tiempo natural.
order	Componente no estacional del modelo. Vector con tres números (p, d, q) que corresponden a la componente auto regresiva (p), la componente de integración (d), y la componente de media móvil (q).
seasonal	Componente estacional del modelo. Vector con tres números (p, d, q) que corresponden a la componente auto regresiva (p), la componente de integración (d), y la componente de media móvil (q).
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
TS	Se accede a la función ts que permite modificar los parámetros de la serie temporal.

PLOTFORECAST1	Se accede a la función plot.forecast que permite modificar los gráficos de las variables en el tiempo.
PLOTFORECAST2	Se accede a la función plot.forecast que permite modificar el gráfico de predicciones.
TSDISPLAY	Se accede a la función tsdisplay que permite modificar el gráfico de residuos y los gráficos de la función de autocorrelación simple (*ACF*) y el coeficiente de correlación parcial (*ACF* parcial o *PACF*).
NDIFFS	Se accede a la función ndiffs que calcula si la serie es estacionaria o no estacionaria.
ARIMA	Se accede a la función Arima que estima el modelo ARIMA manual.
AUTOARIMA	Se accede a la función auto.arima que estima el modelo ARIMA automatizado.
FORECAST	Se accede a la función forecast que realiza las predicciones del modelo.
LINES	Se accede a la función lines que permite modificar las líneas del gráfico de predicciones.
SEASONPLOT	Se accede a la función seasonplot que representa el gráfico estacional.
file	ARCHIVO TXT. Nombre del archivo de salida con los resultados de los análisis.

Details

XV1. SERIES TEMPORALES

Una serie temporal es un conjunto de observaciones de una variable tomadas a lo largo de intervalos regulares de tiempo, como el número de automóviles vendidos por un fabricante cada mes durante los últimos diez años. Las series temporales aparecen prácticamente en todos los campos de actividad. El interés de su análisis estadístico radica en el estudio del comportamiento de la serie, lo que permite explicar sus variaciones y, sobre todo, en la posibilidad de predecir valores futuros.

El principio fundamental que subyace al análisis de las series de tiempo es que los valores actuales (y futuros) de la serie dependen en cierta medida de los valores que ha tomado en el pasado, de modo que podríamos considerar que la serie tiene una cierta inercia que impide -o hace poco probables- cambios demasiado bruscos en períodos cortos de tiempo. Esa inercia puede ser utilizada para la predicción, admitiendo que en el futuro se comportará de modo similar a como lo hizo en el pasado. El éxito en el análisis y en la predicción de valores futuros dependerá sobre todo de la disponibilidad de una muestra amplia y fiable, y de la regularidad de la serie. Lamentablemente muchas series son dinámicas o cambiantes, y no se comportan de acuerdo con esquemas rígidos, de manera que el modelo estadístico utilizado es solamente una aproximación más o menos razonable al verdadero funcionamiento de la serie. En esos casos las predicciones deben hacerse únicamente a plazo corto, y con un grado de fiabilidad limitado.

En una serie temporal los estudios más frecuentes que se suelen hacer son la creación de un modelo y el análisis de los ciclos. Para más detalles sobre series temporales véase Guisande et al. (2011).

XV1.1. MODELO ARIMA

Existen muchos modelos estadísticos aplicados a las series temporales, entre los cuales destaca por su potencia y eficacia el modelo ARIMA (Box & Jenkins, 1976).

A la hora de hacer un modelo se deben tener en cuenta los siguientes análisis: la autocorrelación, si la serie crece, decrece o es estacionaria, la estacionalidad y, por último, ver si hay variables que pueden estar relacionadas con la variable dependiente.

XV1.1.1. Autocorrelación

Para saber algo más de la serie es útil estudiar la correlación o asociación entre los valores de la serie en cada período y los valores en los períodos anteriores, es decir, la relación entre la serie y ella misma (autocorrelación) desplazada uno o más períodos.

Se suele utilizar la función de autocorrelación simple (ACF), la cual muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma.

XV1.1.2. Series estacionarias y no estacionarias

El modelo ARIMA supone como condición previa- que la serie es estacionaria, es decir que su comportamiento se mantiene sin cambios en el tiempo, ya que todo el planteamiento de esta metodología se basa en que la evolución en el futuro será similar a la que ha tenido en el pasado (inercia de la serie). En especial se debe comprobar que su valor medio permanece estable (no tiene tendencia creciente o decreciente) y su variabilidad es también estable, no cambia (no aumenta o disminuye). Si no se cumplen estas condiciones es necesario hacer algún tipo de transformación de la variable, generalmente sencilla, que conduzca a una serie estacionaria. Si no existe ninguna transformación adecuada, el modelo ARIMA no debe ser aplicado.

Dickey & Fuller (1979) desarrollaron un procedimiento para probar si una variable tiene una raíz unitaria o, de manera equivalente, que la variable sigue una ruta aleatoria. Hay una extensión de la prueba de Dickey-Fuller llamada la prueba aumentada Dickey-Fuller (ADF), que elimina todos los efectos estructurales (autocorrelación) en la serie de tiempo. Esta prueba también se usó para determinar si la serie es estacionaria.

XV1.1.3. Estacionalidad

Las series temporales pueden ser estacionales o no estacionales. Muchas series temporales contienen variaciones estacionales, o pautas de comportamiento que se repiten con regularidad: El número de turistas es sistemáticamente mayor en verano que en invierno, el volumen de compras en unos grandes almacenes es más alto durante los primeros días de cada mes, el número de automóviles que transitan por una calle es mayor entre las 19h y las 21h que en el intervalo 22h-24h. Por tanto, la componente estacional se define como aquellos movimientos de la serie que se repiten de forma periódica, siendo la periodicidad inferior al año. La estacionalidad puede estar referida a diferentes períodos de tiempo (meses, días, horas, etc.), y en algunos casos es sumamente regular. Para detectar pautas de estacionalidad es muy útil representar gráficamente la serie.

Los modelos ARIMA permiten incluir una componente estacional, que puede ser utilizada con fines de predicción, y que presenta la misma estructura (p, d, q) que la parte no estacional. Estos componentes (p, d, q) se explican más adelante, cuando se analice la construcción del modelo ARIMA.

En aquellas series temporales en las que existe estacionalidad es necesario corregir este efecto, para una mejor interpretación del modelo, y se recurre a la desestacionalización o corrección estacional. Para realizar esta operación es necesario aislar en primer lugar la componente estacional, lo que posibilitará su posterior eliminación.

XV1.1.4. Covariables

Es posible incluir en el modelo adicionalmente otras variables explicativas o covariables que suponemos que tienen relación causal con nuestra variable dependiente, o que al menos están estrechamente relacionadas con ella.

Con ello, además de identificar posibles variables que influyen en nuestra serie temporal, si la relación entre estas covariables y nuestra variable dependiente es significativa, se debería esperar que el modelo mejore, es decir, que sea más predictivo.

XV1.1.5. Construcción del modelo

En los modelos ARIMA los valores de la serie se obtienen mediante la combinación de tres componentes (*AR*, *I*, *MA*) o también llamados (*p*, *d*, *q*):

1) Componente auto regresiva (*p*). Muestra la parte del valor de la variable que depende de los valores anteriores, de la historia de la serie. Se obtiene mediante un modelo de regresión lineal en el que la variable dependiente es la componente *AR* del período actual, y la variable independiente o explicativa el valor inmediatamente anterior o un cierto número *p* de valores anteriores de la misma serie. Por ejemplo, el valor AR del mes de junio depende del valor de la serie en el mes de mayo (*p* = 1), o de los valores de marzo, abril y mayo (*p* = 3). El valor de *p*, el orden de auto regresión, debe ser fijado para cada aplicación.

2) Componente de integración (*d*). Una serie es estacionaria si su media, su variabilidad y su estructura de correlaciones no cambian con el tiempo. Si la serie no es estacionaria, sino que tiene un comportamiento global creciente o decreciente, existe un componente de integración que debe ser determinado en primer lugar. La eliminación de este componente permite considerar la serie como estacionaria, y debe ser integrado posteriormente mediante un proceso inverso al de su eliminación. Para el cálculo de la componente de integración se obtienen las diferencias entre valores consecutivos de la serie, repitiendo este procedimiento sobre las diferencias obtenidas el número de veces *d* (orden de integración) que sea necesario para conseguir una serie estacionaria.

3) Componente de media móvil (*q*). En general las series no se comportan de modo exacto en relación con el modelo, sino que existe un elemento residual, que explica las diferencias entre el valor observado en cada período y el valor obtenido mediante el modelo auto regresivo. Este elemento representa la innovación en la serie, la parte que no responde a la inercia o al comportamiento histórico anterior, sino que es nueva en cada período. Analizando la serie podemos determinar estos residuos o términos de innovación en los períodos de tiempo ya transcurridos. La componente MA de cada período se obtiene como media de un cierto número *q* de estos residuos en los períodos anteriores, siendo *q* el orden de medias móviles.

En general, los valores de *p*, *d* y *q* son pequeños (cero o uno). Los valores mayores de estos parámetros del modelo permiten un mejor ajuste a los datos de la muestra, pero pueden producir también un efecto de sobreajuste al introducir en el modelo aspectos propios de la muestra disponible, no representativos de la población de la que se ha obtenido sino particulares de la muestra, de manera que el modelo puede ser muy adecuado para la muestra pero poco apropiado para la población de la que procede. En general se prefiere un modelo simple, que puede ser más adecuado para los fines de predicción, especialmente si la muestra no es muy grande, o si la serie no es muy estacionaria en su comportamiento sino ligeramente cambiante.

FUNCIONES

La serie temporal se realiza con la función ts del paquete base stats. El modelo ARIMA manual con la función Arima, el modelo ARIMA automatizado se estima con la función auto.arima, el gráfico representando las variables en el tiempo se realiza con la función plot.forecast, para representar la función de autocorrelación simple (*ACF*) y el coeficiente de correlación parcial (*ACF* parcial o *PACF*) se utiliza la función tsdisplay, para calcular si la serie es estacionaria o no estacionaria se usa la función ndiffs, las predicciones del modelo se estiman con la función forecast y el grafico estacional se realiza con la función seasonplot, todas ellas del paquete forecast (Hyndman et al., 2018). La líneas del gráfico de predicciones se realizan con la función lines del paquete base graphics. La prueba Box-Ljung se realiza con la función Box.test y la normalidad de los residuos se analiza con las función shapiro.test, ambas del paquete base stats. También se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-

Smirnov con la corrección de Lilliefors a los residuos. El test aumentado de Dickey-Fuller se realizó con la función adf.test del paquete tseries (Trapletti & Hornik 2017).

EJEMPLO

El factor de condición mensual (*FC*) de la anchoa (*Engraulis encrasicolus*) a lo largo de varios años es un indicador del estado reproductivo de los individuos. El objetivo es determinar el efecto de los valores medios mensuales de la temperatura (en degree C), clorofila (en mgm^{-3}) y estabilidad de la columna de agua (en m^3s^{-3}) sobre el estado reproductivo de esta especie en el estrecho de Sicilia (Basilone et al., 2006). Como *FC* y el resto de variables pueden covariar con el tiempo, es necesario cuantificar primero el posible efecto del tiempo sobre *FC*, en otras palabras, introducir el potencial efecto del tiempo sobre *FC* en el modelo, para poder determinar el efecto real de las variables mencionadas anteriormente sobre *FC*. Por lo tanto, en primer lugar siempre es conveniente analizar la evolución temporal de la serie, en este caso de *FC*, sin tener en cuenta el efecto de las variables independientes.

Es importante mencionar que no es necesario que la serie este completa, ya que se hace una estimación de los datos que faltan. Sin embargo, es necesario que exista esa fila indicando la fecha a la que corresponde. Por ejemplo, si faltase el dato de enero de 1998, es necesario que aparezca la fila indicando que es enero de 1998, y luego dejando en blanco las columnas que falten.

MODELO ARIMA AUTOMATIZADO SIN VARIABLES INDEPENDIENTES

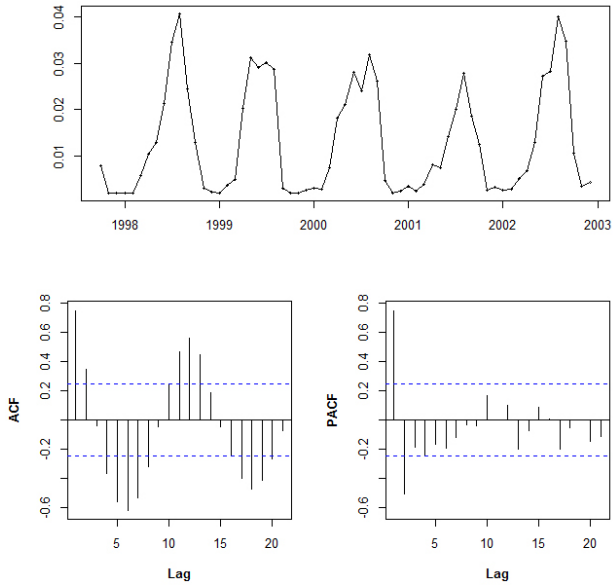
En el script el argumento *frequency=12* porque al ser los muestreos mensuales, se considera el año como unidad de tiempo. El argumento *start=c(1997,10)* porque el primer muestro fue en octubre del año 1997. Al no especificar nada en los argumentos *order* y *seasonal*, que son las condiciones por defecto, entonces se realiza el modelo ARIMA automatizado, es decir, se construye un modelo ARIMA mediante un modelizador experto.

Autocorrelación

Se observa una clara estacionalidad anual (Figura XV1.1), ya que todos los años se repite la misma pauta, con valores bajos de *FC* en invierno y altos en verano. Por otra parte la serie en su conjunto parece que no muestra tendencia (es horizontal), ni ciclo. La función de autocorrelación simple (*ACF*), que se observa en la Figura XV1.1, muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma. Cuando el valor de la serie Y_t depende fuertemente del valor anterior $Y_t - 1$ existirá un coeficiente de correlación alto R_1 entre la serie y la misma serie desplazada un período, pero dado que $Y_t - 1$ depende a su vez del valor anterior a él $Y_t - 2$, también será alto el coeficiente de correlación con la serie desplazada dos períodos, R_2, y así sucesivamente. El gráfico muestra coeficientes que disminuyen gradualmente. Se representa en esta función el coeficiente correspondiente al retardo cero, que siempre vale uno (no debe ser interpretado, solo sirve como referencia de comparación).

Es posible que una parte de esa dependencia que mide el coeficiente de correlación simple sea en realidad indirecta, a través de los valores intermedios, por lo que solamente los valores más próximos al actual deben ser considerados (tal vez uno o dos solamente). Por ello, el coeficiente de correlación parcial (*ACF* parcial o *PACF*) es más adecuado para analizar estas relaciones de dependencia que el coeficiente de correlación simple, ya que al eliminar la influencia de las restantes variables (los valores intermedios) aísla el efecto propio de cada retardo sobre el valor actual de la serie, y permite estudiar la correlación entre la serie y la misma serie retardada sin incluir los efectos indirectos. En la función de autocorrelación parcial (Figura XV1.1) comprobamos que solamente los dos primeros son significativos, y el primero tiene un valor elevado. Con estos resultados, un modelo adecuado podría tener $p = 1$ o $p = 2$ y $q = 0$.

Figura XV1.1. Representación temporal del factor de condición de la anchoa, y los valores de ACF y ACFP.



Serie estacionaria o no estacionaria

Los resultados muestran que la serie tiene un valor $d = 0$, por lo tanto, la serie es estacionaria. Si la serie no fuera estacionaria sería suficiente con poner el valor de d recomendado en el modelo ARIMA, ya que lo que hace el modelo ARIMA es diferenciar la serie d veces, con lo que la serie se convierte en estacionaria, y una vez aplicado el modelo integra la serie resultante d veces para deshacer el cambio. El resultado de la prueba aumentada Dickey-Fuller (ADF) corrobora que la serie es estacionaria, porque la probabilidad es 0,99, por lo que se acepta la hipótesis nula de que la serie temporal es estacionaria.

```
        Augmented Dickey-Fuller Test

data:  serie1[, 1]
Dickey-Fuller = -5.6377, Lag order = 3, p-value = 0.99
alternative hypothesis: explosive
```

Construcción del modelo ARIMA automatizado

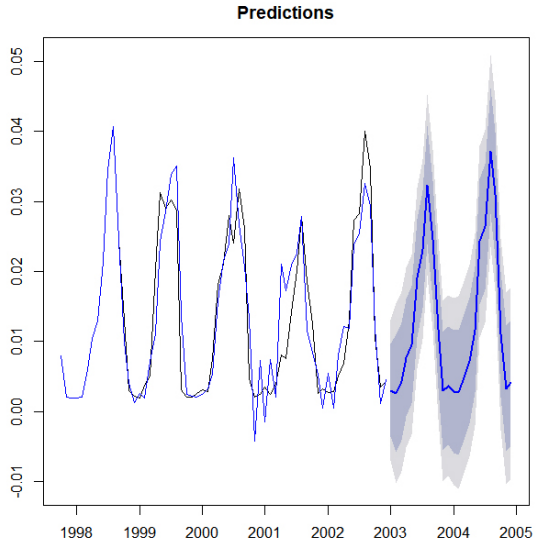
La función dispone de un procedimiento que busca entre los diferentes valores de los parámetros p, q, y d, tanto para la parte regular como para la parte estacional, el modelo más adecuado. Emplea el algoritmo de Hyndman & Khandakar (2008), que elige los parámetros minimizando el *AIC*, y los valores de d y D mediante test de raíz unitaria. Un método paso a paso permite limitar el campo de búsqueda, reduciendo el esfuerzo de cálculo y el tiempo total necesario (en todo caso, el proceso puede tardar a veces varios minutos). En nuestro ejemplo encuentra y ajusta un modelo ARIMA(0,0,1)(1,1,0), que considera óptimo.

```
ARIMA(0,0,1)(1,1,0)[12]

Coefficients:
          ma1     sar1
      0.7959  -0.6327
s.e.  0.1165   0.1056

sigma^2 estimated as 2.637e-05:  log likelihood=193.9
AIC=-381.8   AICc=-381.28   BIC=-376
```

Predicción de valores futuros

La función calcula las predicciones para el número de períodos *h* deseado, el cual se puede modificar accediendo a la función forecast con el argumento *FORECAST* y modificando el valor de *h*. Estas predicciones se muestran en la Figura XV1.2 con los intervalos de confianza. Se pueden colorear las bandas de confianza especificando *shaded=TRUE* dentro del argumento *FORECAST*, o simplemente dibujar líneas. El argumento *LINES* permite modificar los valores ajustados, o predicciones del modelo para todos los períodos observados.

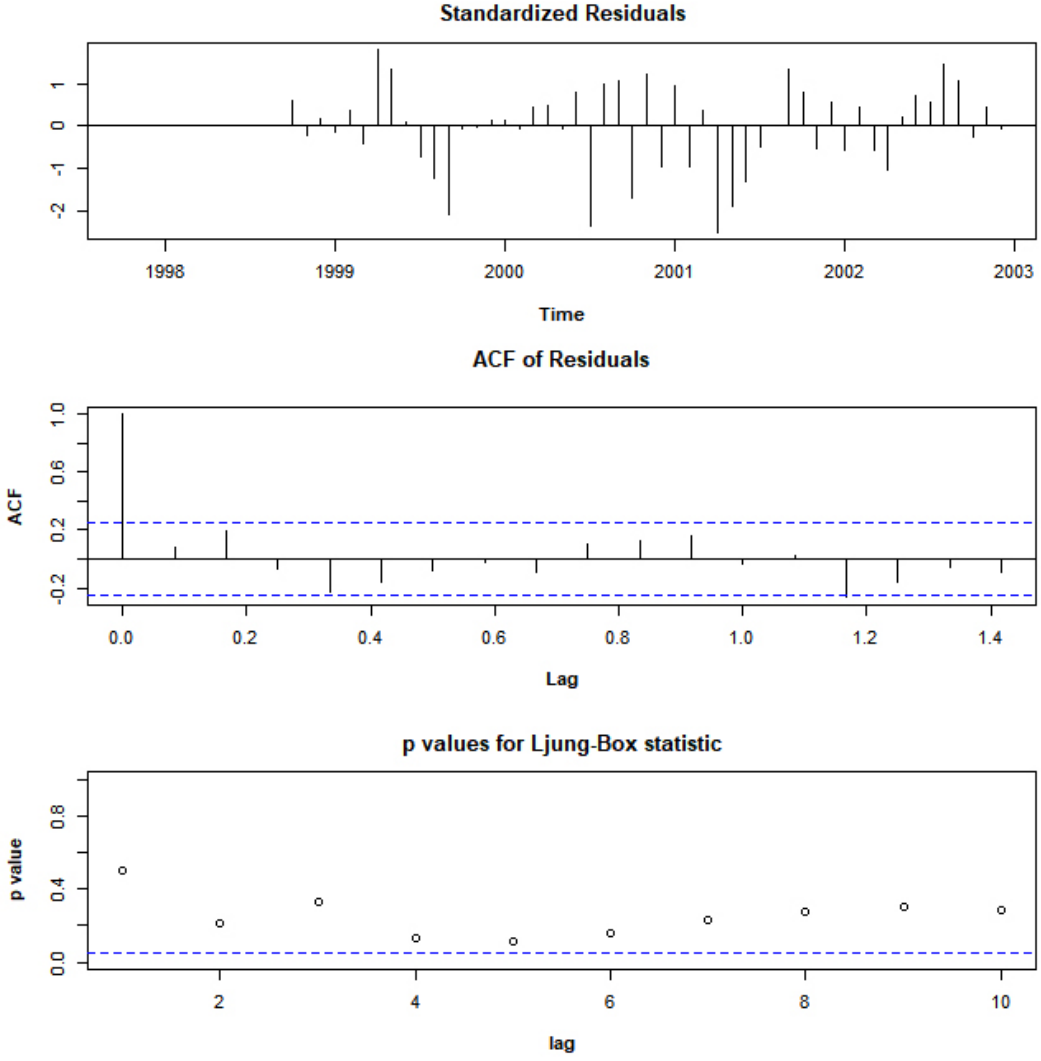
Figura XV1.2. Predicciones del modelo ARIMA.



Diagnósticos del modelo

Es conveniente comprobar que el modelo elegido es completo y no deben añadirse nuevos elementos. Para ello comenzamos analizando los residuos con un gráfico (Figura XV1.3). Podemos comprobar como los residuos parecen comportarse aleatoriamente, sin secuencias predecibles o una estructura de autocorrelaciones, es decir, de acuerdo con el modelo conocido como "ruido blanco". El gráfico *ACF* no debe mostrar autocorrelaciones importantes (recordemos que la primera, siempre igual a 1, no cuenta), como en este ejemplo. Al observar los *p*-valores correspondientes a la prueba de Ljung-Box, para todos los retardos no se rechaza la hipótesis de que las autocorrelaciones estimadas son normales con media cero y varianza $\frac{1}{n}$. Todos o casi todos los *p* valores deben estar por encima de la línea azul punteada del 5%. Por lo tanto, parece razonable concluir que los residuos se comportan como un ruido blanco, y el modelo es adecuado.

Figura XV1.3. Gráficos de residuos, *FAS* de residuos, y valores *p* del test Ljung-Box para cada retardo del modelo ARIMA automatizado.



La prueba Box-Ljung, que se muestra a continuación, realiza un test conjunto para contrastar la hipótesis nula de que se trata de ruido blanco. El valor p = 0,305 mayor que el nivel de significación habitual 0,05 indica que se puede aceptar que se trata de ruido blanco y, por lo tanto, los residuos no conservan una estructura que deba ser explicada mediante un modelo más complejo.

```
        Box-Ljung test

data:  t1$resid
X-squared = 13.94, df = 12, p-value = 0.3046
```

Es posible comprobar también la distribución Normal de los residuos aplicando el test de Shapiro-Wilk y el de Kolmogorov-Smirnov con la corrección de Lilliefors a los residuos. Los residuos no

tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov $p < 0{,}05$.

```
[1] "Shapiro-Wilk TEST"

[[15]]

     Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.94088, p-value = 0.004555


[[16]]
[1] "Kolmogorov-Smirnov TEST WITH Lilliefors CORRECTION"

[[17]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.174, p-value = 6.065e-05
```

Desestacionalización

En las series con estacionalidad acusada, como la de este ejemplo, es difícil apreciar las variaciones que se ocultan o enmascaran detrás de la estacionalidad y, por lo tanto, averiguar los cambios que se producen. Para ello podemos analizar con más detalle las variaciones estacionales y eliminar la estacionalidad de la serie para estudiar el resto.

La Figura XV1.4 muestra los valores de estacionalidad mensual. Como se puede apreciar claramente, los valores de estacionalidad alta son los que corresponden al mes de agosto y próximos a este mes, y la estacionalidad baja ocurre en los meses que van de noviembre a febrero, todos ellos en niveles similares.

La Figura XV1.5 muestra la serie desestacionalizada, donde es posible ver los cambios en la serie descontando el efecto de la estacionalidad.

Figura XV1.4. Valores mensuales de estacionalidad de la serie.

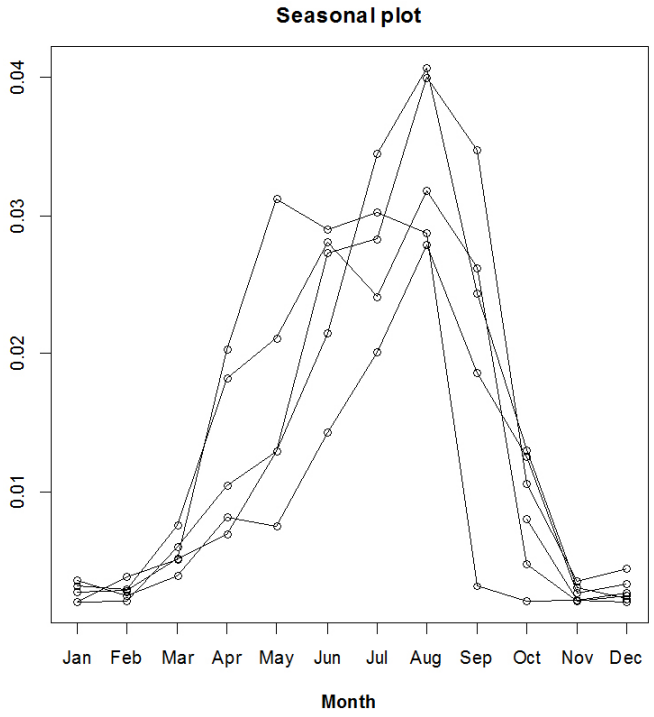
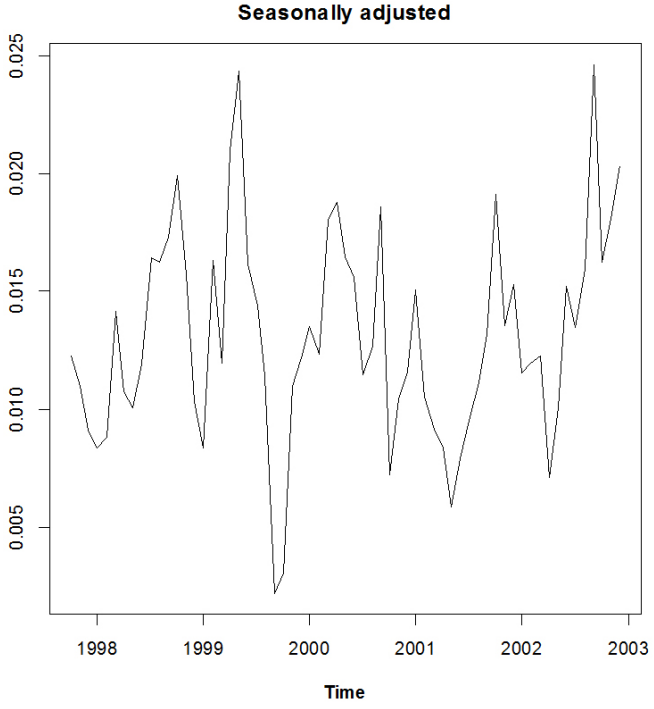


Figura XV1.5. Serie desestacionalizada.



MODELO ARIMA MANUAL SIN VARIABLES INDEPENDIENTES

En lugar de utilizar el modelizador experto, podemos decidir -a partir de los gráficos de autocorrelación o del conocimiento de la serie- un modelo concreto, sobre todo en este caso que no se cumplía la Normalidad de los residuos.

Para ello es necesario definir los argumentos *order*, que define la componente no estacional del modelo, y *seasonal*, que define la componente estacional del modelo. Ambos argumentos son un vector con tres números (*p*, *d*, *q*) que corresponden a la componente auto regresiva (*p*), la componente de integración (*d*), y la componente de media móvil (*q*).

El modelo ARIMA automatizado seleccionó los parámetros (0,0,1) para la componente no estacional y (1,1,0) para la componente estacional. En el modelo manual elegimos los siguientes parámetros: *order=c(0,0,1)* y *seasonal=c(2,1,0)*. Es decir, la única modificación es que cambiamos en la componente estacional en valor de *p*, que era 1, a un valor de 2. Por tanto, consideramos que la autocorrelación se da no solamente con el mes anterior, sino que un mes también está correlacionado con el valor de hace dos meses.

No tiene sentido cambiar la componente de integración (*d*) en el modelo manual, ya que la serie era estacionaria y, por tanto, el valor cero es el más adecuado.

Sería posible hacer pruebas y cambiar también la componente *q*, si consideramos que existe innovación en la serie, es decir, que existen unos cambios que no responden a la inercia o al comportamiento histórico anterior, sino que son nuevos en cada período. Como ya se explicó en la introducción, siempre es conveniente que el modelo sea lo más sencillo posible, es decir, que los valores de *p*, *d* y *q* sean pequeños.

La autocorrelación y si la serie es estacionaria o no estacionaria son los mismos resultados que se mostraron con el modelo automatizado.

Para ver si es mejor modelo que el obtenido con el sistema experto, la varianza estimada de los residuos *sigma^2 estimated* debe ser pequeña, y el valor de *AIC* es mejor cuanto más pequeño (o más negativo). En el caso del modelo automatizado el valor de AIC era -381,8. En los resultados que se muestran a continuación del modelo manual se observa que el valor de *AIC* es de -380,1. Por lo tanto, el modelo obtenido por el sistema manual es poco mejor, pero el cambio es muy pequeño.

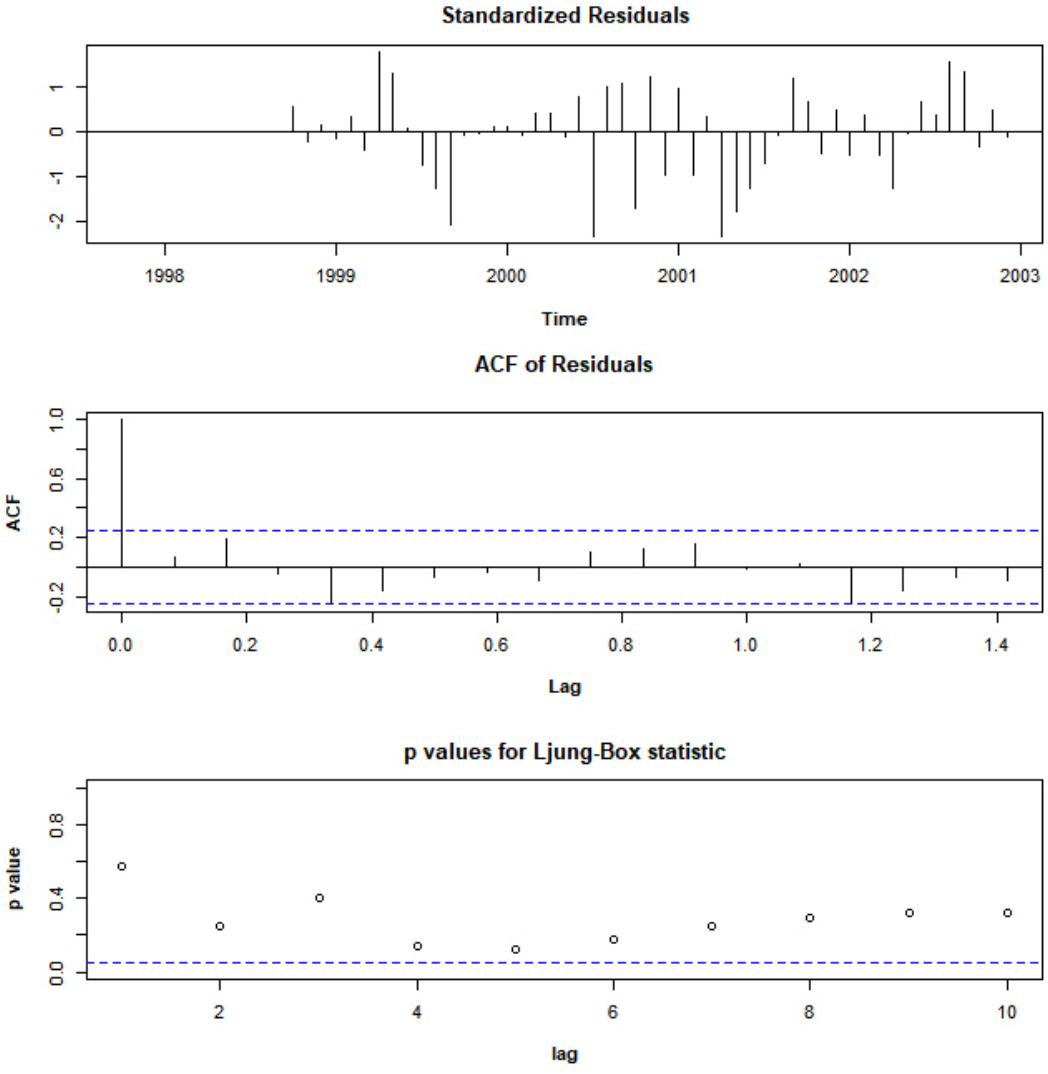
```
[1] "ARIMA MODEL"

[[8]]
Series: serie1[, 1]
ARIMA(0,0,1)(2,1,0)[12]

Coefficients:
         ma1     sar1     sar2
      0.7978  -0.6927  -0.1109
s.e.  0.1183   0.1497   0.1937

sigma^2 estimated as 2.67e-05:  log likelihood=194.06
AIC=-380.12   AICc=-379.25   BIC=-372.4
```

Diagnósticos del modelo

Al igual que antes, es necesario comprobar que el modelo elegido es completo y no deben añadirse nuevos elementos. Para ello, de nuevo comenzamos analizando los residuos con un gráfico (Figura

XV1.6). Podemos comprobar como los residuos parecen comportarse aleatoriamente, sin secuencias predecibles o una estructura de autocorrelaciones, es decir, de acuerdo con el modelo mencionado anteriormente como "ruido blanco". El gráfico *ACF* no debe mostrar autocorrelaciones importantes (recordemos que la primera, siempre igual a 1, no cuenta), como en este ejemplo. Al observar los *p*-valores correspondientes a la prueba de Ljung-Box, para todos los retardos no se rechaza la hipótesis de que las autocorrelaciones estimadas son normales con media cero y varianza $\frac{1}{n}$. Todos o casi todos los *p* valores deben estar por encima de la línea azul punteada del 5%, como en este ejemplo.

Figura XV1.6. Gráficos de residuos, *FAS* de residuos,
y valores *p* del test Ljung-Box para cada retardo del modelo ARIMA manual.



La prueba Box-Ljung, que se muestra a continuación, realiza un test conjunto para contrastar la hipótesis nula de que se trata de ruido blanco. El valor p = 0,344 mayor que el nivel de significación

habitual 0,05 indica que se puede aceptar que se trata de ruido blanco y, por lo tanto, los residuos no conservan una estructura que deba ser explicada mediante un modelo más complejo.

A diferencia del modelo automatizado, en este caso los residuos tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov p > 0,05.

```
     Box-Ljung test

data:  t1$resid
X-squared = 13.351, df = 12, p-value = 0.344


[[13]]
[1] "NORMALITY"

[[14]]
[1] "Shapiro-Wilk TEST"

[[15]]

     Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.94535, p-value = 0.007383


[[16]]
[1] "Kolmogorov-Smirnov TEST WITH Lilliefors CORRECTION"

[[17]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.17036, p-value = 9.892e-05
```

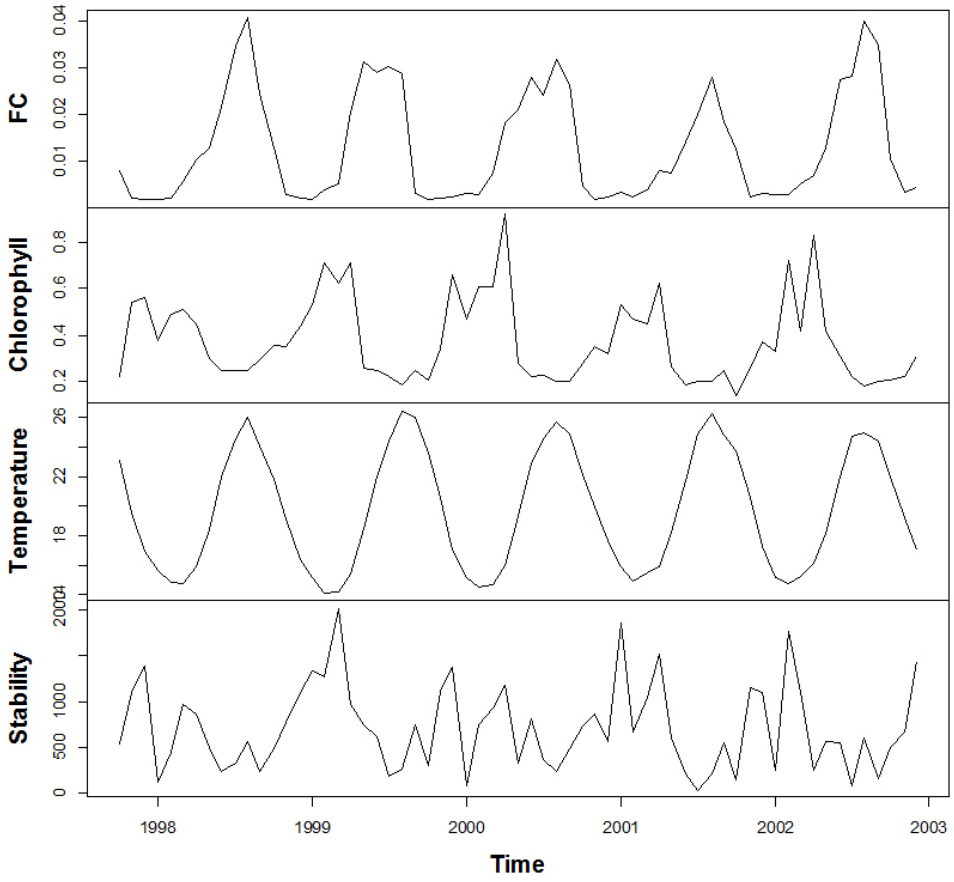
MODELO ARIMA MANUAL CON VARIABLES INDEPENDIENTES

Una vez que ya hemos analizado nuestra serie temporal, ahora podemos determinar el efecto de las variables sobre ella. En nuestro ejemplo, el objetivo era determinar el efecto de los valores medios mensuales de la temperatura, clorofila y estabilidad de la columna de agua sobre el *FC* de la anchoa.

Para incluir variables independientes en el modelo, simplemente hay que especificarlas en el argumento *varY*. Estas variables se pueden incluir tanto en el modelo automatizado como en el manual. Ya que el modelo manual era un poco mejor que el automatizado, realizaremos el ejemplo incluyendo las variables independientes con el modelo manual.

Cuando se incluyen variables independientes en el modelo sale un gráfico a mayores de los mostrados anteriormente (Figura XV1.7), que muestra la evolución temporal de la serie (Factor de condición de la anchoa, *FC*) y también de las variables independientes (clorofila, temperatura y estabilidad).

Figura XV1.7. Representación temporal del factor de condición de la anchoa, estabilidad, temperatura y clorofila, en el estrecho de Sicilia.



El valor de AIC con las covariables es -376,5, mientras que sin ellas era (-380,1). La diferencia es tan pequeña que podemos afirmar que las covariables no añaden información útil y, por lo tanto, el *FC* de la anchoa no depende de la temperatura, clorofila, ni estabilidad de la columna de agua.

```
[1] "ARIMA MODEL"

[[8]]
Series: serie1[, 1]
Regression with ARIMA(0,0,1)(2,1,0)[12] errors

Coefficients:
         ma1     sar1    sar2  Chlorophyll  Temperature  Stability
      0.8677  -0.6467  0.0046       0.0063        1e-03          0
s.e.  0.0809   0.1501  0.1944       0.0047        3e-04          0

sigma^2 estimated as 2.671e-05:  log likelihood=195.25
AIC=-376.5   AICc=-373.9   BIC=-362.98
```

Value

Se representan los siguientes gráficos: 1) evolución de las variables en el tiempo (solamente si se incluyen variables independientes), 2) los residuos, la función de autocorrelación simple (*ACF*) y el coeficiente de correlación parcial (*PACF*), 3) las predicciones del modelo, 4)test Ljung-Box para cada retardo del modelo, 5) el grafico estacional y 6) el gráfico desestacionalizado, los dos últimos se muestran si el argumento *frequency* es mayor de 1.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A., Vergara, A.R. & Maneiro, I. (2006) Effect of habitat conditions on reproduction of the European anchovy (*Engraulis encrasicolus*) in the Strait of Sicily. *Fisheries Oceanography*, 15: 271-280.

Box, G.E.P. & Jenkins, J.M. (1976) *Time series analysis: Forecasting and control*. Holden-Day,San Francisco, CA.

Dickey, D.A. & Fuller, W.A. (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427-431.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Hyndman, R.J. & Khandakar, Y. (2008) Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27: 1-22.

Hyndman, R.J., Athanasopoulos, G.,Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C. & Wang, E. (2018) Forecasting functions for time series and linear models. R package version 8.3. Available at: http://CRAN.R-project.org/package=forecast.

Trapletti, A & Hornik, K. (2017) tseries: Time Series Analysis and Computational Finance. R package version 0.10-38. Available at: https://CRAN.R-project.org/package=tseries.

Examples

```
## Not run:

data(ZXV1)

#Modelo ARIMA automatizado sin variables independientes

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10))

#Modelo ARIMA manual sin variables independientes

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10), order=c(0,0,1),
seasonal=c(2,1,0))

#Modelo ARIMA manual con variables independientes

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10), varX=c("Chlorophyll",
"Temperature", "Stability"), order=c(0,0,1), seasonal=c(2,1,0))
```

```
## End(Not run)
```

XV2 *CICLOS*

Description

Identificación de ciclos y análisis de la relación entre ciclos.

Usage

```
XV2(data, varY, varX=NULL, frequency, start, ResetPAR=TRUE, PAR=NULL,
TS=NULL, PLOTSERIE=NULL, SPECPGRAM=NULL, SPECAR=NULL, COHERENCE=NULL,
SPECTRUM=NULL, PLOTSPECTRUM=NULL, file="Output.txt")
```

Arguments

`data`	Archivo de datos.
`varY`	Variable dependiente.
`varX`	Variable independiente.
`frequency`	Número de observaciones por unidad de tiempo.
`start`	La fecha de la primera observación. Puede ser un único número o un vector de dos números enteros, que especifican una unidad de tiempo natural.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`TS`	Se accede a la función ts que permite modificar los parámetros de la serie temporal.
`PLOTSERIE`	Se accede a la función que permite modificar el primer gráfico con las series temporales.
`SPECPGRAM`	Se accede a la función spec.pgram que permite modificar el periodograma de la variable dependiente.
`SPECAR`	Se accede a la función spec.ar que permite modificar el gráfico de densidad de la variable dependiente.
`COHERENCE`	Se accede a la función coh que permite modificar el gráfico de coherencia entre la variable dependiente e independiente.
`SPECTRUM`	Se accede a la función spectrum que calcula el desfase entre la variable dependiente y la independiente.
`PLOTSPECTRUM`	Se accede a la función que permite modificar el gráfico que muestra el desfase entre la variable dependiente y la independiente.
`file`	ARCHIVO TXT. Nombre del archivo de salida con los resultados de análisis de los ciclos.

Details

XV1. SERIES TEMPORALES

XV1.2 CICLOS

El ciclo es una componente habitual en las series temporales (Guisande et al., 2011). Su forma típica es sinusoidal, con una fase de crecimiento hasta alcanzar el valor máximo, a la que sigue una fase de decrecimiento hasta alcanzar su valor mínimo, repitiéndose este esquema indefinidamente. A diferencia de las variaciones estacionales, no siempre es regular, y la duración de las distintas fases a menudo es cambiante, lo que dificulta su análisis. Está presente en prácticamente todas las series económicas, sin que se conozcan muy bien las razones de su aparición o las causas profundas de la existencia de ciclos.

El análisis espectral es un procedimiento estadístico que puede ser utilizado para identificar comportamientos periódicos en una serie temporal, y para estudiar la existencia de ciclos o variaciones regulares cuya longitud (período) es generalmente superior a un año. El procedimiento descompone el conjunto de todas las observaciones de la serie en elementos (componentes) periódicos de diferentes frecuencias, y permite identificar las más relevantes. Es particularmente útil para detectar estructuras cíclicas complejas, es decir cuando existen varios ciclos superpuestos de diferente período, ya que cuando existe un único patrón cíclico puede ser suficiente una representación gráfica (gráfico secuencial) para identificarlo.

FUNCIONES

La serie temporal se realiza con la función ts, la función spec.pgram se usa para obtener el periodograma de la variable dependiente, la función spec.ar se usa para representar el gráfico de densidad de la variable dependiente y el desfase entre la variable dependiente y la independiente se obtiene con la función spectrum, todas ellas del paquete base stats. El gráfico de coherencia entre la variable dependiente e independiente se realiza con la función coh del paquete seewave (Sueur et al., 2014). La estimación de los valores perdidos se realiza con la función na.approx del paquete zoo (Zeileis et al., 2014).

EJEMPLO

El ángulo que forman el sol y la luna es uno de los principales responsables de los cambios que se producen en la mareas. Cuando el sol y la luna están alineados, es decir, el ángulo es 0 (luna llena y luna nueva), las diferencias en altura entre pleamar y bajamar son máximas. Por el contrario, cuando el sol y la luna forman un ángulo de 90º (cuarto creciente y cuarto menguante) las diferencias en altura entre pleamar y bajamar son mínimas. Los datos que se usan son los ángulos entre el sol y la luna diarios, y la diferencia en metros entre pleamar y bajamar diarias, para 42º 8´Norte y 14º 13´Oeste del 1/1/2004 al 14/3/2004.

El primer objetivo es determinar el ciclo mareal. El segundo objetivo es demostrar el efecto del ángulo entre el sol y la luna sobre las mareas.

La función estima los valores perdidos, en el caso de que existan, pero es necesario que exista la fila con la fecha. En este ejemplo hay una fecha que falta, en concreto el día 3/1/2004.

La Figura XV2.1 muestra una clara variación cíclica en ambas series temporales, con 5 oscilaciones en el conjunto de los 74 días, lo que corresponde a un período de aproximadamente 15 días.

Podemos profundizar un poco más mediante un análisis espectral, que estudia el comportamiento o la respuesta de la serie a cada frecuencia diferente. En el periodograma que se muestra en la Figura XV2.2, se puede observar claramente un pico que destaca sobre el resto de valores. Ese máximo indica la frecuencia del ciclo que hemos observado anteriormente en el gráfico de secuencia. Si

hubiese más ciclos superpuestos, encontraríamos otros picos en el espectrograma mostrando las frecuencias asociadas a ellos.

Figura XV2.1. Evolución temporal del ángulo que forman el sol y la luna, y la diferencia entre pleamar y bajamar.

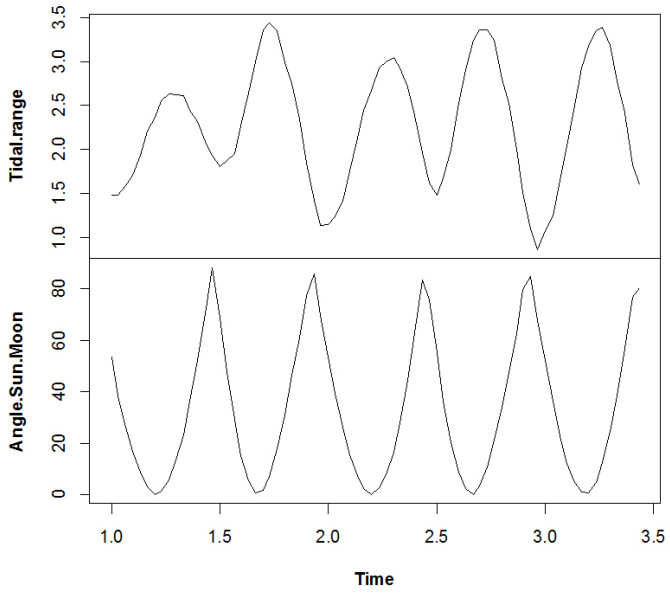
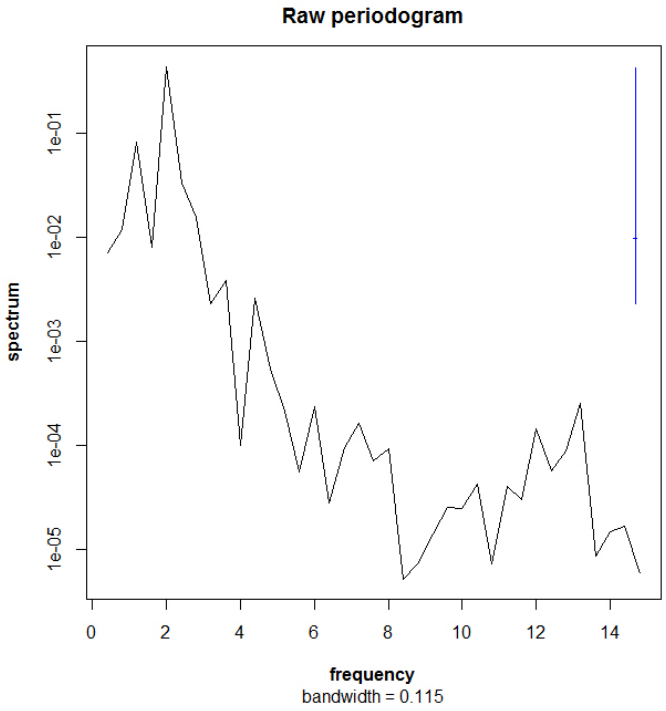
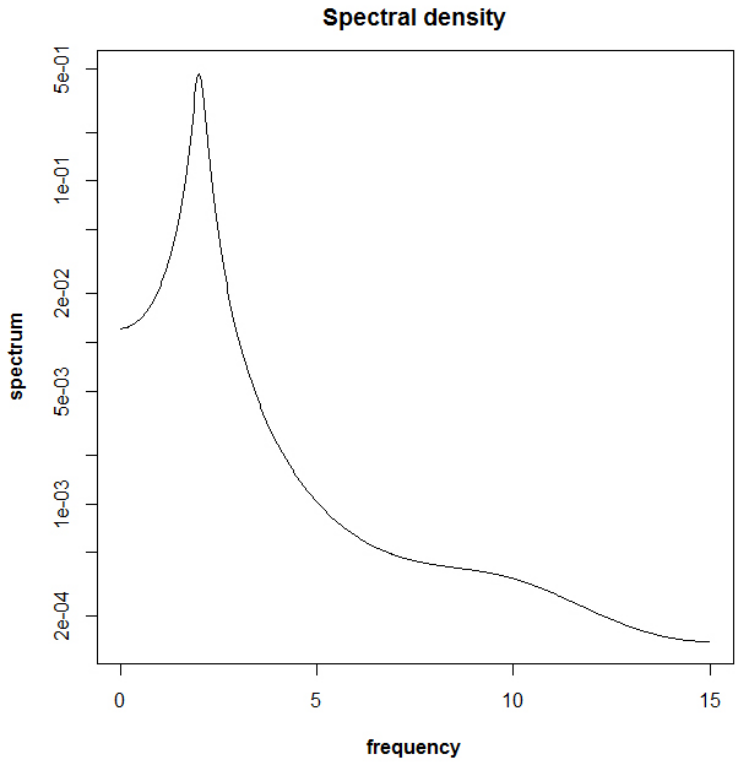


Figura XV2.2. Periodograma de la serie temporal de la diferencia entre pleamar y bajamar.



El periodograma anterior se puede representar mediante el gráfico de densidad espectral (Figura XV2.3), el cual es simplemente una versión suavizada del periodograma, eliminando mediante filtrado el ruido formado por las variaciones aleatorias, o en general cualquier variación de alta frecuencia. El efecto de suavizar los datos nos permite ver de forma más clara el pico que se observaba en el periodograma.

Figura XV2.3. Gráfico de densidad espectral.



Los resultados muestran que usando el primer gráfico (*Raw periodogram*) el periodo es de 15 días, mientras que con el segundo gráfico (*Spectral density*) es de 14,89 días. Es posible que cuando falten muchos valores en la serie, sea más conveniente utilizar el primer gráfico, es decir, el periodograma.

El segundo objetivo es averiguar si la diferencia entre pleamar y bajamar está relacionada con el ángulo entre el sol y la luna. Para ello debemos construir el espectro cruzado de ambas variables. El espectro cruzado se interpreta habitualmente a partir de algunas variables transformadas. Una de las más importantes es la coherencia cuadrada, representada en el diagrama siguiente, que toma valores entre 0 y 1. Un valor próximo a 0 indica que ambas variables no están relacionadas, y próximo a 1 indica una relación prácticamente lineal entre ambas. La escala horizontal muestra la frecuencia en kHz, ya que es la unidad más útil para el estudio de señales en el campo de la electrónica en el que esta técnica es usualmente aplicada. Deben multiplicarse por mil los valores para interpretarlos directamente como fracción de la unidad de tiempo empleada.

Como se puede observar en la Figura XV2.4 existe un valor próximo a 1 que destaca claramente, mostrando para la frecuencia correspondiente una acusada relación cíclica entre ambas variables. Los resultados muestran que se obtiene una coherencia de 14,4 días, que sería el ciclo compartido entre ambas series temporales.

Figura XV2.4. Gráfico de coherencia entre el ángulo que forman el sol y la luna, y el ciclo mareal.

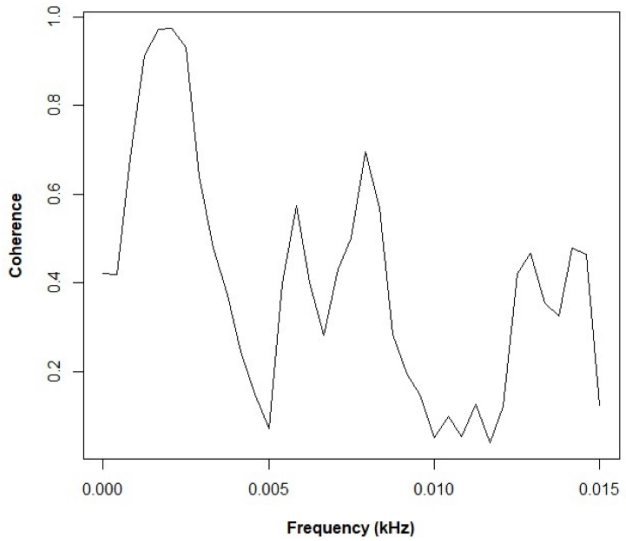
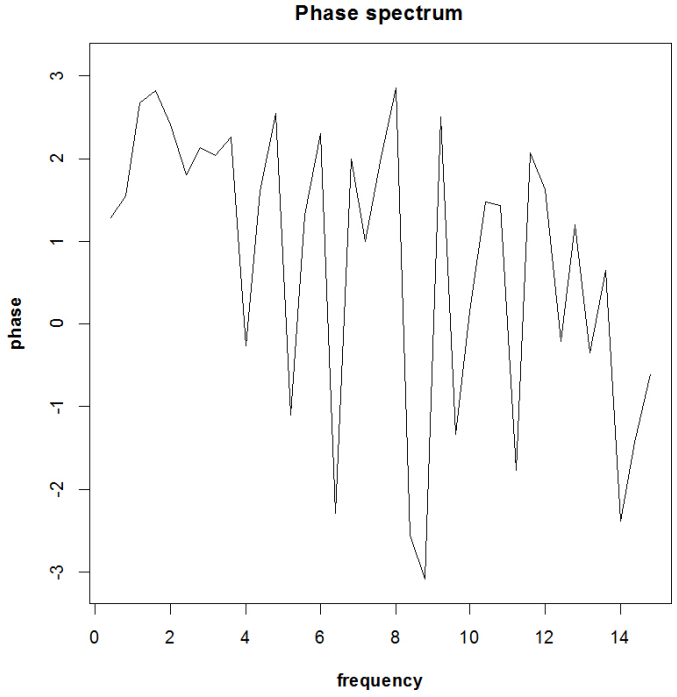


Figura XV2.5. Diagrama de fase entre el ángulo que forman el sol y la luna, y el ciclo mareal.



Aún existiendo un ciclo coincidente en duración para las dos variables, es posible que exista un desfase o retardo entre ambos. Para averiguarlo construimos el diagrama de fase que se muestra

en la Figura XV2.5. En el diagrama se observan oscilaciones irregulares -de apariencia aleatoria- en torno al cero, con amplitud entre -3 y +3 días, para la mayoría de las frecuencias, pero para la frecuencia coincidente que hemos detectado, por debajo de la frecuencia 0,1, existen varios valores que parecen estabilizarse alrededor de -2.

El desfase que se obtiene (*lag*) es 1,8, lo cual significa que el efecto del ángulo entre el sol y la luna sobre las mareas se produce con un retraso de 1,8 días.

Value

Se representan los siguientes gráficos: 1) evolución de las variables en el tiempo, 2) periodograma de la variable dependiente, 3) gráfico de densidad espectral de la variable dependiente, 4) gráfico de coherencia entre la variable dependiente e independiente y 5) diagrama de fase entre la variable dependiente e independiente. Los dos últimos gráficos se muestran si existe variable independiente.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Sueur, J., Aubin, T., Simonis, C., Brown, E.C., Desjonqueres, C., Gasc, A., LaZerte, S., Lees, J., Lellouch, L., Pavoine, S., Villanueva-Rivera, L.J., Ross, Z. & Witthoft, C.G. (2014) Sound analysis and synthesis. R package version 1.7.6. Available at: http://CRAN.R-project.org/package=seewave.

Zeileis, A., Grothendieck, G., Ryan, J.A. & Andrews, F. (2014) Sound analysis and synthesis. R package version 1.7-11. Available at: http://CRAN.R-project.org/package=zoo.

Examples

```
## Not run:

data(ZXV2)

XV2(data=ZXV2, varY="Tidal.range", varX="Angle.Sun.Moon", frequency=30, start=c(1,1))

## End(Not run)
```

XV3 *CAPACIDAD DE PREDICCION DE UNA REGRESION LOGISTICA BINOMIAL*

Description

Permite evaluar la capacidad de predicción de una regresión logística binomial.

Usage

```
XV3(data, cat, var, gruporesp, start=NULL, nsubsets=20, ResetPAR=TRUE, PAR=NULL,
PLOT=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1, COLABLINE="black", COLOR="black", PCH=15,
file1="Test.csv", file2="Coefficients.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`cat`	Variable dependiente.
`var`	Variables independientes.
`gruporesp`	Categoría que será referencia en la variable dependiente.
`start`	Valor inicial de datos que se cogen aleatoriamente.
`nsubsets`	Número de veces que se eligen submuestras aleatorias.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`PLOT`	Se accede a la función plot.default que permite modificar muchos aspectos diferentes del gráfico plot.
`YLAB`	Leyenda del eje Y.
`XLAB`	Leyenda del eje X.
`CEXPCH`	Tamaño de los símbolos del gráfico.
`COLABLINE`	Color de la línea del modelo de regresión.
`COLOR`	Color de los símbolos.
`PCH`	Tipo de símbolo.
`file1`	ARCHIVO CSV. Nombre del archivo con el porcentaje de aciertos de la regresión para cada una de las submuestras de datos escogidos aleatoriamente.
`file2`	ARCHIVOS CSV. Nombre del archivo con los coeficientes de las regresiones de acumulación.
`na`	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

Una vez que se obtiene un modelo de regresión logística, que en principio puede ser válido para poder predecir a que categoría pertenece un caso utilizando una serie de variables cualitativas y cuantitativas, la idea es poder evaluar la capacidad de predicción del modelo.

El algoritmo consiste en ir cogiendo un número de muestras aleatorias del conjunto de datos, elaborar el modelo de regresión logística con esos datos y estimar el porcentaje de aciertos considerando el conjunto total de datos.

Por último, se aplica las cuatro regresiones de acumulación descritas en la función XI2, para relacionar el número de datos escogidos aleatoriamente en la submuestra y el porcentaje de aciertos.

Una regresión tendrá una gran capacidad de predicción, si se consigue un porcentaje de acierto cercano al 100% utilizando la mitad de los datos.

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors y la función nls del paquete base stats para estimar los modelos, de regresión de acumulación.

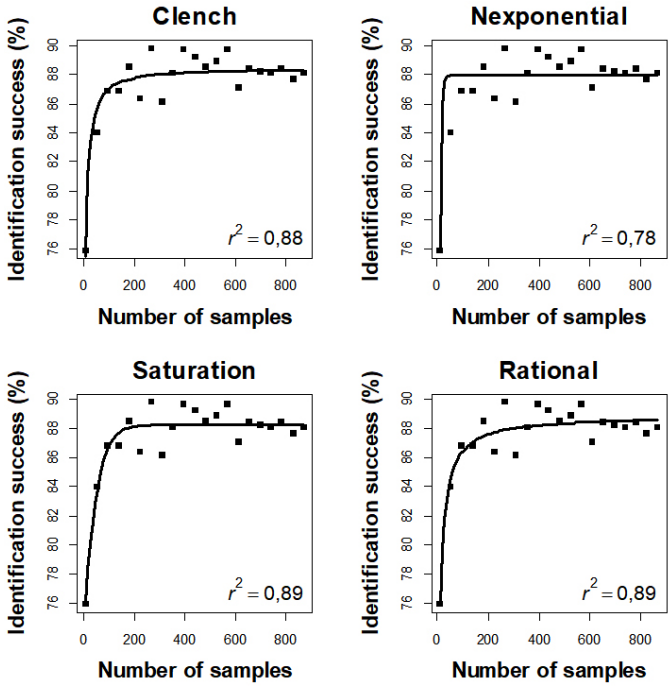
Para estimar la regresión logística binomial se usa la función glm del paquete stats. Para estimar el modelo por pasos se usó la función stepAIC del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2014).

EJEMPLO

El estudio consistió en examinar pacientes a los que se les ha diagnosticado un tipo de cáncer, así como una muestra de personas que no lo tienen. El objetivo es determinar si variables como el sexo, edad, estado civil (soltero, casado y separado) y presencia/ausencia de sangre en la orina, pueden ser indicadoras de la presencia del cáncer.

Al ser muestras que se escogen aleatoriamente, cada vez que se ejecuta la función, los modelos de regresión logística que se estiman pueden variar y, por tanto, también las curvas de acumulación.

En la gráfica se observa que los mejores ajustes los dan las curvas Racional, Clench y Saturación. Cuando se utiliza la mitad de los datos para elaborar el modelo de regresión logística, ya se alcanza la asíntota y el modelo predice alrededor del 88% de los casos correctamente. Por lo tanto, se ha podido identificar claramente la potencia de predicción del modelo.



Value

Se obtiene un archivo TXT, que se llama "Output. TXT" con las ecuaciones de las cuatro regresiones de acumulación. Un archivo CSV con los coeficientes de las regresiones de acumulación. Un gráfico con las cuatro regresiones de acumulación relacionando el número de datos escogidos

aleatoriamente y el porcentaje de casos que se identifican correctamente. También se obtiene un archivo CSV, que se llama "Model predictions.CSV" con las predicciones para cada caso, resultantes de aplicar la regresión logística a la última submuestra aleatoria, que lógicamente tiene el mayor número de datos. Por último, en la ventana de replies se muestra el porcentaje de aciertos que predice cada regresión de acumulación en la asíntota.

References

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI7)

XV3(data=ZXI7, cat="Cancer" , var=c("Age", "Gender", "Status", "Blood"),
gruporesp="Si", start=10)

## End(Not run)
```

XV4 *CAPACIDAD DE PREDICCION DE UNA REGRESION LOGISTICA MULTINOMIAL*

Description

Permite evaluar la capacidad de predicción de una regresión logística multinomial.

Usage

```
XV4(data, cat, var, start=NULL, nsubsets=20, ResetPAR=TRUE, PAR=NULL, PLOT=NULL,
YLAB=NULL, XLAB=NULL, CEXPCH=1, COLABLINE="black", COLOR="black", PCH=15,
file1="Test.csv", file2="Coefficients.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
cat	Variable dependiente.
var	Variables independientes.
start	Valor inicial de datos que se cogen aleatoriamente.

nsubsets	Número de veces que se eligen submuestras aleatorias.
ResetPAR	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
PAR	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
PLOT	Se accede a la función plot.default que permite modificar muchos aspectos diferentes del gráfico plot.
YLAB	Leyenda del eje Y.
XLAB	Leyenda del eje X.
CEXPCH	Tamaño de los símbolos del gráfico.
COLABLINE	Color de la línea del modelo de regresión.
COLOR	Color de los símbolos.
PCH	Tipo de símbolo.
file1	ARCHIVO CSV. Nombre del archivo con el porcentaje de aciertos de la regresión para cada una de las submuestras de datos escogidos aleatoriamente.
file2	ARCHIVOS CSV. Nombre del archivo con los coeficientes de las regresiones de acumulación.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

Como se menciona en la misma sección de la función XV3.

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2013) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors y la función nls del paquete base stats para estimar los modelos, de regresión de acumulación.

Para estimar la regresión logística binomial se usa la función glm del paquete stats. Para estimar el modelo por pasos se usó la función stepAIC del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2014).

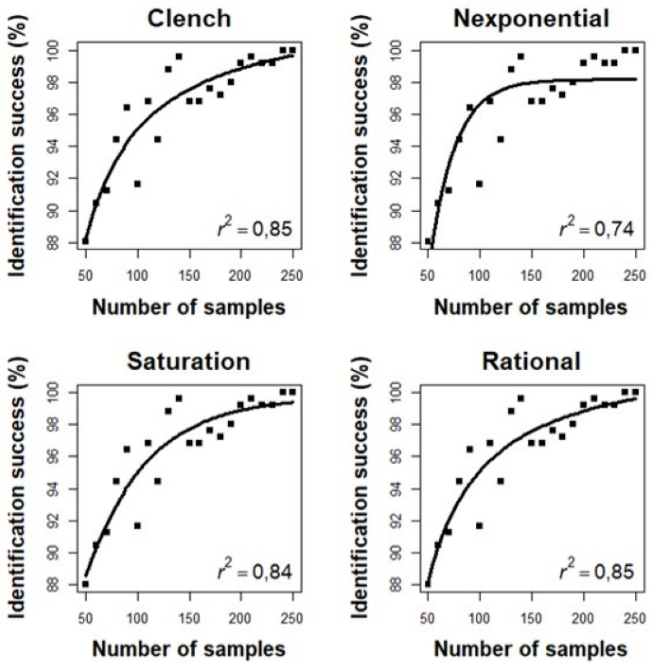
EJEMPLO

Los datos son medidas corporales de varias especies de peces (variables continuas), así como el tipo de boca que tienen (variable categórica). El objetivo es determinar si es posible identificar a que familia pertenece un individuo en función de la biometría.

Al ser muestras que se escogen aleatoriamente, cada vez que se ejecuta la función, los modelos de regresión logística que se estiman pueden variar para cada submuestra de datos y, por tanto, también las curvas de acumulación.

En la gráfica se observa que los mejores ajustes los dan las curvas Clench y Saturación. Cuando se utiliza la mitad de los datos para elaborar el modelo de regresión logística, se obtiene un valor

alrededor del 97% y, según las curvas de acumulación, la asíntota es alrededor del 100%. Por tanto, es posible que con más datos se pueda demostrar que con la mistad de los datos, el modelo ya es capazar de obtener una identificación correcta alrededor del 100% para todo el conjunto de datos.



Value

Como se menciona en la misma sección de la función XV3.

References

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:
data(ZXI9)

XV4(data = ZXI9, cat="Familia", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Boca"), start=50)
## End(Not run)
```

XV5 *ESTANDARIZACIÓN DE DATOS*

Description

Se estandarizan variables a cualquier rango que defina el usuario (por defecto entre 0 y 1) o se puede simplemente estandarizar todas las variables en función de una de ellas.

Usage

```
XV5(data, var, varRef=NULL, varCode=NULL, min=0, max=1, file="Standardized.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Archivo de datos.
`var`	Variables a estandarizar.
`varRef`	Si se especifica una variable, todas se estandarizan en función de ella, es decir, todas se dividen por esa variable.
`varCode`	Opcionalmente se puede seleccionar variables de la matriz original, que no se estandarizan, y que se exportarán en el archivo de salida con las variables estandarizadas. Por ejemplo, permite seleccionar variables que son códigos de las filas.
`min, max`	Valores mínimo y máximo en el rango a estandarizar.
`file`	ARCHIVO CSV. Nombre del archivo con las variables estandarizadas.
`na`	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

EJEMPLO

Los datos corresponden a parámetros socio-económicos de 57 países de Europa, África y América. Las variables usadas fueron la esperanza de vida masculina y femenina al nacer (en años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el producto interior bruto per cápita (en miles de dólares anuales) y la tasa de alfabetización de hombres y mujeres (en porcentaje) en el año 2000. Los datos fueron obtenidos del Banco Mundial (http://datos.bancomundial.org/). Se estandarizan todas las variables al rango por defecto de 0 a 1 y se añaden los códigos del continente y el país.

Value

Se exporta un archivo CSV con las variables estandarizadas y con las variables incluidas en el argumento *varCode* si se incluyó alguna variable en este argumento.

Examples

```
## Not run:

data(ZXIII2)

XV5(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap",
"Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF"),
varCode=c("Continent", "Country"))

## End(Not run)
```

XV6 *DENDROGRAMA A UN COMPONENTES PRINCIPALES*

Description

Se aplica un Dendrograma sobre un análisis de Componentes Principales.

Usage

```
XV6(data, var, labels, cat=NULL, por=80, k=NULL, pthreshold=0.05,
ellipse=FALSE, convex=FALSE, dim=c(1,2), size=c(1,5), showCluster=TRUE,
VIF=FALSE, threshold=10, method="overlap", minimum=TRUE, ResetPAR=TRUE,
PAR=NULL, PCA=NULL, SCATTERPLOT=NULL, HCLUST=NULL, CLUSTER=NULL, BOXPLOT=NULL,
mfrowBOXPLOT=NULL, LabelCat=NULL, COLOR=NULL, COLORC=NULL, COLORB=NULL,
PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL, ylabBOXPLOT=NULL,
LEGEND=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL, arrows=TRUE,
larrow=0.7, colArrows="black", quadratic=FALSE, file1="Output.txt",
file2="Cat loadings.csv", file3="Descriptive statistics of clusters.csv",
file4="Original data and cluster number.csv", file5="Var loadings-Linear.csv",
file6="Cat loadings-Linear.csv", file7="Table cross-validation-Linear.csv",
file8="Cases cross-validation-Linear.csv", file9="Table cross-validation-Quadratic.csv",
file10="Cases cross-validation-Quadratic.csv", file11="Plots VARSEDIG.pdf",
file12="U Mann-Whitney test.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis.
labels	Variable que permite mostrar una etiqueta para cada caso.
cat	Opcionalmente se puede especificar una variable para mostrar una agrupación en el gráfico de Componentes Principales.
por	Umbral de corte especificando el porcentaje de varianza acumulada, para determinar cuantos ejes se seleccionan del Componentes Principales. Por defecto es 80%, que significa que se seleccionan los ejes hasta alcanzar un porcentaje acumulado de varianza del 80%.

`k` Número de clústeres en que se divide el Dendrograma. Si es NULL el algoritmo estima automáticamente el número máximo de clústeres en que se puede dividir el Dendrograma, cuyos grupos sean estadísticamente diferentes en al menos una variable según el test U Mann-Whitney.

`pthreshold` Probabilidad de corte del test U Mann-Whitney.

`ellipse` Si es TRUE, se calculan las elipses de los niveles de significación en el gráfico de Componentes Principales al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *cat*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*. Si es TRUE, también se calculan las elipses de los clústeres en el Discriminante y en el gráfico de coordenadas polares del algoritmo VARSEDIG.

`convex` Si es TRUE, la envolvente convexa se calcula para cada categoría en el gráfico de Componentes Principales, pero solamente si se ha seleccionado alguna variable en el argumento *cat*. Si es TRUE, se calcula la envolvente convexa de los clústeres en el Discriminante y en el gráfico de coordenadas polares del algoritmo VARSEDIG.

`dim` Vector con dos valores indicando los ejes que se representan en el gráfico de Componentes Principales.

`size` Rango de tamaño de las burbujas. Dos valores: mínimo y máximo tamaño.

`showCluster` Si es TRUE se muestra en el Dendrograma el número de cada cúster.

`VIF` Si es TRUE se utiliza el factor de inflación de la varianza (VIF) para seleccionar las variables altamente correlacionadas y, por tanto, las variables no correlacionadas son excluidas del análisis de Componentes Principales.

`threshold` Valor de corte para el VIF.

`method` Hay tres métodos para priorizar las variables en el algoritmo VARSEDIG de acuerdo a su capacidad de discriminación. Si el método es "overlap", se estima la curva de densidad para cada variable y el solapamiento del área bajo la curva entre cada dos clústeres para todas las variables. Aquellas variables con bajo solapamiento deberían tener una mayor capacidad de discriminación y, por tanto, las variables se ordenan de menor a mayor solapamiento; en otras palabras, de mayor a menor capacidad de discriminación. Si el método es "Monte-Carlo", se realiza un test de Monte-Carlo comparando todos los valores del grupo 1 con el grupo 2, y todos los valores de grupo 2 con el 1. Las variables se priorizan desde la variable con el valor medio de probabilidad menor (mayor capacidad de discriminación) a la variable con el valor medio de probabilidad más alto (menor capacidad de discriminación). Si el método es "logistic regression", entonces se calcula una regresión logística binomial, y solamente las variables que son significativas son seleccionadas para posteriores análisis con la regresión por pasos usando el *Akaike Information Criterion* (AIC).

`minimum` Si es TRUE, el algoritmo VARSEDIG está diseñado para encontrar una discriminación significativa entre ambos grupos con el mínimo número posible de variables significativas. Por lo tanto, solo se seleccionan las variables con mayor capacidad de discriminación. Si es FALSE, el algoritmo selecciona todas las

	variables significativas, y no solo aquellas con mayor capacidad de discriminación. Este argumento solo es válido con los métodos "Monte-Carlo" y "overlap" y es útil en esos casos que la discriminación entre los grupos es difícil y requiere incluir tantas variables como sea posible.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`PCA`	Se accede a la función prcomp del paquete stats.
`SCATTERPLOT`	Se accede a la función scatterplot del paquete car.
`HCLUST`	Se accede a la función hclust del paquete stats.
`CLUSTER`	Se accede a la función que permite modificar la representación gráfica del dendrograma.
`BOXPLOT`	Permite especificar las características del boxplot.
`mfrowBOXPLOT`	Permite especificar el panel del boxplot. Es un vector con dos números, por ejemplo c(2,5) que significa que los boxplots se ponen en 2 filas y 5 columnas.
`LabelCat`	Permite especificar un vector con los nombres de los clústeres en los boxplots. Deben ser tantos como clústeres.
`COLOR`	Permite modificar los colores del gráfico en el Componentes Principales, pero deben ser tantos como grupos diferentes tenga la variable *cat*.
`COLORC`	Permite modificar los colores de los clústeres en el Dendrograma, pero deben ser tantos como clústeres.
`COLORB`	Permite modificar los colores de los clústeres en los boxplots, pero deben ser tantos clústeres.
`PCH`	Vector con los símbolos del gráfico de Componentes Principales, que deben ser tantos como grupos diferentes tenga la variable *cat*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
`XLIM, YLIM`	Vectores con los límites de los ejes *X* e *Y* del Componentes Principales.
`XLAB, YLAB`	Leyendas de los ejes *X* e *Y* del Componentes Principales.
`ylabBOXPLOT`	Se puede especificar un vector con las leyendas de los ejes *Y* de los boxplots. Deben ser tantos como el número de variables.
`LEGEND`	Permite incorporar o modificar la leyenda del Componentes Principales.
`MTEXT`	Permite añadir textos en los márgenes del gráfico leyenda del Componentes Principales.
`TEXTvar`	Se accede a la función que permite modificar las etiquetas de las variables del gráfico de Componentes Principales.
`TEXTlabels`	Se accede a la función que permite modificar las etiquetas de los casos del gráfico de Componentes Principales.
`arrows`	Si es TRUE se muestran las flechas en el gráfico de Componentes Principales.
`larrow`	Permite modificar la longitud de las flechas.
`colArrows`	Colores de las flechas.

quadratic	Si es TRUE se realiza un Análisis Discriminante Cuadrático, además del Análisis Discriminante Lineal.
file1	ARCHIVO TXT. Nombre del archivo de salida con los resultados.
file2	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de los casos del gráfico de Componentes Principales.
file3	ARCHIVO CSV. Nombre del archivo de salida con la estadística descriptiva de cada variable para cada uno de los clústeres obtenidos en el Dendrograma.
file4	ARCHIVO CSV. Nombre del archivo de salida con los datos originales de las variables y el clúster al que pertenece cada caso.
file5	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las variables en el gráfico del Análisis Discriminante Lineal.
file6	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las categorías en el gráfico del Análisis Discriminante Lineal.
file7	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Lineal.
file8	ARCHIVO CSV. Nombre del archivo de salida con el grupo al que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Lineal.
file9	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Cuadrático.
file10	ARCHIVO CSV. Nombre del archivo de salida con el grupo al que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Cuadrático.
file11	ARCHIVO PDF. Nombre del archivo de salida con los gráficos obtenidos del algoritmo VARSEDIG.
file12	ARCHIVO CSV. Nombre del archivo de salida con las probabilidades obtenidas de comparar todas las variables entre todos los clústeres con el test U Mann-Whitney.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

El objetivo de este análisis es determinar que grupos estadísticamente diferentes se forman al aplicar un Componentes Principales.

El primer eje en un Componentes Principales es la combinación lineal de las variables originales que tiene máxima varianza. La segunda componente es la combinación lineal de las variables originales con máxima varianza con la condición añadida de que sea independiente de la primera (ortogonal), y así sucesivamente se pueden obtener todas las componentes principales, que siendo independientes entre sí contienen información distinta, ya que la independencia o ausencia de correlación significa que las nuevas variables o componentes no comparten información común. Cada

componente principal explica por lo tanto la máxima variabilidad residual posible (la que no han explicado ya las anteriores). Por lo tanto, en un Componentes Principales los casos se diferencian en función de las variables que tienen mayor variabilidad. La idea del análisis es determinar si se forman grupos estadísticamente diferentes asociados a la variabilidad observada en las variables.

Este análisis puede ser útil para encontrar grupos diferentes cuando realmente no se saben cuales son. Por ejemplo, encontrar diferentes especies usando variables morfométricas, sin saber realmente cuantas potenciales especies hay y a que especie pertenece cada individuo. Sin embargo, es importante resaltar que solo se detectaran grupos diferentes si las variables que más variabilidad tienen dan lugar a que se diferencien grupos. Es muy posible que una variable no presente una gran variabilidad, pero sea importante para discriminar entre grupos. Este tipo de diferenciaciones basadas en variables que no tienen alta varianza, no serían detectadas en este análisis.

Para detectar los grupos potenciales que se forman se aplica un dendrograma a las puntuaciones obtenidas de los ejes que absorben una mayor varianza. Por defecto se escogen los ejes que absorben un 80% de la variabilidad, pero este valor puede ser modificado por el usuario.

Posteriormente, se realiza un Análisis Discriminante para determinar si se discriminan bien los clústeres que se han generado, es decir, determinar el número de casos correctamente identificados en cada clúster.

A continuación, se realiza un test de la U Mann-Whitney para determinar si existen diferencias significativas en las variables entre los clústeres.

Por último, se aplica el algoritmo de la función VARSEDIG (véase para más detalles Guisande et al. 2016: Guisande, 2018). Con este algoritmo es posible determinar si todos los casos de cada clúster son estadísticamente diferentes de los otros clústeres.

La idea de esta función es buscar el mayor número posible de clústeres con el porcentaje de discriminación más alto. Para ello el usuario debería realizar pruebas, modificando el umbral de corte especificando el porcentaje de varianza acumulada para determinar cuantos ejes se seleccionan del Componentes Principales (por defecto *por=80*) y las variables a incluir, eliminando aquellas que no estén correlacionadas y no sean útiles en el Componentes Principales, así como aquellas que tengan poco poder de discriminación en el Análisis Discriminante.

FUNCIONES

El Análisis de Componentes Principales se realizó con la función prcomp del paquete base stats. Para el cáculo del VIF se usó la función vif del paquete usdm (Naimi, 2013; Naimi et al., 2014). Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2018). Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014). La envolvente convexa se estima con la función chull del paquete grDevices. Las correlaciones entre variables se calcularon con la función rcor.test del paquete ltm (Rizopoulos, 2018). El test KMO se realiza con la función KMO del paquete psych (Revelle, 2018). El test de esfericidad de Bartlett se realizó con la función bart_spher del paquete REdaS (Maier, 2015). El test de la U Mann-Whitney se realiza con la función *wilcox.test* del paquete base stats. La comparación entre clústeres con el algoritmo VARSEDIG se realiza con la función VARSEDIM del paquete VARSEDIG (Guisande et al. 2016: Guisande, 2018). El Análisis Discriminante Lineal se realizó con las funciones candisc del paquete candisc (Friendly, 2007; Friendly & Fox, 2017) y lda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2018). El Análisis Discriminante Cuadrático se realizó con la función qda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2018). El gráfico del Análisis Discriminante con una dimensión se realizó con la función plot.cancor del paquete candisc (Friendly, 2007; Friendly & Fox, 2017).

EJEMPLO

El estudio consistió en analizar los parámetros demográficos de 57 países de Europa, África y América. Las variables usadas fueron la esperanza de vida masculina y femenina al nacer (en años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el producto interior bruto per cápita (en miles de dólares anuales) y la tasa de alfabetización de hombres y mujeres (en porcentaje) en el año 2000. Los datos fueron obtenidos del Banco Mundial (http://datos.bancomundial.org/). El objetivo es ver que grupos se forman a partir del Análisis de Componentes Principales.

En primer lugar se utilizan todas las variables *var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF")*, y como etiquetas *labels="Country"*. Con fines solo de presentación gráfica en el Componentes Principales se utiliza el continente como categoría *cat="Continent"*. Es importante resaltar que la categoría no se utiliza para ningún análisis estadístico y simplemente se usa para agrupar los casos con elipses o con la envolvente convexa en el gráfico de Componentes Principales.

Siempre es conveniente realizar el análisis eliminando las variables que no están correlacionadas, para lo que debemos especificar *VIF=TRUE*. Sin embargo, no se ha hecho en el ejemplo, porque el KMO de todas las variables fue significativo. Por ello, no se muestran los valores de VIF en el archivo TXT de salida.

El primer resultado que se obtiene es el test KMO, que nos indica si las variables son adecuadas para el Componentes Principales. El valor debe ser mayor de 0,5. Por tanto, todas las variables que no tengan un valor mayor a 0,5, podrían ser eliminadas del análisis. En el caso de que el valor sea exactamente 0,5 es que no fue posible estimar el KMO.

```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf   PIB_cap Birthrate Mortality Fertility LiteracyM LiteracyF
     0.83      0.89      0.99      0.85      0.85      0.77      0.88      0.83      0.84
```

```
Bartlett's Test of Sphericity

Call: REdaS::bart_spher(x = datos1)

     X2 = 1084.808

     df = 36

p-value < 2.22e-16
```

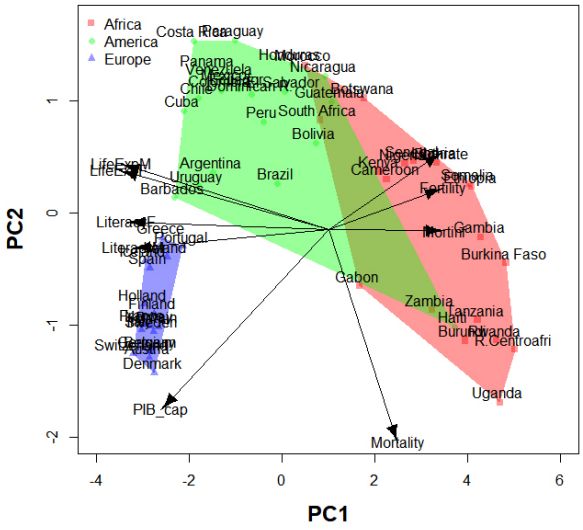
El segundo estadístico que parece es el test de esfericidad de Bartlett, el cual está basado en el determinante de la matriz de correlaciones y contrasta la hipótesis de "esfericidad" o ausencia de correlación con un único test. Un valor *p* del contraste menor que el nivel de significación permite rechazar la hipótesis y concluir que existe correlación. Por tanto, para que el Componentes Principales sea válido, la probabilidad deber ser menor de 0,05, como es en este caso.

Como podemos observar en la matriz de correlación, casi todos los valores *p* son menores que 0,05, lo que indica que ninguno de los coeficientes de correlación parece ser nulo. Por tanto, todas las variables están correlacionadas.

```
"Correlation matrix"

          LifeExpF LifeExpM Mortinf PIB_cap Birthrate Mortality Fertility
LifeExpF   *****    0.993   -0.962   0.673  -0.921     -0.774    -0.926
LifeExpM  <0.001    *****   -0.958   0.659  -0.900     -0.778    -0.904
Mortinf   <0.001   <0.001    *****  -0.685   0.912      0.704     0.915
PIB_cap   <0.001   <0.001   <0.001   *****  -0.795     -0.152    -0.716
Birthrate <0.001   <0.001   <0.001  <0.001   *****      0.538     0.978
Mortality <0.001   <0.001   <0.001   0.260  <0.001      *****     0.611
Fertility <0.001   <0.001   <0.001  <0.001  <0.001     <0.001     *****
LiteracyM <0.001   <0.001   <0.001  <0.001  <0.001     <0.001    <0.001
LiteracyF <0.001   <0.001   <0.001  <0.001  <0.001     <0.001    <0.001
          LiteracyM LiteracyF
LifeExpF    0.802     0.859
LifeExpM    0.794     0.849
Mortinf    -0.850    -0.888
PIB_cap     0.620     0.606
Birthrate  -0.836    -0.867
Mortality  -0.508    -0.587
Fertility  -0.846    -0.887
LiteracyM   *****     0.975
LiteracyF  <0.001     *****
```

En la Figura XV6.1 se observa que en el primer eje se diferencian perfectamente los países de Europa, América y África. Los de Europa se caracterizan por alta esperanza de vida y alta alfabetización, tanto en hombres como en mujeres. Los de África por su alta mortalidad infantil y alta tasa de natalidad.

Figura XV6.1. Análisis de Componentes Principales.

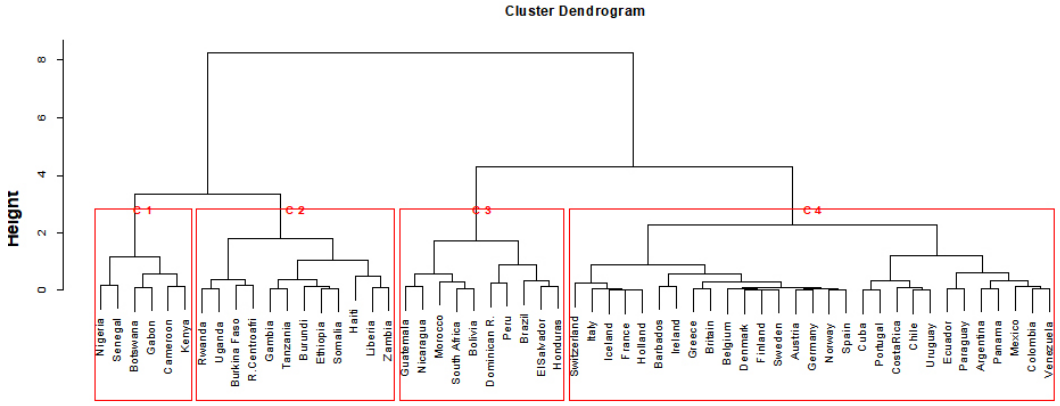


El primer eje explica el 81,86%, el segundo el 10,27% y el tercero el 4,7% de la variación observada. Los dos primeros ejes explican el 92,14% de la varianza. Como se dejó el valor por defecto de *por=80*, solamente se selecciona el primer eje de Componentes Principales.

```
"Summary Multivariate Analysis"

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.7144 0.9615 0.6552 0.37512 0.26068 0.18871 0.13331
Proportion of Variance 0.8186 0.1027 0.0477 0.01563 0.00755 0.00396 0.00197
Cumulative Proportion  0.8186 0.9214 0.9691 0.98471 0.99226 0.99621 0.99819
                          PC8     PC9
Standard deviation     0.10906 0.06650
Proportion of Variance 0.00132 0.00049
Cumulative Proportion  0.99951 1.00000
```

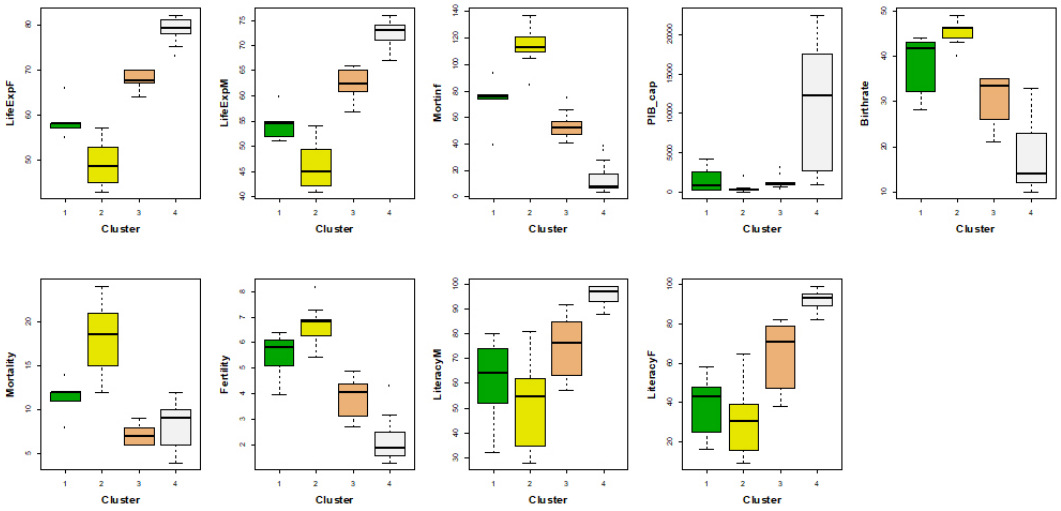
En la Figura XV6.2 se muestra el Dendrograma donde se agrupan 4 clústeres, ya que se especificó el argumento *k=4*.

Figura XV6.2. Dendrograma con las puntuaciones de los ejes seleccionados del Componentes Principales.



En la Figura XV6.3 se observan las diferencias entre clústeres para cada una de las variables. Es clara la diferencia en el PIB del clúster 4, que agrupa a la mayoría de países de Europa. También destacar los valores altos de mortalidad y mortalidad infantil del grupo de países del clúster 2.

Figura XV6.3. Boxplot obtenido para cada una de las variables con valores promediados para cada clúster.



El Análisis Discriminante muestra que es posible discriminar correctamente el 100% de los casos por validación cruzada con el método Lineal. El primer eje discriminante absorbe la mayor parte de la variabilidad y discrimina bien entre los 4 clústeres (Figura XV6.4). Todas las variables parecen ser importantes para la discriminación ya que ninguna de las flechas es pequeña. La Figura XV6.5 muestra los dos primeros ejes discriminantes y se ven las diferencias entre los 4 clústeres.

Figura XV6.4. Eje I del Análisis Discriminante

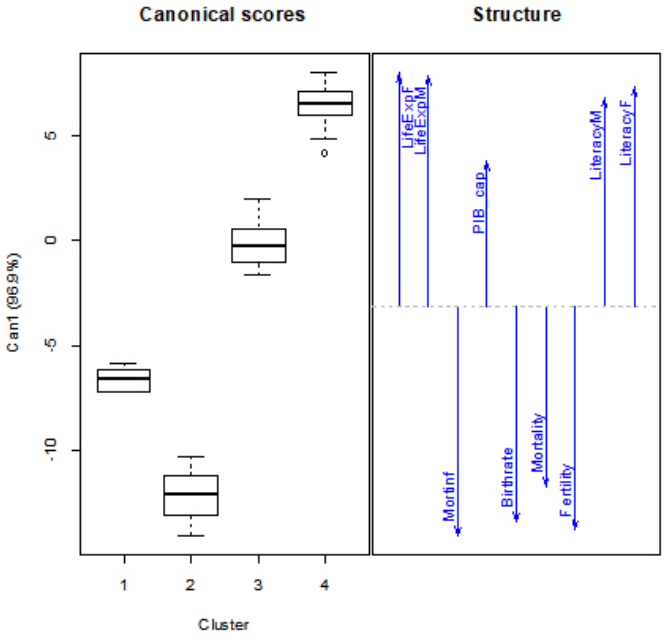
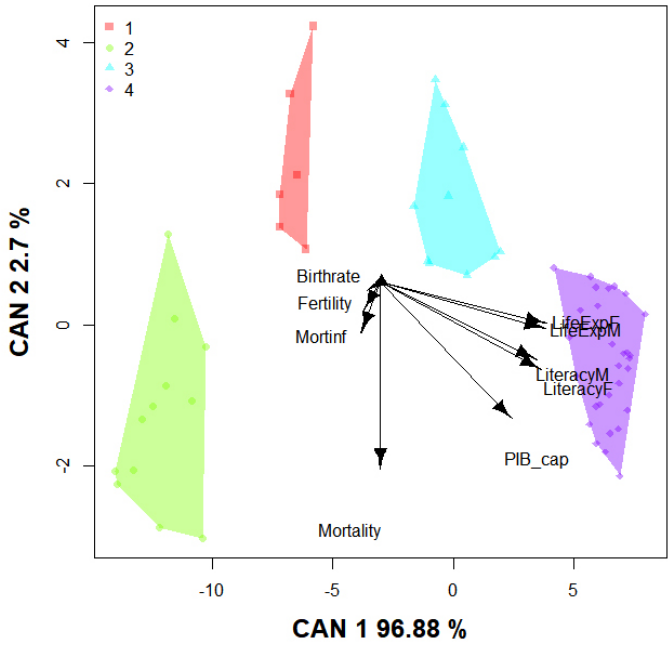


Figura XV6.5. Ejes I y II del Análisis Discriminante

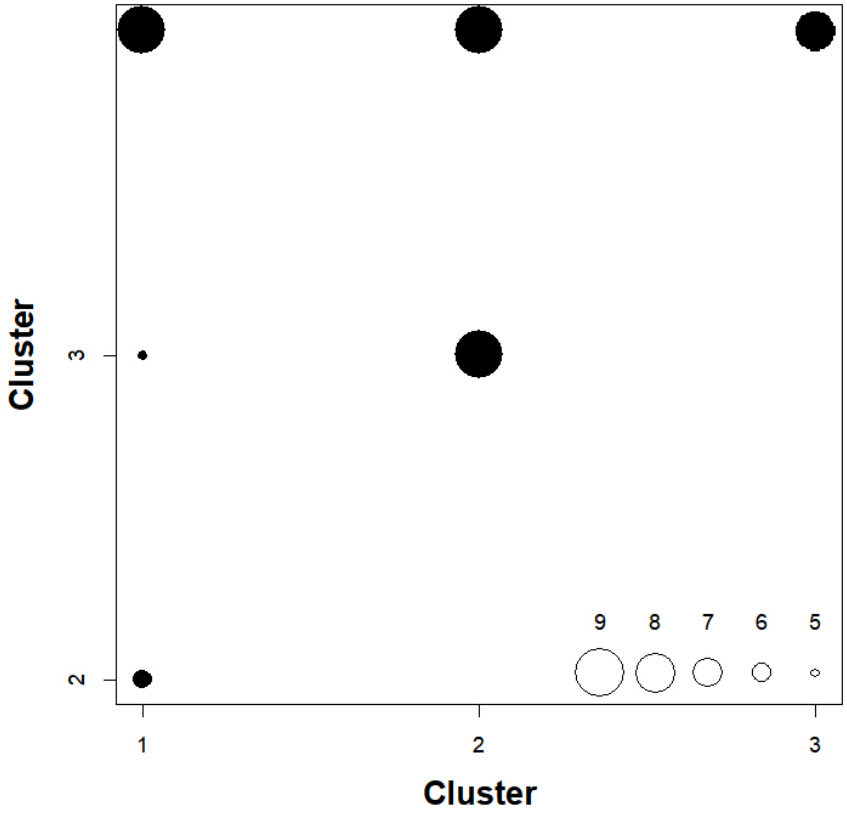


La siguiente prueba para determinar si los clústeres son estadísticamente diferentes fue la comparación de las variables entre los clústeres. En Figura XV6.6 se muestran los resultados del test U Mann-Whitney.

Para que los clústeres sean diferentes es necesario que exista al menos una variable estadísticamente diferente al comparar cada clúster con todos los demás. En la gráfica se oberva que en la comparación entre todos los clústeres siempre hay un punto, es decir, siempre hay al menos una variable que es diferente. En realidad, entre el clúster 1 y el clúster 3 fue donde se observó un menor número de variables estadísticamente diferentes, un total de 5.

Por lo tanto, de la comparación de las variables entre clústeres con el test U Mann-Whitney, se concluye que los clústeres son estadísticamente diferentes entre si.

Figura XV6.6. Gráfico donde las burbujas representan el número de variables.
que son estadísticamente diferentes (p<=0,05) entre clústeres.



Por tanto, de acuerdo al Análisis Discriminante y a las pruebas realizadas con el test de la U Mann-Whitney, los clústeres son estadísticamente diferentes entre sí, pero el algoritmo VARSEDIG mostró que no todos los clústeres son estadísticamente diferentes. Sin embargo, es muy importante resaltar que el algoritmo VARSEDIG considera dos grupos estadísticamente diferentes si el caso que más se parece de cada grupo al otro, es estadísticamente diferente usando el test de Monte-Carlo. El test de Monte-Carlo necesita un número de casos grande en cada grupo, para que realmente se pueda detectar diferencias significativas. Es decir, es posible que, como se mostró en la comparación del clúster 1 con el 2, los casos de ambos grupos que más se parecen no están dentro de la nube de puntos del otro grupo, pero debido al bajo número de casos de los grupos, no es posible determinar que la diferencia no sea debido al azar.

Value

Se obtiene:

1. Un archivo TXT con el VIF (si el argumento *VIF=TRUE*), las correlaciones entre variables, el factor de adecuación Kaiser-Meyer-Olkin (KMO), el test de esfericidad de Bartlett y los resultados del Análisis de Componentes Principales. El archivo se llama por defecto "Output.TXT".

2. Un archivo CSV con las coordenadas para cada caso de los ejes de Componentes Principales. El archivo se llama por defecto "Cat loadings.CSV".

3. Un archivo CSV con la estadística descriptiva de cada variable para cada uno de los clústeres

obtenidos en el Dendrograma. El archivo se llama por defecto "Descriptive statistics of clusters.CSV".

4. Un archivo CSV con los datos originales de las variables y el clúster al que pertenece cada caso. El archivo se llama por defecto "Original data and cluster number.CSV".

5. Un archivo CSV con las coordenadas de las variables en el gráfico del Análisis Discriminante Lineal. El archivo se llama por defecto "Var loadings-Linear.csv".

6. Un archivo CSV con las coordenadas de las categorías en el gráfico del Análisis Discriminante Lineal. El archivo se llama por defecto "Cat loadings-Linear.csv".

7. Un archivo CSV con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Lineal. El archivo se llama por defecto "Table cross-validation-Linear.csv".

8. Un archivo CSV con el grupo al que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Lineal. El archivo se llama por defecto "Cases cross-validation-Linear.csv".

9. Un archivo CSV con la tabla de predicciones usando la validación cruzada del Análisis Discriminante Cuadrático. El archivo se llama por defecto "Table cross-validation-Quadratic.csv".

10. Un archivo CSV con el grupo al que pertenece cada caso y la predicción del Análisis Discriminante usando la validación cruzada del Análisis Discriminante Cuadrático. El archivo se llama por defecto "Cases cross-validation-Quadratic.csv".

11. Un archivo CSV con las probabilidades obtenidas de comparar todas las variables entre todos los clústeres con el test U Mann-Whitney.

12. Un archivo PDF con los gráficos obtenidos del algoritmo VARSEDIG.

13. Un gráfico del Componentes Principales.

14. El Dendrograma agrupando por clústeres en función de las puntuaciones del Componentes Principales.

15. Un panel gráfico con un boxplot por cada variable comparando los valores de esas variables entre cada uno de los clústeres obtenidos en el Dendrograma.

16. Un gráfico del Análisis Discriminante mostrando la influencia de las variables en el eje discriminante I, diferenciando los diferentes clústeres.

17. Un gráfico del Análisis Discriminante mostrando las puntuaciones de los ejes discriminante I y II, diferenciando los diferentes clústeres.

18. Un gráfico de burbujas con el número de variables que son estadísticamente diferentes entre clústeres.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2017) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. Available at: http://CRAN.R-project.org/package=candisc.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162: 571-580.

Guisande, C. (2018) An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy. R package version 1.8. Available at: http://CRAN.R-project.org/package=VARSEDIG.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Maier, M.J. (2015) Companion Package to the Book 'R: Einführung durch angewandte Statistik. R package version 0.9.3. Available at: http://CRAN.R-project.org/package=REdaS.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Revelle,W. (2018) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.8.4. Available at: http://CRAN.R-project.org/package=psych.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2018) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-50. Available at: http://CRAN.R-project.org/package=MASS.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17: 1-25.

Rizopoulos, D. (2018) Latent Trait Models under IRT. R package version 1.1-1. Available at: http://CRAN.R-project.org/package=ltm.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXIII2)

XV6(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent", labels="Country",
k=4, convex=TRUE)


## End(Not run)
```

XV7 *DETECCION DE VALORES ATIPICOS EN VARIABLES INDEPENDIENTES*

Description

Se detectan valores atípicos y, si el usuario lo elige, se borran automáticamente.

Usage

```
XV7(data, var, cat=NULL, varCode=NULL, remove=FALSE, Nan=FALSE,
quant1=0.05, quant2=0.95, file="Outliers.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
var	Variables a detectar los valores atípicos.
cat	Si se especifica una variable, los valores atípicos se calculan considerando cada grupo de esa variable. Por ejemplo, si la variable incluye diferentes especies, en vez de buscar los atípicos en función de todos los datos de la variable, se buscan considerando los datos de cada especie por separado.
varCode	Opcionalmente se puede seleccionar variables de la matriz original, que se exportarán en el archivo de salida con las variables en las que se detectan valores atípicos. Por ejemplo, permite seleccionar variables que son códigos de las filas.
remove	Si es TRUE, se eliminan las filas que tengan alguna variable con valor atípico. Si es FALSE, los atípicos se indican con el valor -9999.
Nan	Si es TRUE los valores atípicos se muestran como NA en vez de -9999.
quant1	Cuantil del extremo inferior para la eliminación de los datos atípicos.
quant2	Cuantil del extremo superior para la eliminación de los datos atípicos.
file	ARCHIVO CSV. Nombre del archivo con las variables mostrando los valores atípicos o con las variables con los valores atípicos eliminados.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

Los valores atípicos se consideran aquellos que quedan fuera de 1,5 * IQR (rango intercuartílico). IQR es la diferencia entre el valor especificado en los argumentos *quant2* y *quant1*. Por defecto es 0,95 y 0,05, pero por ejemplo en el boxplot es normalmente 0,75 y 0,25.

EJEMPLO

Se detectan valores atípicos en unas variables morfométricas de peces y se ha seleccionado que se eliminen con el argumento *remove=TRUE*.

Value

Se exporta un archivo CSV con las variables mostrando los valores atípicos y con las variables incluidas en el argumento *varCode* si se incluyó alguna variable en este argumento. Si la opción *remove=TRUE* se eliminan las filas si alguna variable tiene un valor atípico.

Examples

```
## Not run:

data(ZX3)

XV7(data=ZX3, var=c("M2", "M3", "M4", "M5", "M6", "M7", "M8"), cat="Species",
varCode=c("Order", "Family", "Genus", "Species"), remove=TRUE)


## End(Not run)
```

XV8 *DENDROGRAMA A UN ANALISIS DE CORRESPONDENCIAS*

Description

Se aplica un Dendrograma sobre un análisis de Correspondencias.

Usage

```
XV8(data, var, labels, cat=NULL, por=80, k=NULL, pthreshold=0.05,
ellipse=FALSE, convex=FALSE, dim=c(1,2), size=c(1,5), showCluster=TRUE,
method="overlap", minimum=TRUE, ResetPAR=TRUE, PAR=NULL, SCATTERPLOT=NULL,
HCLUST=NULL, CLUSTER=NULL, BOXPLOT=NULL, mfrowBOXPLOT=NULL,
LabelCat=NULL, COLOR=NULL, COLORC=NULL, COLORB=NULL,
PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL, ylabBOXPLOT=NULL,
LEGEND=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL, arrows=TRUE,
larrow=0.7, colArrows="black", quadratic=FALSE, file1="Output.txt",
file2="Cat loadings.csv", file3="Descriptive statistics of clusters.csv",
file4="Original data and cluster number.csv", file5="Var loadings-Linear.csv",
file6="Cat loadings-Linear.csv", file7="Table cross-validation-Linear.csv",
file8="Cases cross-validation-Linear.csv", file9="Table cross-validation-Quadratic.csv",
file10="Cases cross-validation-Quadratic.csv", file11="Plots VARSEDIG.pdf",
file12="U Mann-Whitney test.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
var	Variables que se incluyen en el análisis. Si son factores o caracteres se transforman a numéricas y se muestra, en los resultados, una tabla con el valor numérico asignado a cada categoría.
labels	Variable que permite mostrar una etiqueta para cada caso.
cat	Opcionalmente se puede especificar una variable para mostrar una agrupación en el gráfico del análisis de Correspondencias.

`por` Umbral de corte especificando el porcentaje de varianza acumulada, para determinar cuantos ejes se seleccionan del análisis de Correspondencias. Por defecto es 80%, que significa que se seleccionan los ejes hasta alcanzar un porcentaje acumulado de varianza del 80%.

`k` Número de clústeres en que se divide el Dendrograma. Si es NULL el algoritmo estima automáticamente el número máximo de clústeres en que se puede dividir el Dendrograma, cuyos grupos sean estadísticamente diferentes en al menos una variable según el test U Mann-Whitney.

`pthreshold` Probabilidad de corte del test U Mann-Whitney.

`ellipse` Si es TRUE, se calculan las elipses de los niveles de significación en el gráfico del análisis de Correspondencias al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías de la variable *cat*. Estos niveles se pueden modificar entrando en la función scatterplot con el argumento *SCATTERPLOT* y modificando el argumento *levels=c(0.5,0.95)*. Si es TRUE, también se calculan las elipses de los clústeres en el Discriminante y en el gráfico de coordenadas polares del algoritmo VARSEDIG.

`convex` Si es TRUE, la envolvente convexa se calcula para cada categoría en el gráfico de análisis del análisis de Correspondencias, pero solamente si se ha seleccionado alguna variable en el argumento *cat*. Si es TRUE, se calcula la envolvente convexa de los clústeres en el Discriminante y en el gráfico de coordenadas polares del algoritmo VARSEDIG.

`dim` Vector con dos valores indicando los ejes que se representan en el gráfico del análisis de Correspondencias.

`size` Rango de tamaño de las burbujas. Dos valores: mínimo y máximo tamaño.

`showCluster` Si es TRUE se muestra en el Dendrograma el número de cada clúster.

`method` Hay tres métodos para priorizar las variables en el algoritmo VARSEDIG de acuerdo a su capacidad de discriminación. Si el método es "overlap", se estima la curva de densidad para cada variable y el solapamiento del área bajo la curva entre cada dos clústeres para todas las variables. Aquellas variables con bajo solapamiento deberían tener una mayor capacidad de discriminación y, por tanto, las variables se ordenan de menor a mayor solapamiento; en otras palabras, de mayor a menor capacidad de discriminación. Si el método es "Monte-Carlo", se realiza un test de Monte-Carlo comparando todos los valores del grupo 1 con el grupo 2, y todos los valores de grupo 2 con el 1. Las variables se priorizan desde la variable con el valor medio de probabilidad menor (mayor capacidad de discriminación) a la variable con el valor medio de probabilidad más alto (menor capacidad de discriminación). Si el método es "logistic regression", entonces se calcula una regresión logística binomial, y si el argumento *stepwise=TRUE* (por defecto), solamente las variables que son significativas son seleccionadas para posteriores análisis con la regresión por pasos usando el *Akaike Information Criterion* (AIC).

`minimum` Si es TRUE, el algoritmo VARSEDIG está diseñado para encontrar una discriminación significativa entre ambos grupos con el mínimo número posible de variables significativas. Por lo tanto, solo se seleccionan las variables con mayor capacidad de discriminación. Si es FALSE, el algoritmo selecciona todas las

	variables significativas, y no solo aquellas con mayor capacidad de discriminación. Este argumento solo es válido con los métodos "Monte-Carlo" y "overlap" y es útil en esos casos que la discriminación entre los grupos es difícil y requiere incluir tantas variables como sea posible.
`ResetPAR`	Si es FALSE no se ponen las condiciones por defecto de la función PAR y se mantienen las definidas por el usuario en gráficos anteriores.
`PAR`	Se accede a la función PAR que permite modificar muchos aspectos diferentes del gráfico.
`SCATTERPLOT`	Se accede a la función scatterplot del paquete car.
`HCLUST`	Se accede a la función hclust del paquete stats.
`CLUSTER`	Se accede a la función que permite modificar la representación gráfica del dendrograma.
`BOXPLOT`	Permite especificar las características del boxplot.
`mfrowBOXPLOT`	Permite especificar el panel del boxplot. Es un vector con dos números, por ejemplo c(2,5) que significa que los boxplots se ponen en 2 filas y 5 columnas.
`LabelCat`	Permite especificar un vector con los nombres de los clústeres en los boxplots. Deben ser tantos como clústeres.
`COLOR`	Permite modificar los colores del gráfico en el análisis de Correspondencias, pero deben ser tantos como grupos diferentes tenga la variable *cat*.
`COLORC`	Permite modificar los colores de los clústeres en el Dendrograma, pero deben ser tantos como cústeres.
`COLORB`	Permite modificar los colores de los clústeres en los boxplots, pero deben ser tantos como clústeres.
`PCH`	Vector con los símbolos del gráfico de análisis de Correspondencias, que deben ser tantos como grupos diferentes tenga la variable *cat*. Si es NULL se calculan automáticamente empezando por el símbolo 15.
`XLIM, YLIM`	Vectores con los límites de los ejes X e Y del análisis de Correspondencias.
`XLAB, YLAB`	Leyendas de los ejes X e Y del análisis de Correspondencias.
`ylabBOXPLOT`	Se puede especificar un vector con las leyendas de los ejes Y de los boxplots. Deben ser tantos como el número de variables.
`LEGEND`	Permite incorporar o modificar la leyenda del análisis de Correspondencias.
`MTEXT`	Permite añadir textos en los márgenes del gráfico leyenda del análisis de Correspondencias.
`TEXTvar`	Se accede a la función que permite modificar las etiquetas de las variables del gráfico de análisis de Correspondencias.
`TEXTlabels`	Se accede a la función que permite modificar las etiquetas de los casos del gráfico de análisis de Correspondencias.
`arrows`	Si es TRUE se muestran las flechas en el gráfico de análisis de Correspondencias.
`larrow`	Permite modificar la longitud de las flechas.
`colArrows`	Colores de las flechas.

quadratic	Si es TRUE se realiza un análisis Discriminante Cuadrático, además del análisis Discriminante Lineal.
file1	ARCHIVO TXT. Nombre del archivo de salida con los resultados.
file2	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de los casos del gráfico de análisis de Correspondencias.
file3	ARCHIVO CSV. Nombre del archivo de salida con la estadística descriptiva de cada variable para cada uno de los clústeres obtenidos en el Dendrograma.
file4	ARCHIVO CSV. Nombre del archivo de salida con los datos originales de las variables y el clúster al que pertenece cada caso.
file5	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las variables en el gráfico del análisis Discriminante Lineal.
file6	ARCHIVO CSV. Nombre del archivo de salida con las coordenadas de las categorías en el gráfico del análisis Discriminante Lineal.
file7	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del análisis Discriminante Lineal.
file8	ARCHIVO CSV. Nombre del archivo de salida con el grupo al que pertenece cada caso y la predicción del análisis Discriminante usando la validación cruzada del análisis Discriminante Lineal.
file9	ARCHIVO CSV. Nombre del archivo de salida con la tabla de predicciones usando la validación cruzada del análisis Discriminante Cuadrático.
file10	ARCHIVO CSV. Nombre del archivo de salida con el grupo al que pertenece cada caso y la predicción del análisis Discriminante usando la validación cruzada del análisis Discriminante Cuadrático.
file11	ARCHIVO PDF. Nombre del archivo de salida con los gráficos obtenidos del algoritmo VARSEDIG.
file12	ARCHIVO CSV. Nombre del archivo de salida con las probabilides obtenidas de comparar todas las variables entre todos los clústeres con el test U Mann-Whitney.
na	ARCHIVOS CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVOS CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVOS CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

El objetivo de este análisis es determinar que grupos estadísticamente diferentes se forman al aplicar un análisis de Correspondencias.

Este análisis puede ser útil para encontrar grupos diferentes cuando realmente no se saben cuales son. Sin embargo, es importante resaltar que solo se detectaran grupos diferentes si las variables que más variabilidad tienen dan lugar a que se diferencien grupos. Es muy posible que una variable no presente una gran variabilidad, pero sea importante para discriminar entre grupos. Este tipo de diferenciaciones basadas en variables que no tienen alta varianza, no serían detectadas en este análisis.

Para detectar los grupos potenciales que se forman se aplica un dendrograma a las puntuaciones obtenidas de los ejes que absorben una mayor varianza. Por defecto se escogen los ejes que absorben un 80% de la variabilidad, pero este valor puede ser modificado por el usuario.

Posteriormente, se realiza un análisis Discriminante para determinar si se discriminan bien los clústeres que se han generado, es decir, determinar el número de casos correctamente identificados en cada clúster.

A continuación, se realiza un test de la U Mann-Whitney para determinar si existen diferencias significativas en las variables entre los clústeres.

Por último, se aplica el algoritmo de la función VARSEDIG (véase para más detalles Guisande et al. 2016: Guisande, 2018). Con este algoritmo es posible determinar si todos los casos de cada clúster son estadísticamente diferentes de los otros clústeres.

La idea de esta función es buscar el mayor número posible de clústeres con el porcentaje de discriminación más alto. Para ello el usuario debería realizar pruebas, modificando el umbral de corte especificando el porcentaje de varianza acumulada para determinar cuantos ejes se seleccionan del análisis de Correspondencias (por defecto *por=80*) y las variables a incluir, eliminando aquellas que no sean útiles en el análisis de Correspondencias, así como aquellas que tengan poco poder de discriminación en el análisis Discriminante.

FUNCIONES

El análisis de correspondencias se realiza con la función ca del paquete ca (Greenacre & Pardo, 2006; Greenacre, 2007; Nenadic & Greenacre, 2007; Greenacre, 2013).

Para realizar el gráfico *biplot* se usó la función scatterplot del paquete car (Fox et al., 2018).

Las flechas se realizan con la función Arrows del paquete IDPmisc (Locher & Ruckstuhl, 2014).

La envolvente convexa se estima con la función chull del paquete grDevices.

El test de la U Mann-Whitney se realiza con la función *wilcox.test* del paquete base stats.

La comparación entre clústeres con el algoritmo VARSEDIG se realiza con la función VARSEDIM del paquete VARSEDIG (Guisande et al. 2016: Guisande, 2018).

El Análisis Discriminante Lineal se realizó con las funciones candisc del paquete candisc (Friendly, 2007; Friendly & Fox, 2017) y lda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2018).

El Análisis Discriminante Cuadrático se realizó con la función qda del paquete MASS (Venables & Ripley, 2002; Ripley et al., 2018).

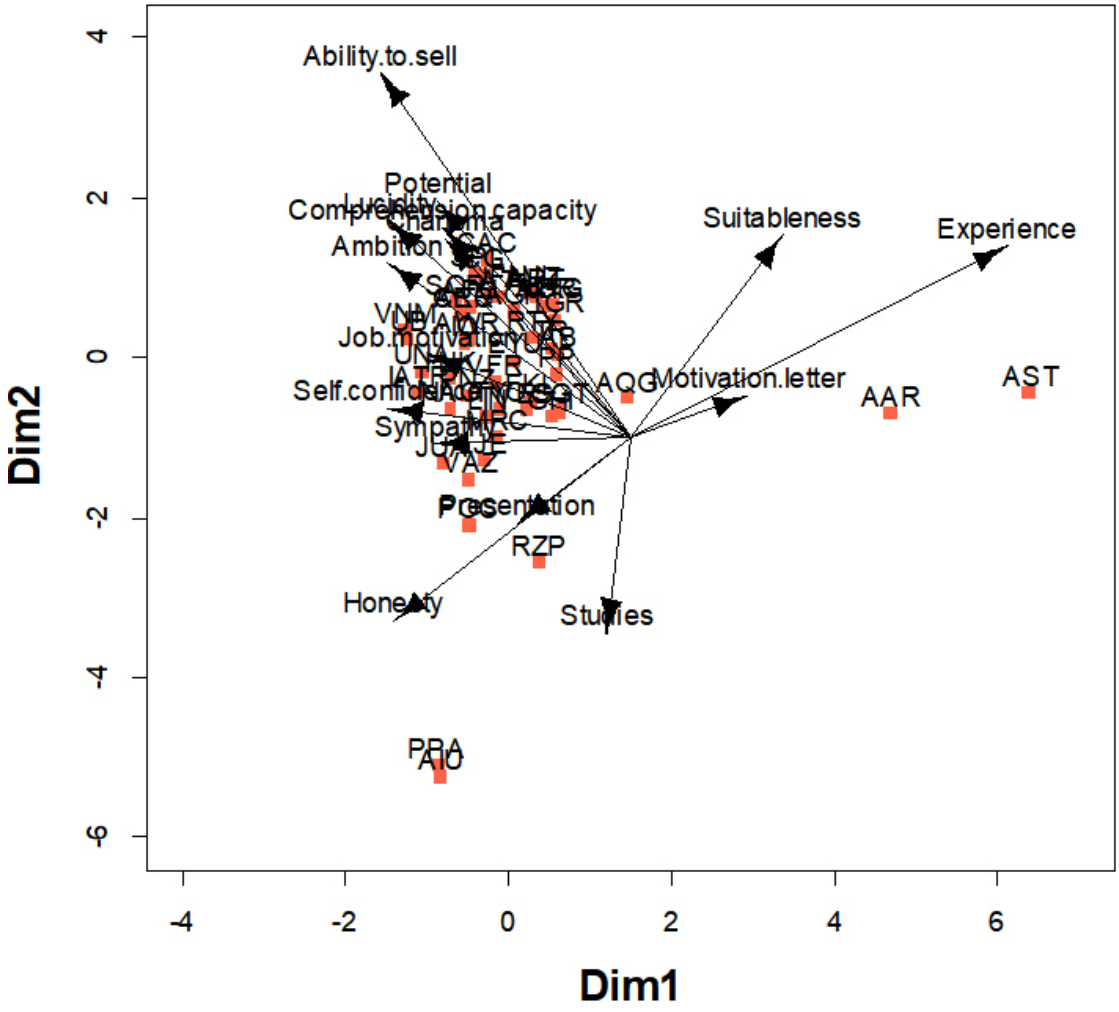
El gráfico del Análisis Discriminante con una dimensión se realizó con la función plot.cancor del paquete candisc (Friendly, 2007; Friendly & Fox, 2017).

EJEMPLO

En el ejemplo se utilizan criterios de valoración, en una escala de 0 a 10, de 48 candidatos que optan a un puesto de trabajo. El objetivo es ver que grupos se forman a partir del análisis de Correspondencias.

La Figura XV8.1 muestran que en el extremo derecho del primer eje está la experiencia junto con la adecuación al puesto y carta de motivación. En el extremo izquierdo del primer eje están asociadas la motivación y confianza. El segundo eje viene asociado al sentido comercial en el extremo superior, mientras que en el inferior están a la honestidad y nivel de estudios. Los autovalores muestran el porcentaje de varianza de los ejes: 32,9% primer eje, 25,7% el segundo, etc.

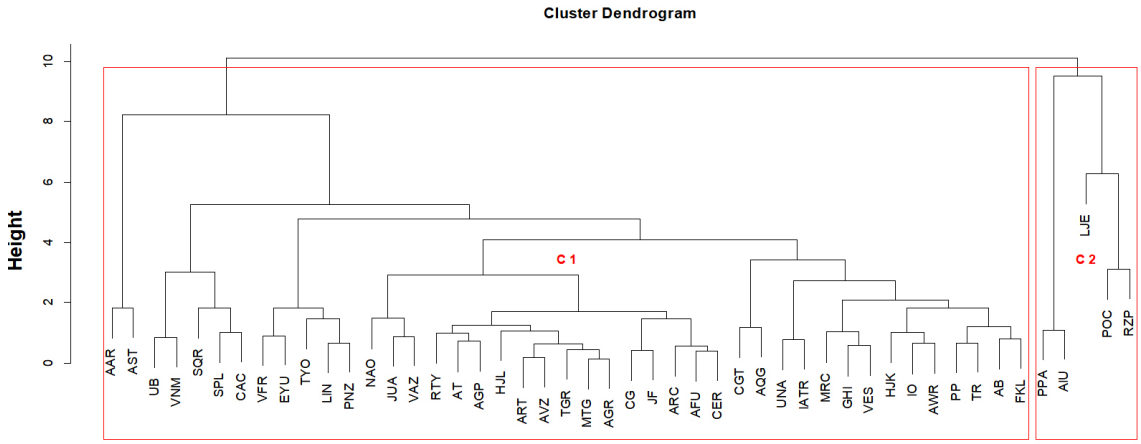
Figura XIII.1. Análisis de Correspondencias mostrando la variabilidad observada en las cualidades de candidatos a un puesto de trabajo.



La inercia (*Inertia*) indica el peso relativo o la distinta importancia de cada variable. Los mayores valores son los de la experiencia (0,0271), nivel de estudios (0,0168), honestidad (0,0155), carta de motivación (0,015), adecuación (0,0149), sentido comercial (0,012, *Ability.to.sell*), etc.

Se observa en la Figura XV8.2 que solamente hay 2 clústeres estadísticamente significativos con al menos una variable diferente según el test de la U Mann-Whitney.

Figura XV8.2. Dendrograma con las puntuaciones de los ejes seleccionados del análisis de Correspondencias.



En la Figura XV8.3 se observan las diferencias entre clústeres para cada una de las variables. Ambos clústeres se diferencian significativamente en casi todas las variables, en concreto en 13 de ellas (Figura XV8.4). Solamente no hay diferencias en la honestidad y en el nivel de estudios.

Figura XV8.3. Boxplot obtenido para cada una de las variables con valores promediados para cada clúster.

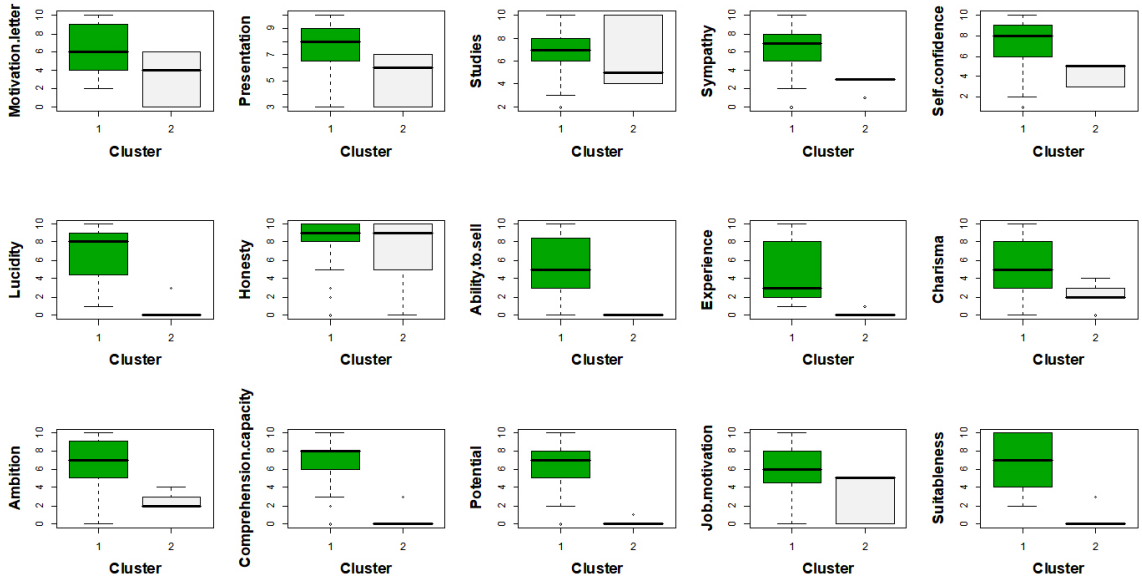
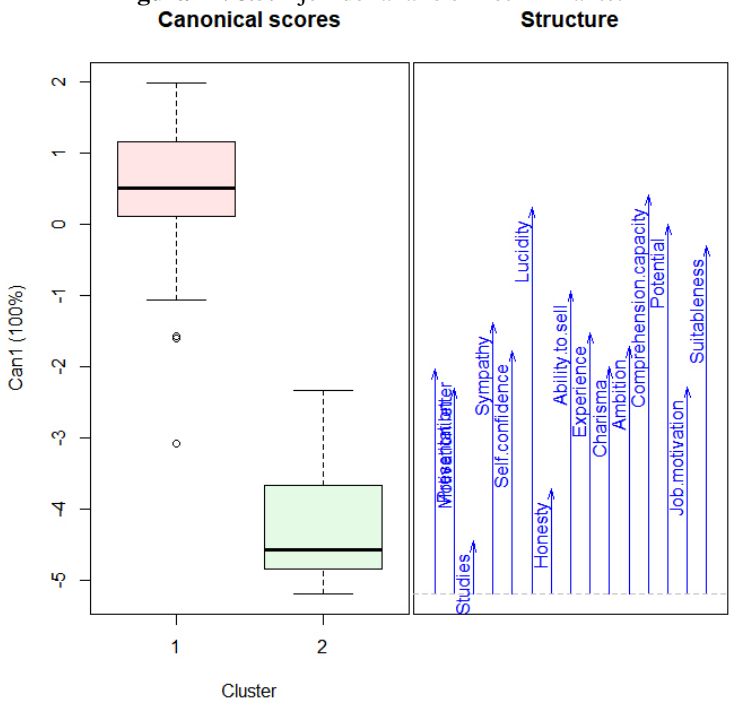


Figura XV8.4. Gráfico donde las burbujas representan el número de variables. que son estadísticamente diferentes (p<=0,05) entre clústeres.

El análisis Discriminante muestra que es posible discriminar correctamente el 87,5% de los casos por validación cruzada con el método Lineal. Todas las variables parecen ser importantes para la discriminación, salvo el nivel de estudios y la honestidad, en las cuales las flechas son pequeñas (Figura XV8.5).

Figura XV8.5. Eje I del análisis Discriminante.



Por tanto, según las pruebas realizadas con el test de la U Mann-Whitney, los clústeres son estadísti-

camente diferentes entre sí. En el Discriminante no se consiguió una completa separación de ambos grupos. En el algoritmo VARSEDIG mostró que todos los candidatos del clúster 2 son diferentes de clúster 1, pero no todos del clúster 1 son diferentes de clúster 2. Sin embargo, es muy importante resaltar que el algoritmo VARSEDIG considera dos grupos estadísticamente diferentes si el caso que más se parece de cada grupo al otro, es estadísticamente diferente usando el test de Monte-Carlo. El test de Monte-Carlo necesita un número de casos grande en cada grupo, para que realmente se pueda detectar diferencias significativas. En el caso del grupo 2 el número de candidatos es muy pequeño y, por tanto, no es posible determinar que las diferencias no sean por azar.

Value

Se obtiene:

1. Un archivo TXT con los resultados del análisis de Correspondencias. El archivo se llama por defecto "Output.TXT".

2. Un archivo CSV con las coordenadas para cada caso de los ejes de análisis de Correspondencias. El archivo se llama por defecto "Cat loadings.CSV".

3. Un archivo CSV con la estadística descriptiva de cada variable para cada uno de los clústeres obtenidos en el Dendrograma. El archivo se llama por defecto "Descriptive statistics of clusters.CSV".

4. Un archivo CSV con los datos originales de las variables y el clúster al que pertenece cada caso. El archivo se llama por defecto "Original data and cluster number.CSV".

5. Un archivo CSV con las coordenadas de las variables en el gráfico del análisis Discriminante Lineal. El archivo se llama por defecto "Var loadings-Linear.csv".

6. Un archivo CSV con las coordenadas de las categorías en el gráfico del análisis Discriminante Lineal. El archivo se llama por defecto "Cat loadings-Linear.csv".

7. Un archivo CSV con la tabla de predicciones usando la validación cruzada del análisis Discriminante Lineal. El archivo se llama por defecto "Table cross-validation-Linear.csv".

8. Un archivo CSV con el grupo al que pertenece cada caso y la predicción del análisis Discriminante usando la validación cruzada del análisis Discriminante Lineal. El archivo se llama por defecto "Cases cross-validation-Linear.csv".

9. Un archivo CSV con la tabla de predicciones usando la validación cruzada del análisis Discriminante Cuadrático. El archivo se llama por defecto "Table cross-validation-Quadratic.csv".

10. Un archivo CSV con el grupo al que pertenece cada caso y la predicción del análisis Discriminante usando la validación cruzada del análisis Discriminante Cuadrático. El archivo se llama por defecto "Cases cross-validation-Quadratic.csv".

11. Un archivo CSV con las probabilidades obtenidas de comparar todas las variables entre todos los clústeres con el test U Mann-Whitney.

12. Un archivo PDF con los gráficos obtenidos del algoritmo VARSEDIG.

13. Un gráfico del análisis de Correspondencias.

14. El Dendrograma agrupando por clústeres en función de las puntuaciones del análisis de Correspondencias.

15. Un panel gráfico con un boxplot por cada variable comparando los valores de esas variables entre cada uno de los clústeres obtenidos en el Dendrograma.

16. Un gráfico del análisis Discriminante mostrando la influencia de las variables en el eje discriminante I, diferenciando los diferentes clústeres.

17. Un gráfico del análisis Discriminante mostrando las puntuaciones de los ejes discriminante I y II, diferenciando los diferentes clústeres.

18. Un gráfico de burbujas con el número de variables que son estadísticamente diferentes entre clústeres.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2017) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. Available at: http://CRAN.R-project.org/package=candisc.

Greenacre, M. (2013). Simple, Multiple and Joint Correspondence Analysis. R package version 0.53. Available at: http://CRAN.R-project.org/package=ca.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC.

Greenacre, M.J. & Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, 35: 193-218.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162: 571-580.

Guisande, C. (2018) An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy. R package version 1.8. Available at: http://CRAN.R-project.org/package=VARSEDIG.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Nenadic, O. & Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20: 1-13.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2018) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-50. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:
```

```
data(ZXIII1)

XV8(data=ZXIII1, var=c("Motivation.letter", "Presentation", "Studies",
"Sympathy", "Self.confidence", "Lucidity", "Honesty", "Ability.to.sell",
"Experience", "Charisma", "Ambition", "Comprehension.capacity",
"Potential", "Job.motivation", "Suitableness"), labels="Candidate",
XLIM=c(-4,7), YLIM=c(-6,4))

## End(Not run)
```

XV9 *DETECCION DE VALORES ATIPICOS EN UNA COMBINACIÓN DE VARIABLES*

Description

Se detectan valores atípicos en una combinación de variables y, si el usuario lo elige, se borran automáticamente.

Usage

```
XV9(data, var, cat=NULL, varCode=NULL, remove=FALSE, Nan=FALSE,
quant=0.95, file="Outliers.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Archivo de datos.
var	Variables a combinar para detectar los valores atípicos.
cat	Si se especifica una variable, los valores atípicos se calculan considerando cada grupo de esa combinación de variables. Por ejemplo, si las variables incluyen diferentes especies, en vez de buscar los atípicos en función de todos los datos de las variables, se buscan considerando los datos de cada especie por separado.
varCode	Opcionalmente se puede seleccionar variables de la matriz original, que se exportarán en el archivo de salida con las variables en las que se detectan valores atípicos. Por ejemplo, permite seleccionar variables que son códigos de las filas.
remove	Si es TRUE, en el archivo de salida se eliminan las filas que tengan alguna variable con valor atípico, pero los valores atípicos se siguen mostrando en el gráfico. Si es FALSE, los atípicos se indican con el valor -9999.
Nan	Si es TRUE los valores atípicos se muestran como NA en vez de -9999.
quant	Cuantil del extremo superior para la eliminación de los datos atípicos.
file	ARCHIVO CSV. Nombre del archivo con las variables mostrando los valores atípicos o con las variables con los valores atípicos eliminados.
na	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
dec	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
row.names	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

En algunos análisis estadísticos como MANOVA y Componentes Principales se trabaja con combinación de variables, en vez de con las variables de forma independiente. Es posible que unas variables no tengan valores atípicos cuando se analizan de forma independiente, pero la combinación de ellas si puede dar lugar a que resulten valores atípicos. Por ello, es necesario un método estadístico para detectar valores atípicos en combinación de variables y se utiliza frecuentemente la distancia de Mahalanobis (Mahalanobis, 1936).

La distancia de Mahalanobis es una medida de la distancia entre un punto P y una distribución D. Es una generalización multidimensional de la idea de medir cuántas desviaciones estándar se alejan P de la media de D. Esta distancia es cero si P está en la media de D, y crece a medida que P se aleja de la media a lo largo de cada eje. Por tanto, la distancia Mahalanobis mide el número de desviaciones estándar de P a la media de D. Si cada uno de estos ejes se escala para tener varianza unitaria, entonces la distancia Mahalanobis corresponde a la distancia euclidiana estándar en el espacio transformado. La distancia Mahalanobis, por lo tanto, no tiene unidades, es invariante de escala y tiene en cuenta las correlaciones del conjunto de datos. La ecuación es la siguiente:

$$D^2 = (x - \mu)' C^{-1} (x - \mu)$$

donde C^{-1} es la matriz inversa de covarianzas y μ es el centro.

Una vez calculada la distancia nos queda el problema de como detectar los valores atípicos. Para muestras relativamente grandes, las distancias de Mahalanobis estimadas tienen una distribución que se aproxima a una chi-cuadrado. Este resultado se puede usar para evaluar (subjetivamente) si un punto de datos puede ser un valor atípico. Para ello normalmente se usa un gráfico Q-Q para representar las distancias de Mahalanobis de la muestra. La idea básica es la misma que en un gráfico de distribución de Normalidad. Para datos multivariados, se grafican las distancias de Mahalanobis ordenadas versus cuantiles estimados para una muestra de tamaño *n* a partir de una distribución chi-cuadrado con *p* grados de libertad (número de variables que se combinan). Esto debería parecerse a una línea recta para los datos de una distribución normal multi-variable. Los valores atípicos se mostrarán como puntos en el lado superior derecho de la gráfica para los cuales la distancia de Mahalanobis es notablemente mayor que el valor de cuantil de chi-cuadrado. Los cuantiles se determinan de la siguiente forma:

1. El cuantil *ith* se determina como el valor de chi-cuadrado (con df = número de variables) para el cual la probabilidad acumulada es $(i - 0.5)/n$.

2. El conjunto total de cuantiles de chi-cuadrado se estiman para cada valor *i* de 1 a *n* (número total de datos).

El problema es que este método es subjetivo, es decir, es a criterio del usuario decidir, en base al gráfico obtenido, que puntos se eliminan. Para eliminar esta subjetividad, en esta función se introduce un algoritmo que usando el criterio descrito para detectar valores atípicos en variables independientes definido en la función XV7, se consideran atípicos aquellas distancias de Mahalanobis que quedan fuera del cuantil superior (por defecto 0,95).

Los valores considerados como atípicos se destacan en rojo en la gráfica Q-Q, para que el usuario pueda decidir si la selección ha sido correcta.

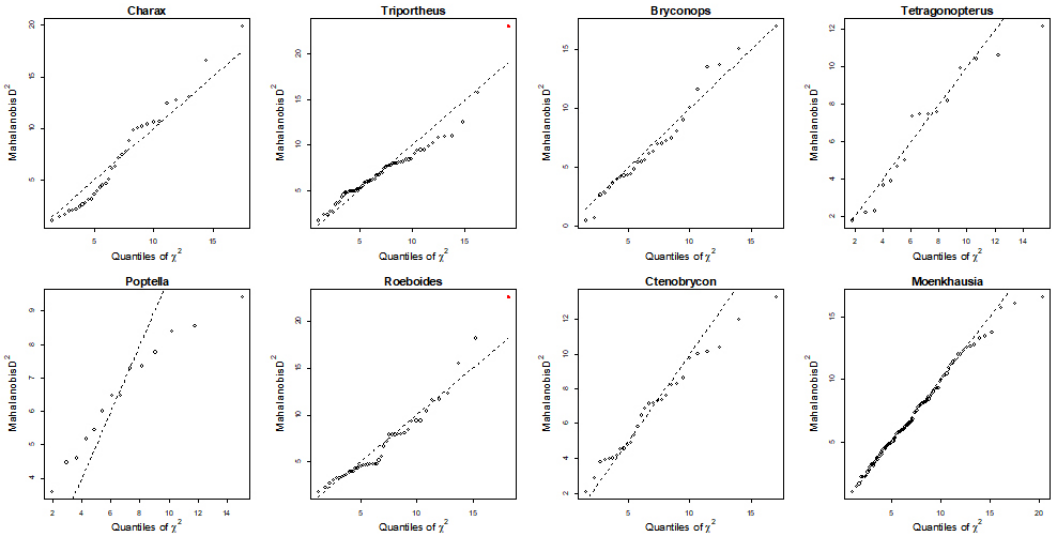
FUNCIONES

La distancia de Mahalanobis se calcula con la función link[stats]mahalanobis del paquete base stats.

EJEMPLO

Se detectan valores atípicos en la combinación de unas variables morfométricas de peces. Los valores atípicos se muestran con el valor -9999 en el archivo obtenido.

Como en el script se seleccionó realizar los cálculos por género, *cat="Genus"*, se muestra una gráfica Q-Q para cada género.



En las gráficas Q_Q que se muestran en el panel anterior, en las cuales se relaciona la distancia de Mahalanobis con los cuantiles estimados a partir de una distribución chi-cuadrado con 7 grados de libertad (ya que son 7 las variables a combinar), solamente se detectan valores atípicos en los géneros *Triporheus* y *Roeboides*, con el cuantil por defecto de 0,95.

Es interesante resaltar que en el caso del género *Poptella*, como el número de datos no es muy grande, los puntos no se ajustan a la línea recta de la distribución chi-cuadrado del gráfico Q-Q.

Value

Se exporta:

1. Un archivo CSV con las variables y la distancia de Malahanobis resultante de la combinación de las variables, donde se muestran los valores atípicos. También se exportan las variables incluidas en el argumento *varCode* si se incluyó alguna variable en este argumento. Si la opción *remove=TRUE* se eliminan las filas si la distancia de Malahanobis es atípica.

2. Un gráfico Q-Q con las distancias de Mahalanobis ordenadas versus cuantiles estimados para una muestra de tamaño *n* a partir de una distribución chi-cuadrado con *p* grados de libertad (número de variables que se combinan). En el gráfico se indica con puntos en rojo los valores atípicos. Si se eligió realizar la estimación por categorías, se muestra un gráfico Q-Q para cada categoría.

References

Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2: 49-55.

Examples

```
## Not run:
data(ZX3)

XV9(data=ZX3, var=c("M2", "M3", "M4", "M5", "M6", "M7", "M8"), cat="Genus",
varCode=c("Order", "Family", "Genus", "Species"))

## End(Not run)
```

XVI1 *ANALISIS MULTIVARIANTE DE LA VARIANZA (MANOVA)*

Description

Se aplica un análisis de varianza considerando varias variables cuantitativas dependientes.

Usage

```
XVI1(data, var, Factor, SS="III", trans=NULL, pthreshold=0.05, psig=FALSE,
ellipse=FALSE, convex=TRUE, PCH=rep(16,100), order=TRUE, cor=TRUE,
file1="Output.txt", file2="Multinormality.csv", file3="Post hoc pvalues.csv",
na="NA", dec=",", row.names=TRUE)
```

Arguments

data	Archivo de datos.
var	Variables dependientes.
Factor	Variables con los factores.
SS	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar tres tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados). El tipo III es adecuado para la mayoría de las aplicaciones, por lo que será el que utilizaremos por defecto. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del MANOVA con modelos complejos.
trans	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3

9. EXP (exponencial)

10. ASN (arcoseno)

`pthreshold` Probabilidad de corte que se utiliza para agrupar las categorías de cada factor, en función de las probabilidades obtenidas en las pruebas post hoc.

`psig` Si es TRUE, en el archivo de comparación post hoc entre grupos, solamente se muestran las que son significativas con un valor de probabilidad igual o menor al valor de corte especificado en el argumento *pthreshold*.

`ellipse` Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías en los gráficos.

`convex` Si es TRUE, la envolvente convexa se calcula para cada categoría en los gráficos.

`PCH` Símbolos de cada categoría en los gráficos.

`order` Si es TRUE, la primera variable de las coordenadas polares es la que discrimina mejor las categorías. El objetivo es mostrar la mayor separación posible entre las categorías en los gráficos. Si es FALSE, la primera variable es la especificada por el usuario en el argumento *var*.

`cor` Si es TRUE, las variables se ordenan de acuerdo con la correlación entre ellas cuando se estiman las coordenadas polares. Por lo tanto, la siguiente variable a otra variable es la que tiene una mayor correlación positiva. El objetivo es mostrar la mayor separación posible entre las categorías en los gráficos. Si es FALSE y el argumento *order=FALSE*, las variables se organizan como aparecen en el argumento *var*.

`file1` ARCHIVO TXT. Nombre del archivo de salida con los resultados.

`file2` ARCHIVO CSV. Nombre del archivo con las probabilidades obtenidas del análisis de multinormalidad.

`file3` ARCHIVO CSV. Nombre del archivo con las probabilidades obtenidas de la comparación post hoc entre grupos para cada factor.

`na` ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.

`dec` ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".

`row.names` ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

Análisis multivariante de la varianza (MANOVA)

El análisis multivariante de la varianza (MANOVA) es una extensión del análisis univariante de la varianza (ANOVA). En un ANOVA, examinamos las diferencias estadísticas en una variable dependiente continua entre las medias de las categorías de uno o varios factores. MANOVA amplía este análisis teniendo en cuenta múltiples variables dependientes continuas, y las agrupa en una combinación lineal ponderada o variable compuesta. El MANOVA comparará si la combinación creada recientemente difiere o no según los diferentes grupos o niveles de cada factor. De esta forma, el MANOVA, esencialmente, prueba si la variable de agrupación independiente (factor)

explica una cantidad estadísticamente significativa de varianza en la variable dependiente creada como combinación lineal de todas las variables independientes.

Para la aplicación de un MANOVA es necesario cumplir los siguientes supuestos:

· Las variables dependientes son continuas, aleatorias e independientes, es decir, no debe existir relación entre los grupos ni dentro de los grupos de cada factor. Por ejemplo, no es válido que los datos de un grupo deriven a partir de los datos de otro grupo (muestras relacionadas) o que se repitan datos entre grupos.

· No existen valores atípicos univariados ni multivariados. Para detectar los valores atípicos dentro de cada variable dependiente se puede usar la función XV7 y para los multivariados, es decir, los que resultan de la combinación de variables, la función XV9, ambas de StatR. Los valores atípicos multivariados son casos que tienen una combinación inusual de puntajes en las variables dependientes y, por tanto, la combinación que resulta de las variables dependientes da lugar a un valor atípico.

· No debe existir colinealidad entre las variables dependientes cuantitativas. Un valor superior a alrededor 10-12 de VIF significaría que hay colinealidad.

· La relación entre las variables dependientes cuantitativas es lineal.

· Se asume normalidad multivariada, lo cual significa que los residuos son normales para cada variable dependiente, en cada uno de los grupos o categorías de los factores.

· Se asume igualdad de las matrices de varianzas-covarianzas en los residuos de los grupos de los factores considerados.

Una contribución original de esta función es que se realiza unas pruebas post hoc, que permite agrupar las categorías que no son estadísticamente diferentes y, por tanto, mostrar cuantos grupos reales hay diferentes dentro de cada factor. Solamente se muestran las pruebas post hoc si hay más de dos categorías en el factor y si hay diferencias entre algunas de las categorías del factor.

El algoritmo consiste en realizar el MANOVA dentro de cada factor, pero en vez de comparar todos los grupos, se compara cada grupo con el resto. Las probabilidades obtenidas se graban en el archivo que se llama por defecto "Post hoc pvalues.CSV". Posteriormente se genera una matriz donde las filas son las categorías y las columnas también son categorías. En cada fila se señalan con un 1 las categorías que son estadísticamente diferentes y con 0 si no lo son. Por tanto, cada fila muestra con 1 y 0, que categorías son o no son estadísticamente diferentes de la categoría que se considera en esa fila. Esta matriz permite conocer que categorías son más parecidas entre sí, en función de que tengan más categorías en significativas en común.

A esta matriz se le va aplicando sucesivos dendrogramas, donde se van realizando diferente número de clústeres. A los clústeres que se van formando se aplica un MANOVA y el algoritmo se para cuando ya no hay diferencias entre alguno de los clústeres. Por tanto, el algoritmo permite obtener el mayor número posible de clústeres diferentes que hay dentro de cada factor, agrupando en cada clúster las categorías que no son estadísticamente diferentes.

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2015) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors.

La función shapiro.test se usa para realizar el test de Shapiro-Wilk, el modelo lineal se realiza con la función lm, ambas del paquete base stats.

El MANOVA tipo II-III se realiza con la función Manova del paquete car (Fox et al., 2018a). El MANOVA tipo I se realiza con la función manova del paquete base stats.

Para determinar la homogeneidad de varianzas entre las categorías de los factores se aplicó el test M de Box, para lo cual se usó la función boxM del paquete heplots (Fox et al., 2018b).

Para el cálculo del VIF se usó la función vif del paquete usdm (Naimi, 2013; Naimi et al., 2014).

EJEMPLO

Los datos son del peso al nacer y al destete de terneros. Los individuos al destete son diferentes de los que se les midió el peso al nacer. Por tanto, las variables dependientes no están relacionadas. El objetivo es determinar si hay diferencias significativas en ambos pesos en función del sexo o del padre. El objetivo final es seleccionar parentales que tengan terneros con mayor peso.

Si se realiza un ANOVA a cada una de las variables dependientes, peso al nacer y peso al destete, no hay diferencias significativas en el peso al nacer entre padres (p =0,726) ni sexos (p = 0,201). Tampoco hay diferencias significativas en el peso al destete entre padres (p= 0,109) ni sexos (p = 0,431). Sin embargo, el MANOVA muestra algo diferente.

En el archivo TXT de resultados, lo primero que se muestra son los valores de VIF, los cuales al ser pequeños nos indican que no existe colinealidad entre las variables dependientes, que es una de los supuestos del MANOVA.

```
        Variables      VIF
1 Weaning.weight 5.290086
2   Birth.weight 5.290086
```

El requisito de igualdad de varianzas entre grupos de factores también se cumple. El test M de Box muestra que hay igualdad en las matrices de varianzas-covarianzas en los residuos de los grupos de los factores considerados, tanto en el caso del sexo (p = 0,734) como entre los padres (p = 0,957).

```
[1] "Box's M-Test"

[[1]][[3]]
[[1]][[3]][[1]]
[[1]][[3]][[1]][[1]]
[1] "Father"

[[1]][[3]][[1]][[2]]

     Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 7.0034, df = 15, p-value = 0.9576


[[1]][[3]][[2]]
[1] "Gender"

[[1]][[3]][[3]]

     Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 1.2758, df = 3, p-value = 0.7349
```

En el archivo que se llama por defecto "Multinormality.csv", se observa que la multinormalidad también se cumple. Las muestras son normales en las dos variables para todos los grupos de los dos factores.

Variable	Factor	Group	p.Shapiro	p.Lillie
Weaning.weight	Father	1	0,99348891	0,98806216
Weaning.weight	Father	2	0,44018032	0,70511091
Weaning.weight	Father	3	0,54544602	0,3992761
Weaning.weight	Father	4	0,53544977	0,13810582
Weaning.weight	Father	5	0,76143615	0,26239204
Weaning.weight	Father	6	0,76023498	0,31492003
Weaning.weight	Gender	Female	0,54449487	0,55550131
Weaning.weight	Gender	Male	0,99909073	0,60009249
Birth.weight	Father	1	0,16884499	0,51136833
Birth.weight	Father	2	0,83346753	0,76979371
Birth.weight	Father	3	0,47084057	0,24041839
Birth.weight	Father	4	0,53414884	0,69173319
Birth.weight	Father	5	0,61053694	0,60533207
Birth.weight	Father	6	0,16164634	0,3535231
Birth.weight	Gender	Female	0,90294659	0,72885732
Birth.weight	Gender	Male	0,59722188	0,3750236

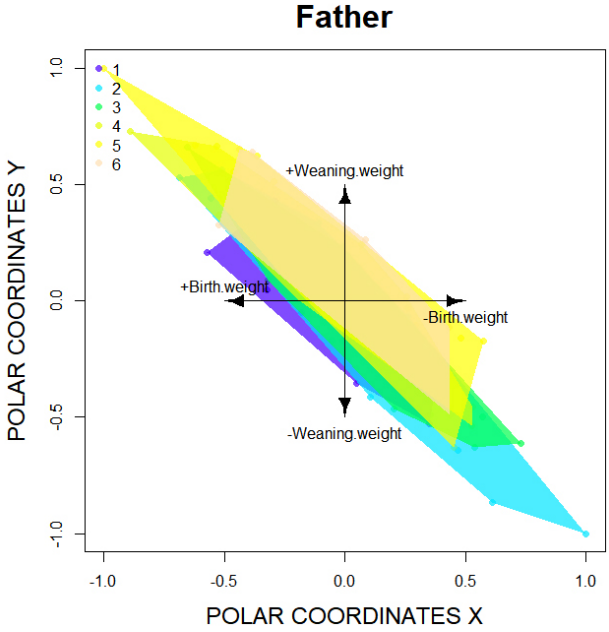
Los siguientes resultados son el análisis de MANOVA. Son cuatro los test que se utilizan más frecuentemente: Pillai, Wilks, Hotelling-Lawley y Roy. Ninguna de estas pruebas se puede identificar claramente como la mejor prueba para ser usada en todas las situaciones (Lee, 1971; Pillai y Jayachandran, 1967). La efectividad comparativa de cada uno de estos test cambia en relación con las características específicas de los datos. Sin embargo, teniendo en cuenta un amplio espectro de condiciones reales de datos, las pruebas de Pillai, Wilks y Hotelling-Lawley generalmente funcionan de manera similar, particularmente cuando se cumplen los supuestos (Johnson & Wichern, 2002; Finch & French, 2013).

En la siguiente tabla se observa que en todos los casos los valores de los test coinciden. En el caso del factor padre hay diferencias significativas, lo cual significa que el peso de los terneros al nacer y al destete varía en función del padre (p = 0,010). Sin embargo, no existen diferencias significativas entre sexos (p = 0,977) ni en la interacción padre-sexo (p = 0,684). La no interacción significa que los terneros nacidos de diferentes padres no tienen un peso diferente en función de si son machos o hembras. En otras palabras, la interacción significativa significaría que en algunos padres, los terneros machos pesan más que las hembras, y viceversa en otros padres.

```
Multivariate Tests: Father
                Df test stat approx F num Df den Df   Pr(>F)
Pillai           1 0.1108180 4.860537      2     78 0.010248 *
Wilks            1 0.8891820 4.860537      2     78 0.010248 *
Hotelling-Lawley 1 0.1246292 4.860537      2     78 0.010248 *
Roy              1 0.1246292 4.860537      2     78 0.010248 *

Multivariate Tests: Gender
                Df test stat   approx F num Df den Df Pr(>F)
Pillai           1 0.0005834 0.02276654      2     78 0.9775
Wilks            1 0.9994166 0.02276654      2     78 0.9775
Hotelling-Lawley 1 0.0005838 0.02276654      2     78 0.9775
Roy              1 0.0005838 0.02276654      2     78 0.9775

Multivariate Tests: Father:Gender
                Df test stat  approx F num Df den Df  Pr(>F)
Pillai           1 0.0096591 0.3803799      2     78 0.68486
Wilks            1 0.9903409 0.3803799      2     78 0.68486
Hotelling-Lawley 1 0.0097533 0.3803799      2     78 0.68486
Roy              1 0.0097533 0.3803799      2     78 0.68486
```

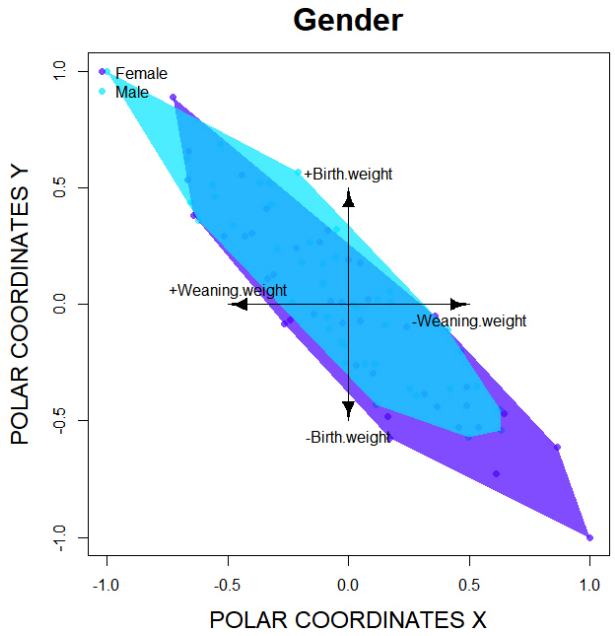
La Figura XVI1.1 muestra los grupos de padres en el gráfico de coordenadas polares donde se observa que hay grupos diferentes, lo que corrobora las diferencias entre padres que mostró el MANOVA en los pesos al nacer y al destete.

Figura XVI1.1. Coordenadas polares de los diferentes padres en función de las dos variables dependientes analizadas.



Sin embargo, en el caso del sexo (Figura XVI1.2) se observa dos grupos que solapan, lo que también permite visualizar lo mostrado en el MANOVA, que no hay diferencias entre sexos en el peso de los terneros al nacer y al destete.

Figura XVI1.2. Coordenadas polares de ambos sexos en función de las dos variables dependientes analizadas.

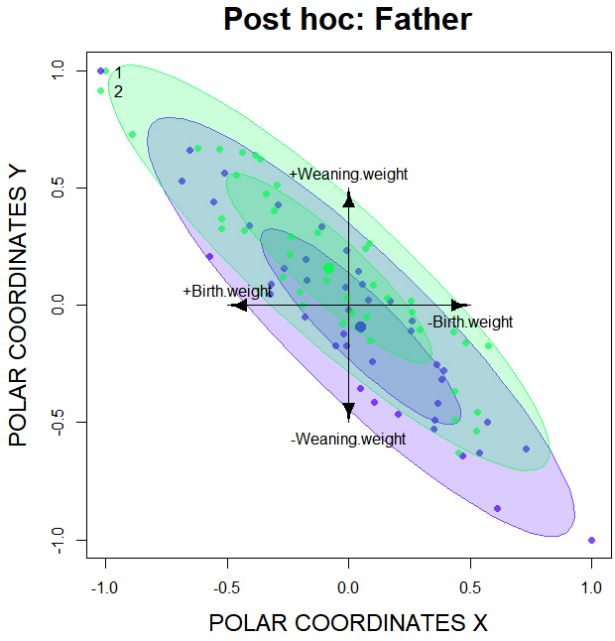


En la función se ha implementado un algoritmo que realiza unas pruebas post hoc, las cuales se detallan en la sección *details*. Solamente se muestran las pruebas post hoc si hay más de dos categorías en el factor y si hay diferencias entre algunas de las categorías del factor. Por ello, en este ejemplo, no sale información sobre pruebas post hoc del factor sexo, ya que solo hay dos categorías. En el caso del factor padre se generan dos grupos. El grupo 1 incluye a los padres 1, 2 y 3, y el grupo 2 a los padres 4, 5 y 6, como se observa en tabla que se muestra a continuación y sale en el archivo de resultados. Por lo tanto, dentro de las diferencias detectadas entre padres, existen dos grupos diferenciados significativamente de padres (Figura XVI1.3).

```
"Post hoc groups: Father"

Father Cluster
  1       1
  2       1
  3       1
  4       2
  5       2
  6       2
```

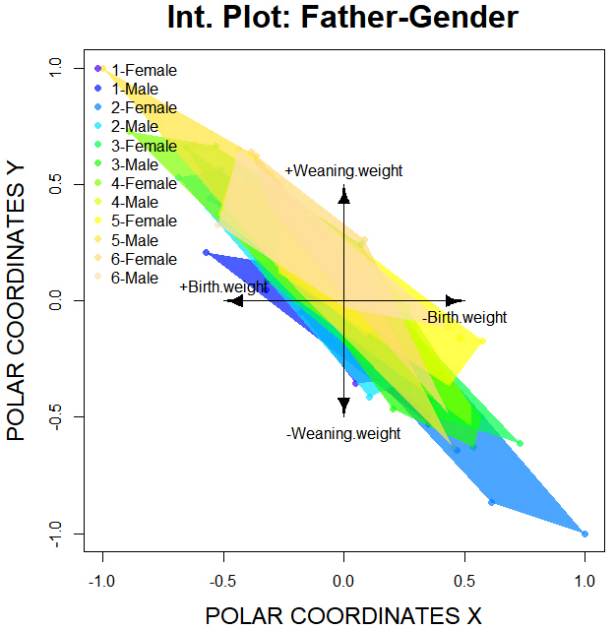
La Figura XVI1.3 muestra el gráfico post hoc para el factor padre. Para obtener el gráfico con las elipses es necesario correr el script añadiendo los argumentos *convex=FALSE* y *ellipse=TRUE*. El punto central de las elipses de cada grupo permite diferenciar mejor que el grupo 2 tiene un peso promedio mayor, tanto al nacimiento como al destete.

Figura XVI1.3. Coordenadas polares de los grupos que se forman de padres, que muestran diferencias significativas en las pruebas post hoc.



El último gráfico es el de interacción entre ambos factores (Figura XVI1.4). Es más difícil de visualizar, pero se puede entrever que, para un mismo padre, por ejemplo el padre 6, los valores son muy semejantes en machos y hembras. Así es parecido para el resto de padres, lo cual permite visualizar que no hay una interacción significativa entre los factores.

Figura XVI1.4. Coordenadas polares de la interacción entre ambos factores



Todos estos gráficos de coordenadas polares se pueden modificar a gusto del usuario usando los

datos originales y la función F46 para realizar coordenadas polares disponible en el paquete PlotsR.

Value

Se obtiene:

1. Un archivo TXT con los resultados del VIF, el test M de Box y los resultados del MANOVA.

2. Un archivo CSV con las probabilidades obtenidas del análisis de multinormalidad.

3. Un archivo CSV con las probabilidades obtenidas de la comparación post hoc entre los grupos de cada factor.

4. Gráficos de coordenadas polares mostrando la separación entre categorías para cada factor, de los grupos resultantes de las pruebas post hoc y de las interacciones entre factores.

References

Finch, H & French, B. (2013) A Monte Carlo comparison of robust MANOVA. *Test Statistics Journal of Modern Applied Statistical Methods*, 12: 35-81.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018a) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Fox, J, Friendly, M. & Monette, G. (2108b) Visualizing Hypothesis Tests in Multivariate Linear Models. R package version 1.3.5. Available at: http://CRAN.R-project.org/package=heplots.

Gross, J. (2015) Tests for Normality. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=nortest.

Johnson, R.A. & Wichern, D.W. (2002) *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.

Lee, Y.S. (1971) Asymptotic formulae for the distribution of a multivariate test statistic: Power comparisons of certain multivariate tests. *Biometrika*, 58: 647-651.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Pillai, K.C.S. & Jayachandran, K. (1967) Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika*, 54: 195-210.

Examples

```
## Not run:

data(ZXV3)

XVI1(data=ZXV3, var=c("Weaning.weight", "Birth.weight"), Factor=c("Father","Gender"))

## End(Not run)
```

XVI2 *ANALISIS MULTIVARIANTE DE LA COVARIANZA (MANCOVA)*

Description

Se aplica un análisis de covarianza considerando varias variables cuantitativas dependientes.

Usage

```
XVI2(data, var, Factor, Covar, SS="III", trans=NULL, pthreshold=0.05,
psig=FALSE, ellipse=FALSE, convex=TRUE, PCH=rep(16,100), order=TRUE,
cor=TRUE, file1="Output.txt", file2="Multinormality.csv",
file3="Post hoc pvalues.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data`	Archivo de datos.
`var`	Variables dependientes.
`Factor`	Variables con los factores.
`Covar`	Covariables.
`SS`	Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar tres tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados). El tipo III es adecuado para la mayoría de las aplicaciones, por lo que será el que utilizaremos por defecto. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del MANCOVA con modelos complejos.
`trans`	Tipo de transformación que se aplica a los datos: 1. NULL (sin transformar) 2. 1/x2 3. 1/x 4. LN 5. LOG 6. SQR (raíz cuadrada) 7. x2 8. x3 9. EXP (exponencial) 10. ASN (arcoseno)
`pthreshold`	Probabilidad de corte que se utiliza para agrupar las categorías de cada factor, en función de las probabilidades obtenidas en las pruebas post hoc.
`psig`	Si es TRUE, en el archivo de comparación post hoc entre grupos, solamente se muestran las que son significativas con un valor de probabilidad igual o menor al valor de corte especificado en el argumento *pthreshold*.

`ellipse`	Si es TRUE, se calculan las elipses de los niveles de significación al 0,5 (elipse interna) y al 0,95 (elipse externa) de cada una de las categorías en los gráficos.
`convex`	Si es TRUE, la envolvente convexa se calcula para cada categoría en los gráficos.
`PCH`	Símbolos de cada categoría en los gráficos.
`order`	Si es TRUE, la primera variable de las coordenadas polares es la que discrimina mejor las categorías. El objetivo es mostrar la mayor separación posible entre las categorías en los gráficos. Si es FALSE, la primera variable es la especificada por el usuario en el argumento *var*.
`cor`	Si es TRUE, las variables se ordenan de acuerdo con la correlación entre ellas cuando se estiman las coordenadas polares. Por lo tanto, la siguiente variable a otra variable es la que tiene una mayor correlación positiva. El objetivo es mostrar la mayor separación posible entre las categorías en los gráficos. Si es FALSE y el argumento *order=FALSE*, las variables se organizan como aparecen en el argumento *var* y *Covar*.
`file1`	ARCHIVO TXT. Nombre del archivo de salida con los resultados.
`file2`	ARCHIVO CSV. Nombre del archivo con las probabilidades obtenidas del análisis de multinormalidad.
`file3`	ARCHIVO CSV. Nombre del archivo con las probabilidades obtenidas de la comparación post hoc entre grupos para cada factor.
`na`	ARCHIVO CSV. Texto que se utiliza en las celdas sin datos.
`dec`	ARCHIVO CSV. Define si se utiliza como separador de decimales la coma "," o el punto ".".
`row.names`	ARCHIVO CSV. Valor lógico que define si se ponen identificadores en las filas o un vector con un texto para cada una de las filas.

Details

CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

Análisis multivariante de la covarianza (MANCOVA)

El análisis multivariante de la covarianza (MANCOVA) es una extensión del análisis univariante de la covarianza (ANCOVA). Parte de la variabilidad observada en las variables dependientes puede ser debida a otras variables. Por lo tanto, si queremos saber realmente si uno o varios factores explican la variabilidad de unas variables dependientes, es necesario eliminar cualquier tipo de variación asociada a otras variables, lo cual se hace introduciendo estas variables como covariables. Se aconseja al lector leer la sección *details* de la función IX4 sobre ANCOVA, ya que, como se mencionó anteriormente, el MANCOVA es una extensión del ANCOVA.

Los supuestos del MANCOVA son iguales a los del MANOVA (ver función XVI1 para más detalles), con el añadido de que las covariables también deben ser independientes y debe existir una relación linear entre las variables dependientes y las covariables.

FUNCIONES

Se utiliza la función lillie.test del paquete nortest (Gross, 2015) para realizar el test de Normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors.

La función shapiro.test se usa para realizar el test de Shapiro-Wilk, el modelo lineal se realiza con la función lm, ambas del paquete base stats.

El MANCOVA tipo II-III se realiza con la función Manova del paquete car (Fox et al., 2018a). El MANCOVA tipo I se realiza con la función manova del paquete base stats.

Para determinar la homogeneidad de varianzas entre las categorías de los factores se aplicó el test M de Box, para lo cual se usó la función boxM del paquete heplots (Fox et al., 2018b).

Para el cálculo del VIF se usó la función vif del paquete usdm (Naimi, 2013; Naimi et al., 2014).

EJEMPLO

Datos biométricos de dos especies de tiburones (*Scyliorhinus canicula* y *Galeus melastomus*), los cuales fueron tomados por dos investigadores en cuatro zonas (dos en el Mediterráneo y dos en el Atlántico).

El objetivo es determinar si hay diferencias en la base de la aleta anal y la altura de la aleta anal, entre especies e investigadores, considerando el efecto de la longitud total.

En el script se incluyó el argumento *trans="SQR"*, es decir, se transforman las variables dependientes con la raíz cuadrada, para poder cumplir con los requisitos de multinormalidad e igualdad de las matrices de varianzas-covarianzas en los residuos de los grupos de los factores. Sin embargo, la transformación implica que existe colinealidad entre las variables.

En el archivo TXT de resultados, lo primero que se observa son los valores de VIF, los cuales muestran que existe un poco de colinealidad entre las variables dependientes. Si se hace el análisis sin transformar los datos, no se observa esta colinealidad, pero no se cumplen los requisitos de multinormalidad y homogeneidad de varianzas.

```
     Variables      VIF
Base.anal     16.54237
Longitud.anal 16.54237
```

El requisito de igualdad de varianzas entre grupos de factores también se cumple en ambos factores: investigador (p = 0.151) y especie (p = 0,497).

```
"Box's M-Test"

"Especie"

        Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 2.3794, df = 3, p-value = 0.4975

"Investigador"

        Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 5.2949, df = 3, p-value = 0.1514
```

En el archivo que se llama por defecto "Multinormality.csv", se observa que la multinormalidad también se cumple. Las muestras son normales en las dos variables para todos los grupos de los dos factores, aunque en un par de grupos hay discrepancias entre los dos test de normalidad.

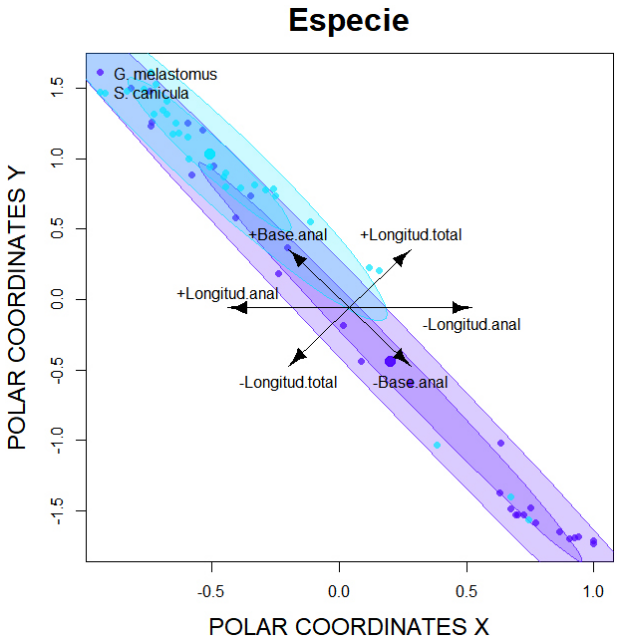
Variable	Factor	Group	p.Shapiro	p.Lillie
Base.anal	Especie	G. melastomus	0,23931599	0,60884961
Base.anal	Especie	S. canicula	0,05688983	0,12299153
Base.anal	Investigador	1	0,33911484	0,40317304
Base.anal	Investigador	2	0,40685597	0,58805688
Longitud.anal	Especie	G. melastomus	0,02955997	0,12842554
Longitud.anal	Especie	S. canicula	0,40485706	0,09361364
Longitud.anal	Investigador	1	0,21989073	0,16032938
Longitud.anal	Investigador	2	0,06922551	0,00440897

Los siguientes resultados son el análisis de MANCOVA. Son cuatro los test que se utilizan más frecuentemente: Pillai, Wilks, Hotelling-Lawley y Roy. En la siguiente tabla se observa que en todos los casos los valores de los test coinciden. La covariable longitud total claramente está relacionada con las variables dependientes ($p < 0{,}001$). Sin embargo, no existen diferencias significativas entre especies ($p = 0{,}207$), ni entre investigadores ($p = 0{,}534$), ni existen interacción investigador-especie ($p = 0{,}27$). La no interacción significa que los dos investigadores miden igual las dos especies.

```
Multivariate Tests: Longitud.total
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1   0.97544  1032.48      2     52 < 2.22e-16 ***
Wilks             1   0.02456  1032.48      2     52 < 2.22e-16 ***
Hotelling-Lawley  1  39.71078  1032.48      2     52 < 2.22e-16 ***
Roy               1  39.71078  1032.48      2     52 < 2.22e-16 ***

Multivariate Tests: Especie
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.0587534 1.622942      2     52 0.20715
Wilks             1 0.9412466 1.622942      2     52 0.20715
Hotelling-Lawley  1 0.0624208 1.622942      2     52 0.20715
Roy               1 0.0624208 1.622942      2     52 0.20715

Multivariate Tests: Investigador
                 Df test stat  approx F num Df den Df  Pr(>F)
Pillai            1 0.0237820 0.6333962      2     52 0.53483
Wilks             1 0.9762180 0.6333962      2     52 0.53483
Hotelling-Lawley  1 0.0243614 0.6333962      2     52 0.53483
Roy               1 0.0243614 0.6333962      2     52 0.53483

Multivariate Tests: Especie:Investigador
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.0490789 1.341911      2     52 0.27024
Wilks             1 0.9509211 1.341911      2     52 0.27024
Hotelling-Lawley  1 0.0516119 1.341911      2     52 0.27024
Roy               1 0.0516119 1.341911      2     52 0.27024
```

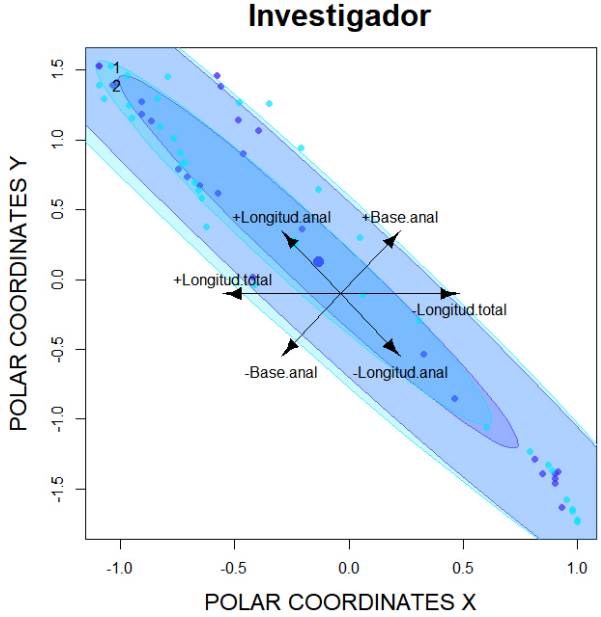
La Figura XVI2.1 muestra que no hay diferencias entre especies, lo que corrobora los resultados del MANCOVA. Es importante mencionar que en el gráfico se incluyen las covariables, para así poder visualizar mejor el efecto de ellas sobre las variables dependientes.

Figura XVI2.1. Coordenadas polares de las especies en función de las variables dependientes y covariables analizadas.



En el caso del investigador (Figura XVI2.2) tampoco se observan diferencias entre ambos.

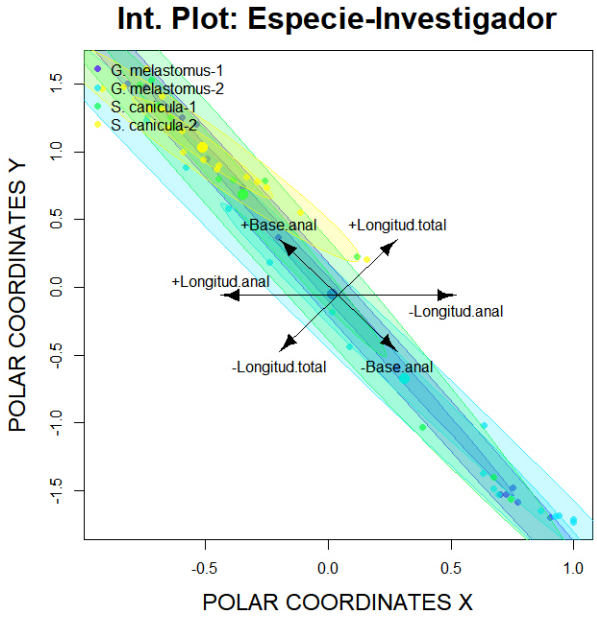
Figura XVI2.2. Coordenadas polares de ambos investigadores en función de las variables dependientes y covariables analizadas.



No sale información sobre las pruebas post hoc, ni el archivo con las probabilidades de comparación ni los gráficos, porque en ambos factores solo hay dos grupos.

El último gráfico es el de interacción entre ambos factores (Figura XVI2.3), donde se ve que ambos investigadores miden igual ambas especies y, por ello, no hay interacción.

Figura XVI2.3. Coordenadas polares de la interacción entre ambos factores



Todos estos gráficos de coordenadas polares se pueden modificar a gusto del usuario usando los datos originales y la función F46 para realizar coordenadas polares disponible en el paquete PlotsR.

Value

Se obtiene:

1. Un archivo TXT con los resultados del VIF, el test M de Box y los resultados del MANCOVA.

2. Un archivo CSV con las probabilidades obtenidas del análisis de multinormalidad.

3. Un archivo CSV con las probabilidades obtenidas de la comparación post hoc entre los grupos de cada factor.

4. Gráficos de coordenadas polares mostrando la separación entre categorías para cada factor, de los grupos resultantes de las pruebas post hoc y de las interacciones entre factores.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018a) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Fox, J, Friendly, M. & Monette, G. (2108b) Visualizing Hypothesis Tests in Multivariate Linear Models. R package version 1.3.5. Available at: http://CRAN.R-project.org/package=heplots.

Gross, J. (2015) Tests for Normality. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=nortest.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:
data(ZIX3)

XVI2(data=ZIX3, var=c("Base.anal", "Longitud.anal"), Factor=c("Especie",
"Investigador"), Covar="Longitud.total", trans="SQR", ellipse=TRUE, convex=FALSE)

## End(Not run)
```

ZII1 *ABUNDANCIA DE PECES*

Description

Abundancia de varias especies de peces en dos zonas de muestreo (S1 y S2), y la temperatura y salinidad en cada una de las zonas.

Usage

```
data(ZII1)
```

Format

Una matriz con 5 columnas con la taxonomía de varias especies de peces, 2 columnas con la abundancia en 2 zonas, y la temperatura y salinidad en ambas zonas.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZIV1 *PESO Y ALTURA*

Description

Peso y altura, en diferentes países y ciudades, de hombres y mujeres.

Usage

```
data(ZIV1)
```

Format

Una matriz con 5 columnas: país, ciudad, sexo, altura (en cm) y peso (en kg).

ZIV3 *TIEMPO DE INMUNIDAD*

Description

Datos del tiempo (en meses) que permanecen inmunes un grupo de mujeres y hombres después de que se les suministra una vacuna.

Usage

```
data(ZIV3)
```

Format

Una matriz con 2 columnas: tiempo de inmunidad y sexo.

ZIX1 *PESO Y ALTURA*

Description

Datos de peso y altura de niños y niñas entre los 2 y 4 años en Italia y España.

Usage

```
data(ZIX1)
```

Format

Una matriz con 5 columnas: país, sexo, edad, peso y altura.

ZIX10 *ERITROCITOS EN SANGRE EN DIFERENTES INDIVIDUOS*

Description

Datos de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos. La diferencia con el archivo ZIX9 es que se identifica a cada persona con una etiqueta.

Usage

```
data(ZIX10)
```

Format

Una matriz con 5 columnas: individuo y las concentraciones de eritrocitos al inicio, y en los meses 1, 2 y 3 después de inciar el tratamiento.

ZIX11 *DATOS CLIMATICOS*

Description

Datos climáticos diarios en los años 1990 y 2000, en tres ciudades de España: Huelva, Palma de Mallorca y Vigo.

Usage

```
data(ZIX11)
```

Format

Una matriz con 14 columnas: Ciudad, Altitud, Año, Mes, Día, T.Max, T.Min, T.media, Dir.Viento, Vel.Viento, Precipitación, Sol, Presión.máx y Presión.min.

Source

http://www.aemet.es/es/portada

ZIX3 *MORFOMETRIA DE TIBURONES*

Description

Datos biométricos de dos especies de tiburones (*Scyliorhinus canicula* y *Galeus melastomus*), los cuales fueron tomados por dos investigadores en cuatro zonas (dos en el Mediterráneo y dos en el Atlántico).

Usage

```
data(ZIX3)
```

Format

Una matriz con 6 columnas: especie, investigador, longitud total, longitud de la base de la aleta anal y altura de la aleta anal.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZIX5 *CAPACIDAD DE RECORDAR*

Description

En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vinos de diferentes regiones. Se midió si la capacidad de acertar la procedencia del vino variaba en el tiempo (después de una hora, un día, una semana y un mes), entre los expertos y no expertos, y si la edad de las personas influía en la capacidad de acierto. Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12.

Usage

```
data(ZIX5)
```

Format

Una matriz con 7 columnas: catador, edad, si tiene o no tiene experiencia (SI/NO), y tiempo de la medida (Hora, Día, Semana y Mes), con el grado de acierto en una escala de 0 a 12.

ZIX6 *LIQUENES COMO INDICADORES DE CONTAMINACION*

Description

Datos del promedio de especies de líquenes encontrados en dos especies de árbol, *Populus alba* (Chopo blanco) y *Aesculus hippocastanum* (Castaño de indias), en tres ciudades de España (Madrid, Barcelona y Sevilla), en tres tipos de ambientes (poca, media y alta contaminación) y por dos investigadores.

Usage

```
data(ZIX6)
```

Format

Una matriz con 5 columnas: ciudad, especies de árbol, grado de contaminación, investigadores y número de especies de líquenes.

ZIX7 *NUTRIENTES EN LAGOS DEL AMAZONAS*

Description

Datos de la concentración de nitrito, nitrato y amonio (microM/L) en lagos de la amazonia colombiana en dos meses diferentes.

Usage

```
data(ZIX7)
```

Format

Una matriz con 5 columnas: mes, lago y las concentraciones de nitrito, nitrato y amonio.

ZIX9 *ERITROCITOS EN SANGRE EN FUNCION DEL SEXO*

Description

Datos de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos.

Usage

```
data(ZIX9)
```

Format

Una matriz con 5 columnas: sexo y las concentraciones de eritrocitos al inicio, y en los meses 1, 2 y 3 después de inciar el tratamiento.

ZVI1 *ABUNDANCIA DE CARACOLES DE MAR*

Description

Abundancia de dos especies de caracoles de mar herbívoras (*Littorina littorea* y *Littorina saxatilis*) y una especie carnívora (*Nucella lapillus*) en dos zonas de muestreo.

Usage

```
data(ZVI1)
```

Format

Una matriz con 4 columnas: género, especie y abundancia de la especies en dos zonas de muestreo.

ZVII1 *ESTUDIO SOBRE FUMADORES*

Description

Datos del grado en que fuman hombres y mujeres en distintos centros de trabajo. Las categorías utilizadas fueron: 1 (no fumador), 2 (entre 1 y 10 cigarrillos diarios), 3 (entre 11 y 20 cigarrillos diarios), 4 (entre 1 y 2 cajetillas diarias) y 5 (más de 2 cajetillas diarias). También hay información de si alguno de los padres de estos trabajadores son fumadores y las categorías son: trabajadores en que alguno de sus padres es fumador (categoría con valor 1) y el otro grupo son aquellos en que ninguno de sus padres es fumador (categoría con valor 0).

Usage

```
data(ZVII1)
```

Format

Una matriz con 4 columnas: sexo, el centro de trabajo, si alguno de los padres fuma y el grado de fumador.

ZVII2 *INCIDENCIA DEL VIRUS DEL HERPES LABIAL*

Description

Datos de la presencia del virus del herpes labial (0 es no portador y 1 es portador) en hombres y mujeres (1 es hombre y 2 es mujer).

Usage

```
data(ZVII2)
```

Format

Una matriz con 2 columnas: sexo y si la persona es o no es portadora del virus.

ZVII3 *CATA DE VINOS*

Description

Datos sobre como evalúan 4 diferentes tipos de vinos 15 enólogos. La categoría 1 es Muy malo, 2 es Malo, 3 es Bueno y 4 Muy bueno.

Usage

```
data(ZVII3)
```

Format

Una matriz con 3 columnas: enólogo, tipo de vino y la calidad.

ZVII4 *RESPUESTA ALERGICA DE UN ANTIBIOTICO EN LA PIEL EN DOS SESIONES DE CONTROL*

Description

Se administró un antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas entre los días 5 y 10 después del tratamiento.

Usage

```
data(ZVII4)
```

Format

Una matriz con 3 columnas: personas que se les suministraron y no suministraron el antibiótico, respuesta a día 5 y respuesta a día 10.

ZVII5 *RESPUESTA ALERGICA DE UN ANTIBIOTICO EN LA PIEL EN VARIAS SESIONES DE CONTROL*

Description

Se administró un antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas entre los días 5, 10, 15 y 20 después del tratamiento.

Usage

```
data(ZVII5)
```

Format

Una matriz con 5 columnas: personas que sí se les suministraron y no suministraron el antibiótico, y las respuestas a los días 5, 10, 15 y 20.

ZVIII1 *RESPUESTA ALERGICA DE UN ANTIBIOTICO EN LA PIEL ANALIZADO EN EL TIEMPO*

Description

Se administró un antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas entre los días 5, 10, 15 y 20 después del tratamiento.

Usage

```
data(ZVIII1)
```

Format

Una matriz con 3 columnas: personas que sí se les suministraron y no suministraron el antibiótico, el tipo de respuesta en los días 5, 10, 15 y 20.

ZVIII2 *MORTALIDAD DE BEBES*

Description

Datos de una muestra de bebés con el peso al nacer: I (menos de 1500 g), II (de 1500 a 2499 g), III(de 2500 a 4199 g) y IV(más de 4200 g), en dos hospitales y si el bebé moría antes del año de vida.

Usage

```
data(ZVIII2)
```

Format

Una matriz con 3 columnas: hospital, grupo en que estaba el bebé en función de su peso al nacer y si el bebé vivía o moría antes del año de vida.

ZX1 *REPRODUCCION EN CODORNIZ*

Description

Datos de un muestreo realizado en varias provincias y localidades de España, de nidadas de una pequeña especie de ave, la codorniz común.

Usage

```
data(ZX1)
```

Format

Una matriz con 6 columnas: provincia, localidad número de huevos por hembra, edad de la hembra, temperatura media en la zona durante la época de cría, y la disponibilidad de alimento para las hembras (cuyos valores son porcentajes respecto al máximo medido).

ZX3 *VARIABLES MORFOMETRICAS EN CHARACIFORMES*

Description

Datos morfométricos de varias especies de characiformes, como la longitud de la base de la aleta dorsal (M12), altura del cuerpo (M11), etc. Para más detales ver Guisande et al. (2010).

Usage

```
data(ZX3)
```

Format

Una matriz con 31 columnas: datos taxonómicos (Orden, familia, género y especie) y 26 variables morfométricas.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXI1 *MORFOMETRIA DE MOLUSCOS*

Description

Variables morfométricas de la concha de varias especies de moluscos.

Usage

```
data(ZXI1)
```

Format

Una matriz con 5 columnas: especie, peso (en gr) y la longitud, altura y ancho de la concha (en cm).

ZXI10 *INDENTIFICACION DE PECES EN BASE A SU MORFOMETRIA-2*

Description

Datos de medidas del cuerpo de peces de agua dulce. Estos datos se usan para determinar la capacidad de predicción, del modelo logístico multinomial, a la hora de identificar especies, ya que no se usaron en el ejemplo de la función XI8.

Usage

```
data(ZXI10)
```

Format

Una matriz con 15 columnas, donde las tres primeras son la taxonomía de varias especies de peces de agua dulces y las siguientes 12 columnas son medidas del cuerpo.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXI11 *EXPOSICION AL ARSENICO Y MORTALIDAD POR ENFERMEDADES RESPIRATORIAS-1*

Description

Conjunto de datos de un estudio de cohortes, sobre la exposición al arsénico de la industria y las muertes por enfermedades respiratorias, disponible en la base de datos Montana del antiguo paquete *epicalc* (Chongsuvivatwong, 2012).

Usage

```
data(ZXI11)
```

Format

Una matriz con 7 columnas: número de fallecidos en cada grupo (*respdeath*), número de personas por años de exposición (*personyrs*), cociente entre el número de fallecidos y el número de personas por años de exposición * 10.000 (*respdeath.personyrs*), grupos de edad(*agegr*), período del empleo(*period*), comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*).

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

ZXI12 *EXPOSICION AL ARSENICO Y MORTALIDAD POR ENFERMEDADES RESPIRATORIAS-2*

Description

Conjunto de datos de un estudio de cohortes, sobre la exposición al arsénico de la industria y las muertes por enfermedades respiratorias, disponible en la base de datos Montana del antiguo paquete *epicalc* (Chongsuvivatwong, 2012). Estos datos se usan para determinar la capacidad de predicción de la regresión de Poisson estimada en el ejemplo de la función XI10, ya que no se utilizaron para la estima del modelo.

Usage

```
data(ZXI12)
```

Format

Una matriz con 7 columnas: número de fallecidos en cada grupo (*respdeath*), número de personas por años de exposición (*personyrs*), cociente entre el número de fallecidos y el número de personas por años de exposición * 10.000 (*respdeath.personyrs*), grupos de edad(*agegr*), período del empleo(*period*), comienzo del empleo (*start*) y tiempo de exposición al arsénico en años(*arsenic*).

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

ZXI2 *RIQUEZA ACUMULADA DE ESPECIES*

Description

Los datos corresponden a un estudio en el que se van muestreando diferentes sitios y se va registrando el número de especies acumuladas totales considerando los sitios muestreados.

Usage

```
data(ZXI2)
```

Format

Una matriz con 2 columnas con el sitio de muestreo y la riqueza acumulada considerando todos los sitios muestreados anteriormente.

ZXI3 *CRECIMIENTO FITOPLANCTON TOXICO*

Description

Experimento en el que se midió la tasa de crecimiento de una especie de fitoplancton tóxico del género *Alexandrium* en función de la concentración de fosfato del cultivo.

Usage

```
data(ZXI3)
```

Format

Una matriz con 2 columnas: tasa de crecimiento en d^{-1} y concentración de fosfato en μM.

References

Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* 3: 131-139.

ZXI4 *CRECIMIENTO EN LA ANCHOA*

Description

Relación entre la longitud y la edad de individuo, cuantificada por medio el conteo de los anillos de los otolitos, en la anchoa (*Engraulis encrasicolus*), en el Mar de Alborán y en el estrecho de Sicilia.

Usage

```
data(ZXI4)
```

Format

Una matriz con 3 columnas: edad (en días), longitud (en cm) y zona.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A. & Kallianiotis, A. (2004). Linking habitat conditions and growth in the European anchovy (*Engraulis encrasicolus*). *Fisheries Research*, 68: 9-19.

ZXI5 *ABUNDANCIA DE FITOPLANCTON*

Description

Evolución de la abundancia de una especie fitoplanctónica en condiciones de cultivo a lo largo del tiempo.

Usage

```
data(ZXI5)
```

Format

Una matriz con 2 columnas: días de cultivo y abundancia en células por mililitro.

References

Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* 3: 131-139.

ZXI6 *PRODUCCION PRIMARIA EN ZONAS DE AFLORAMIENTO*

Description

Datos de clorofila, afloramiento, estabilidad, tamaño de la plataforma, temperatura y nutrientes de diferentes zonas costeras del mundo.

Usage

```
data(ZXI6)
```

Format

Una matriz con 12 columnas: área de afloramiento, latitud, longitud, concentración de clorofila (mg^{-3}), porcentaje de plataforma continental en la celda, transporte de Ekman ($m^3km^{-1}s^{-1}$) que es una medida de afloramiento, temperatura, turbulencia (m^3s^{-1}), estabilidad de la columna de agua ($cicloss^{-1}$) y las concentraciones de fosfatos, nitratos y silicatos en μMl^{-1}.

References

Patti, B., Guisande, C., Vergara, A.R., Riveiro, I., Maneiro, I., Barreiro, A., Bonanno, A., Buscaino, G., Cuttitta, A., Basilone, G. & Mazzola, S. (2008) Factors responsible for the differences in satellite-based chlorophyll a concentration between the major global upwelling areas. *Estuarine, Coastal and Shelf Science*, 76, 775-786.

ZXI7 *PACIENTES CON CANCER-1*

Description

Datos de pacientes a los que se les ha diagnosticado un tipo de cáncer, así como de personas que no lo tienen.

Usage

```
data(ZXI7)
```

Format

Una matriz con 5 columnas: si la persona padece o no padece cancer, edad, sexo, estado civil y si se ha detectado sangre en la orina.

ZXI8 *PACIENTES CON CANCER-2*

Description

Datos de pacientes a los que se les ha diagnosticado un tipo de cáncer, así como de personas que no lo tienen. Estos datos se usan para determinar la capacidad de predicción del modelo logístico binomial estimado en el ejemplo de la función XI6, ya que no se utilizaron para la estima del modelo.

Usage

```
data(ZXI8)
```

Format

Una matriz con 5 columnas: si la persona padece o no padece cancer, edad, sexo, estado civil y si se ha detectado sangre en la orina.

ZXI9 *INDENTIFICACION DE PECES EN BASE A SU MORFOMETRIA-1*

Description

Datos de medidas del cuerpo de peces de agua dulce, las cuales se usan para identificar las especies utilizando como técnica estadística la regresión logística.

Usage

```
data(ZXI9)
```

Format

Una matriz con 15 columnas, donde las tres primeras son la taxonomía de varias especies de peces de agua dulces y las siguientes 12 columnas son medidas del cuerpo.

Source

http://www.ipez.es/ipez/index_country/index.html.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXII1 *INFLUENCIA DEL CONSEJO MEDICO PARA DEJAR DE FUMAR*

Description

Los datos corresponden a un estudio de Silagy & Ketteridge (1997) sobre el efecto que tiene el consejo médico a la hora de dejar de fumar. En el archivo de datos, la variable *event.e* hace referencia al número de personas que dejaron el tabaco por recomendación del médico, *n.e* el número total de personas del grupo a las que el médico recomendó dejar de fumar y, *event.c* y *n.c* son las personas que dejaron de fumar y el total de personas que estaban en el grupo control, a las que el médico no les recomendó dejar el tabaco, respectivamente.

Usage

```
data(ZXII1)
```

Format

Una matriz con 5 columnas: estudio, personas que dejaron de fumar por recomendación del médico (*event.e*), total de personas a las que el médico recomendó dejar de fumar (*n.e*), total de personas que dejaron de fumar en el grupo control (*event.c*) y total de personas en el grupo control (*n.c*). En el grupo control el médico no recomendó a nadie dejar de fumar.

References

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

ZXIII1 *CRITERIOS DE VALORACION DE CANDIDATOS A UN PUESTO DE TRABAJO*

Description

Criterios de valoración, en una escala de 0 a 10, de 48 candidatos que optan a un puesto de trabajo.

Usage

```
data(ZXIII1)
```

Format

Una matriz con 15 columnas con código del candidato en la primera columna y los criterios de valoración que son: carta de motivación, presentación, nivel de estudios, simpatía, confianza en si mismo, lucidez, honestidad, sentido comercial, experiencia carisma, ambición, comprensión, potencial, motivación para el puesto y adecuación.

ZXIII2 *PARAMETROS DEMOGRAFICOS DE PAISES*

Description

Parámetros demográficos de 57 países de Europa, África y América.

Usage

```
data(ZXIII2)
```

Format

Una matriz con 11 columnas: continente, país, esperanza de vida masculina y femenina al nacer (en años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el producto interior bruto per cápita (en miles de dólares anuales) y la tasa de alfabetización de hombres y mujeres (en porcentaje) en el año 2000. Los datos fueron obtenidos del Banco Mundial (http://datos.bancomundial.org/).

ZXIII3 *MORFOMETRIA DE FAMILIAS DE TIBURONES*

Description

Datos biométricos de varias especies pertenecientes a cuatro familias de tiburones.

Usage

```
data(ZXIII3)
```

Format

Una matriz con 24 columnas: la familia y el resto de columnas son medidas morfométricas.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXIII4 *CALIDAD DEL SUELO Y PARAMETROS DE CRECIMIENTO Y REPRODUCCION EN PLANTAS*

Description

Datos sobre calidad de un suelo y parámetros que miden el crecimiento y reproducción en plantas.

Usage

```
data(ZXIII4)
```

Format

Una matriz con 9 columnas: sitio de muestreo, un grupo de variables que nos da información sobre calidad del suelo que son: contenido en agua (en porcentaje del peso del suelo), nitrógeno, fosfato y una serie de oligoelementos minerales (en micro moles por gramo de suelo), y otro grupo de variables nos da una idea sobre condición de las plantas, y son una serie de parámetros independientes relacionados con el crecimiento y la reproducción: cobertura foliar (en m^2), la longitud promedio de entrenudos (en cm) y el porcentaje de semillas viables del total producidas.

ZXIV1 *PARAMETROS DEMOGRAFICOS DE REGIONES DE ESPAÑA*

Description

Datos demográficos del año 2010 de las 19 regiones o comunidades autónomas de España publicados por el Instituto Nacional de Estadística (http://www.ine.es).

Usage

```
data(ZXIV1)
```

Format

Una matriz con 10 columnas: región, provincia, número de hijos por madre, edad media de la madre cuando tiene su primer hijo, porcentaje medio de mujeres no casadas con hijos, la esperanza de vida de hombres y mujeres, y el número de nacimientos, mortalidad de la población y mortalidad infantil (menores de 1 año), los tres último parámetros demográficos por cada 1000 habitantes.

Source

http://www.ine.es

ZXIV2 *NUTRIENTES Y ABUNDANCIA DE FITOPLANCTON EN LAGOS DE COLOMBIA*

Description

Relación de la concentración de nutrientes y abundancia de fitoplancton, en varias estaciones de muestreo en diferentes lagos de Colombia.

Usage

```
data(ZXIV2)
```

Format

Una matriz con 15 columnas: región de Colombia, lago, las concentraciones de nutrientes y la abundancia de diversos grupos del fitoplancton.

ZXIV3 *DATOS DE CITOMETRO DE FLUJO*

Description

Características de la morfología de especies fitoplanctónicas obtenidas de un citómetro de flujo.

Usage

```
data(ZXIV3)
```

Format

Una matriz con 8 columnas: especie, rugosidad, tamaño y emisión a diferentes longitudes de onda.

ZXIV4 *PIGMENTOS DE ALGAS FITOPLANCTONICAS*

Description

Concentración relativa de 19 pigmentos, en relación con la concentración de clorofila a, de especies del fitoplancton que pertenecen a diferentes clases de algas.

Usage

```
data(ZXIV4)
```

Format

Una matriz con 21 columnas: clase, especie y los 19 pigmentos.

ZXV1 *FACTOR DE CONDICION DE LA ANCHOA*

Description

Valores medios mensuales del factor de condición de la anchoa (*Engraulis encrasicolus*), temperatura (en degree C), clorofila (en mgm^{-3}) y estabilidad de la columna de agua (en $m^3 s^{-3}$) a lo largo de varios años en el estrecho de Sicilia (Basilone et al., 2006).

Usage

```
data(ZXV1)
```

Format

Una matriz con 6 columnas: año, mes, factor de condición, clorofila, temperatura y estabilidad de la columna de agua.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A., Vergara, A.R. & Maneiro, I. (2006) Effect of habitat conditions on reproduction of the European anchovy (*Engraulis encrasicolus*) in the Strait of Sicily. *Fisheries Oceanography*, 15: 271-280.

ZXV2 *RANGO DE MAREAS Y ANGULO DEL SOL Y LA LUNA*

Description

Angulos entre el sol y la luna y la diferencia en metros entre pleamar y bajamar para 42º 8´Norte y 14º 13´Oeste del 1/1/2004 al 14/3/2004.

Usage

```
data(ZXV2)
```

Format

Una matriz con 3 columnas: fecha, ángulo entre el sol y la luna, y la diferencia en metros entre marea alta y marea baja.

ZXV3 *PESO DE TERNEROS*

Description

Peso de terneros al nacer y al destete procedentes de diferentes padres y madres.

Usage

```
data(ZXV3)
```

Index

* **II1**
  II1, 4
* **III1**
  III1, 12
* **III2**
  III2, 18
* **III3**
  III3, 20
* **III4**
  III4, 22
* **IV1**
  IV1, 24
* **IV2**
  IV2, 26
* **IV3**
  IV3, 29
* **IV4**
  IV4, 31
* **IX10**
  IX10, 38
* **IX11**
  IX11, 42
* **IX12**
  IX12, 45
* **IX13**
  IX13, 48
* **IX1**
  IX1, 33
* **IX2**
  IX2, 52
* **IX3**
  IX3, 57
* **IX4**
  IX4, 66
* **IX5**
  IX5, 74
* **IX6**
  IX6, 78
* **IX7**
  IX7, 84
* **IX8**
  IX8, 89
* **IX9**
  IX9, 100
* **PAR**
  PAR, 105
* **VI1**
  VI1, 108
* **VI2**
  VI2, 115
* **VI3**
  VI3, 120
* **VII1**
  VII1, 125
* **VII2**
  VII2, 134
* **VII3**
  VII3, 141
* **VII4**
  VII4, 145
* **VII5**
  VII5, 150
* **VIII1**
  VIII1, 154
* **VIII2**
  VIII2, 162
* **X1**
  X1, 170
* **X2**
  X2, 176
* **X3**
  X3, 181
* **XI10**
  XI10, 196
* **XI11**
  XI11, 200
* **XI1**
  XI1, 186

* **XI2**
  XI2, 202
* **XI3**
  XI3, 220
* **XI4**
  XI4, 224
* **XI5**
  XI5, 226
* **XI6**
  XI6, 235
* **XI7**
  XI7, 242
* **XI8**
  XI8, 243
* **XI9**
  XI9, 248
* **XII1**
  XII1, 250
* **XII2**
  XII2, 255
* **XII3**
  XII3, 258
* **XII4**
  XII4, 261
* **XII5**
  XII5, 264
* **XIII1**
  XIII1, 267
* **XIII2**
  XIII2, 275
* **XIII3**
  XIII3, 283
* **XIII4**
  XIII4, 287
* **XIV1**
  XIV1, 293
* **XIV2**
  XIV2, 297
* **XIV4**
  XIV4, 307
* **XIV5**
  XIV5, 310
* **XV1**
  XV1, 313
* **XV2**
  XV2, 328
* **XV3**
  XV3, 334
* **XV4**
  XV4, 337
* **XV5**
  XV5, 340
* **XV6**
  XV6, 341
* **XV7**
  XV7, 353
* **XV8**
  XV8, 355
* **XV9**
  XV9, 365
* **XVI1**
  XVI1, 368
* **XVI2**
  XVI2, 377
* **ZII1**
  ZII1, 383
* **ZIV1**
  ZIV1, 384
* **ZIV3**
  ZIV3, 384
* **ZIX10**
  ZIX10, 385
* **ZIX11**
  ZIX11, 385
* **ZIX1**
  ZIX1, 384
* **ZIX3**
  ZIX3, 386
* **ZIX5**
  ZIX5, 386
* **ZIX6**
  ZIX6, 387
* **ZIX7**
  ZIX7, 387
* **ZIX9**
  ZIX9, 388
* **ZVI1**
  ZVI1, 388
* **ZVII1**
  ZVII1, 389
* **ZVII2**
  ZVII2, 389
* **ZVII3**
  ZVII3, 390
* **ZVII4**
  ZVII4, 390

* **ZVII5**
ZVII5, 391
* **ZVIII1**
ZVIII1, 391
* **ZVIII2**
ZVIII2, 392
* **ZX1**
ZX1, 392
* **ZX3**
ZX3, 393
* **ZXI10**
ZXI10, 394
* **ZXI11**
ZXI11, 394
* **ZXI12**
ZXI12, 395
* **ZXI1**
ZXI1, 393
* **ZXI2**
ZXI2, 396
* **ZXI3**
ZXI3, 396
* **ZXI4**
ZXI4, 397
* **ZXI5**
ZXI5, 397
* **ZXI6**
ZXI6, 398
* **ZXI7**
ZXI7, 398
* **ZXI8**
ZXI8, 399
* **ZXI9**
ZXI9, 399
* **ZXII1**
ZXII1, 400
* **ZXIII1**
ZXIII1, 400
* **ZXIII2**
ZXIII2, 401
* **ZXIII3**
ZXIII3, 401
* **ZXIII4**
ZXIII4, 402
* **ZXIV1**
ZXIV1, 402
* **ZXIV2**
ZXIV2, 403
* **ZXIV3**
ZXIV3, 403
* **ZXIV4**
ZXIV4, 403
* **ZXV1**
ZXV1, 404
* **ZXV2**
ZXV2, 404
* **ZXV3**
ZXV3, 405

adf.test, *317*
Anova, *61*, *70*, *75*, *81*, *93*
aov, *87*
Arima, *314*, *316*
Arrows, *278*, *285*, *290*, *301*, *345*, *359*
assoc, *129*, *137*
auto.arima, *314*, *316*

bart_spher, *278*, *345*
beanplot, *36*, *40*, *44*, *46*, *54*, *62*, *75*, *81*, *87*, *93*, *103*
Box.test, *316*
boxM, *371*, *379*
bptest, *190*, *229*

ca, *270*, *359*
calc.relimp, *227*, *229*
cancor, *288*, *290*
candisc, *298*, *301*, *345*, *359*
chull, *278*, *285*, *290*, *301*, *345*, *359*
coh, *328, 329*
corrplot, *182*, *184*

dwtest, *190*, *229*

forecast, *314*, *316*, *319*
forest.meta, *250, 251*, *255*, *264, 265*
friedman.test, *46*
funnel.meta, *261, 262*

glht, *61*
glm, *198*, *238*, *336*, *338*
gpar, *127*, *136*, *142*, *146*, *151*, *156*, *164*

hclust, *307*, *309*, *343*, *357*
heatmap, *311*
hier.part, *229*

II1, 4, *13*, *24*, *35*

III1, 12
III2, 18
III3, 20
III4, 22
interaction.plot, *61*, *70*, *81*, *87*
IV1, 24, *27*
IV2, 26
IV3, 29
IV4, 31, *35*
IX1, *32*, 33, *122*
IX10, 38
IX11, 42
IX12, 45
IX13, 48
IX2, *32*, 52, *75*
IX3, *32*, 57
IX4, *32*, *49*, 66, *378*
IX5, 74, *81*, *87*
IX6, 78
IX7, 84
IX8, 89
IX9, 100

KMO, *278*, *345*
kruskal.test, *40*
ks.test, *103*
kurtosis, *36*, *54*, *62*, *70*, *93*

labbe.metabin, *258, 259*
labeling_border, *127*, *136*, *142*, *146*, *151*, *156*, *164*
lda, *301*, *345*, *359*
levene.test, *36*, *54*, *62*, *70*, *93*
lillie.test, *61*, *70*, *93*, *190*, *208*, *229*, *316*, *336*, *338*, *370*, *378*
lines, *314*, *316*
lm, *61*, *70*, *75*, *81*, *87*, *370*, *378*

Manova, *370*, *379*
manova, *370*, *379*
metabin, *250, 251*, *255*, *258, 259*, *261, 262*, *264, 265*
metacum, *255*
metainf, *265*
metaMDS, *283*, *285, 286*
multinom, *246*

na.approx, *329*
ndiffs, *314*, *316*

nls, *208*, *336*, *338*

pairs, *182*, *184*
panel.cor, *184*
panel.hist, *184*
panel.reg, *184*
panels, *229*
panels.diag, *229*
PAR, *34*, *39*, *42*, *45, 46*, *48*, *53*, *58–60*, *67, 68*, *75*, *79–81*, *85, 86*, *91, 92*, *101*, 105, *171*, *177*, *182*, *187*, *196*, *203*, *267*, *275*, *283*, *288*, *293*, *298*, *307*, *311*, *313*, *328*, *335*, *338*, *343*, *357*
plot.ca, *270*
plot.cancor, *298*, *301*, *345*, *359*
plot.default, *106*, *187*, *203*, *335*, *338*
plot.forecast, *314*, *316*
plot.ts, *197, 198*
prcomp, *275*, *278*, *343*, *345*

qda, *301*, *345*, *359*

rcor.test, *174*, *179*, *183*, *345*
rpart, *238, 239*, *246*, *293*, *295*
rpart.plot, *239*, *246*, *293*, *295*

scatterplot, *275*, *278*, *280*, *283*, *285, 286*, *288*, *290*, *297, 298*, *301, 302*, *342, 343*, *345*, *356, 357*, *359*
seasonplot, *314*, *316*
shapiro.test, *36*, *54*, *61*, *70*, *93*, *316*, *370*, *378*
skewness, *36*, *54*, *62*, *70*, *93*
smc, *183*
spacings, *127*, *135*, *142*, *146*, *151*, *155*, *163*
spec.ar, *328, 329*
spec.pgram, *328, 329*
spectrum, *328, 329*
stepAIC, *198*, *238*, *246*, *336*, *338*

T.aov, *48, 49*
ts, *313*, *316*, *328, 329*
tsdisplay, *314*, *316*

unit, *127*, *135, 136*, *142, 143*, *146*, *151*, *155, 156*, *163, 164*

VI1, 108, *116*, *121*
VI2, 115
VI3, *35*, 120

vif, *222*, *225*, *229*, *278*, *345*, *371*, *379*
vifcor, *222*, *229*
vifstep, *222*, *229*
VII1, 125, *157*
VII2, 134, *164*
VII3, 141
VII4, 145
VII5, 150, *158*
VIII1, *126*, *135*, *142*, *146*, *151*, *153*, 154, *163*, *165*
VIII2, 162

wilcox.test, *44*, *103*

X1, 170, *178*, *179*
X2, 176
X3, 181
XI1, 186, *204*, *226*, *228*, *232*, *234*
XI10, 196, *200–202*, *395*
XI11, 200
XI2, 202, *335*
XI3, 220, *225*, *227*
XI4, 224
XI5, 226, *236*, *244*
XI6, 235, *242*, *243*, *399*
XI7, 242
XI8, 243, *248*, *249*, *394*
XI9, *244*, *247*, 248
XII1, 250, *255*, *256*, *258*, *259*, *262*, *265*
XII2, 255
XII3, 258
XII4, 261
XII5, 264
XIII1, 267
XIII2, 275
XIII3, 283
XIII4, 287
XIV1, *244*, 293
XIV2, 297
XIV4, 307
XIV5, 310
XV1, 313
XV2, 328
XV3, 334, *338*, *339*
XV4, 337
XV5, 340
XV6, 341
XV7, *61*, 353, *366*, *370*
XV8, 355
XV9, 365, *370*
XVI1, 368, *378*
XVI2, 377

ZII1, 383
ZIV1, 384
ZIV3, 384
ZIX1, 384
ZIX10, 385
ZIX11, 385
ZIX3, 386
ZIX5, 386
ZIX6, 387
ZIX7, 387
ZIX9, 388
ZVI1, 388
ZVII1, 389
ZVII2, 389
ZVII3, 390
ZVII4, 390
ZVII5, 391
ZVIII1, 391
ZVIII2, 392
ZX1, 392
ZX3, 393
ZXI1, 393
ZXI10, 394
ZXI11, 394
ZXI12, 395
ZXI2, 396
ZXI3, 396
ZXI4, 397
ZXI5, 397
ZXI6, 398
ZXI7, 398
ZXI8, 399
ZXI9, 399
ZXII1, 400
ZXIII1, 400
ZXIII2, 401
ZXIII3, 401
ZXIII4, 402
ZXIV1, 402
ZXIV2, 403
ZXIV3, 403
ZXIV4, 403
ZXV1, 404
ZXV2, 404
ZXV3, 405