

# Package ‘datana’

January 6, 2026

**Title** Datasets and Functions to Accompany Analisis De Datos Con R

**Version** 1.1.6

**Description** Datasets and functions to accompany the book 'Analisis de datos con el programa estadístico R: una introducción aplicada' by Salas-Eljatib (2021, ISBN: 9789566086109).  
The package helps carry out data management, exploratory analyses, and model fitting.

**License** GPL (>= 3)

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.3.3

**Depends** R (>= 3.5)

**LazyData** true

**LazyDataCompression** xz

**Imports** ggplot2, graphics, Hmisc, methods, scales, stats, utils

**Suggests** lattice, testthat (>= 3.0.0)

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Christian Salas-Eljatib [aut, cre] (ORCID:  
<<https://orcid.org/0000-0002-8468-0829>>),  
Nicolás Campos [ctb] (ORCID: <<https://orcid.org/0009-0001-9534-9808>>,  
since 2025),  
Joaquín Riquelme [ctb] (up to 2020),  
Nicolás Pino [ctb] (up to 2020)

**Maintainer** Christian Salas-Eljatib <csejlatib@gmail.com>

## Contents

aboutrsq . . . . .	4
aboutrsq2 . . . . .	5
airnyc . . . . .	6
airnyc2 . . . . .	7
annualppCities . . . . .	8

annualppCities2 . . . . .	8
assigncl . . . . .	9
bears . . . . .	10
bears2 . . . . .	12
bearscomp . . . . .	13
bearscomp2 . . . . .	14
beetles . . . . .	15
beetles2 . . . . .	16
cameratrapp . . . . .	17
cameratrapp2 . . . . .	18
carAccidents . . . . .	19
caribou . . . . .	19
caribou2 . . . . .	20
casen . . . . .	21
cdf . . . . .	22
chicksw . . . . .	23
chicksw2 . . . . .	24
co2temp . . . . .	24
contrast . . . . .	27
corkoak . . . . .	29
corkoak2 . . . . .	30
deleteLeft . . . . .	31
deleteRight . . . . .	32
descstat . . . . .	33
education . . . . .	35
election . . . . .	36
election2 . . . . .	36
endfid2 . . . . .	37
eucaleaf . . . . .	38
eucaleaf2 . . . . .	39
eucaleafAll . . . . .	40
eucaleafAll2 . . . . .	41
extractLeft . . . . .	42
extractRight . . . . .	43
fdamage . . . . .	44
fdamage2 . . . . .	44
fertiliza . . . . .	45
fertiliza2 . . . . .	45
ficdiamgr . . . . .	46
ficdiamgr2 . . . . .	47
findColumn.byname . . . . .	48
fishgrowth . . . . .	49
fishgrowth2 . . . . .	50
forestfire . . . . .	50
forestfire2 . . . . .	52
gasoline . . . . .	53
gasoline2 . . . . .	54
gdpcap . . . . .	55

gmean . . . . .	55
hgrdfir . . . . .	56
hgrdfir2 . . . . .	57
histbxp . . . . .	58
hmean . . . . .	61
idahohd . . . . .	62
idahohd2 . . . . .	63
imacec2 . . . . .	64
interp . . . . .	65
kurto . . . . .	67
landcover . . . . .	68
landcover2 . . . . .	69
largetrees . . . . .	71
largetrees2 . . . . .	72
leafw2 . . . . .	73
lifexpect . . . . .	73
llancahue . . . . .	75
llancahue2 . . . . .	76
lrt . . . . .	77
lrt.glm . . . . .	78
maple . . . . .	79
moda . . . . .	80
modresults . . . . .	81
obspredplot . . . . .	82
papersdocstu . . . . .	84
pesohojas . . . . .	85
plotrend . . . . .	85
predstat . . . . .	87
predstatmod . . . . .	89
presenceIce . . . . .	91
president . . . . .	92
pressind . . . . .	93
primarias . . . . .	94
pspLlancahue . . . . .	95
pspruca . . . . .	96
pspruca2 . . . . .	97
ptaeda . . . . .	98
ptaeda2 . . . . .	98
pvalt . . . . .	99
pvalz . . . . .	100
qqgauss . . . . .	101
rainfallCA . . . . .	102
rankmod . . . . .	103
raulihg . . . . .	104
raulihg2 . . . . .	105
rendesc2 . . . . .	106
simce2 . . . . .	107
simmeind . . . . .	108

singleupp . . . . .	109
skewn . . . . .	110
sludge . . . . .	111
sludge2 . . . . .	112
smoothfit . . . . .	113
snaspe . . . . .	114
snaspe2 . . . . .	115
soiltreat . . . . .	116
soiltreat2 . . . . .	117
spataustria . . . . .	118
tabtexanova . . . . .	119
tabtexdescstat . . . . .	121
timeserplot . . . . .	123
treevol . . . . .	125
treevol2 . . . . .	126
treevolroble . . . . .	127
treevolroble2 . . . . .	128
upperleft . . . . .	129
vifx . . . . .	130
xyboxplot . . . . .	131
xyhist . . . . .	134
xymultiplot . . . . .	136

---

aboutrsq

*About the R-Squared statistics: the Anscombe quartet dataset*

---

## Description

A dataset that contains four pairs of columns with the same descriptive statistics; however, there is a difference when representing the points through a graph.

## Usage

```
data(aboutrsq)
```

## Format

The data frame contains four variables as follows:

**X1** Integers values that represent X-axis for Y1, Y2 and Y3 column

**Y1** Float values that represent Y-axis for X1 column

**Y2** Float values that represent Y-axis for X1 column

**Y3** Float values that represent Y-axis for X1 column

**X2** Integers values that represent X-axis for Y4 column

**Y4** Float values that represent Y-axis for X2 column

**Source**

Data were assembled by Dr Christian Salas-Eljatib (Santiago, Chile).

**References**

Anscombe FJ. 1973. Graphs in statistical analysis. The American Statistician 27:17-21. doi:10.2307/2682899

**Examples**

```
data(aboutrsq)
head(aboutrsq)
```

---

aboutrsq2

*Sobre el estadístico R2: los datos del cuarteto de Anscombe*

---

**Description**

Dataset que contiene cuatro pares de columnas con la mismos estadísticos descriptivos, sin embargo, si existe diferencia al representar los puntos mediante un gráfico.

**Usage**

```
data(aboutrsq2)
```

**Format**

Variables se describen a continuación::

**X1** Valores enteros que representan el eje X para las columnas Y1, Y2 e Y3

**Y1** Valores flotantes que representan el eje Y para la columna X1

**Y2** Valores flotantes que representan el eje Y para la columna X1

**Y3** Valores flotantes que representan el eje Y para la columna X1

**X2** Valores enteros que representan el eje X para las columnas Y4

**Y4** Valores flotantes que representan el eje Y para la columna X2

**Source**

Datos fueron contribuidos por el Prof. Christian Salas-Eljatib (Universidad de Chile, Santiago, Chile).

**References**

Anscombe FJ. 1973. Graphs in statistical analysis. The American Statistician 27:17-21. doi:10.2307/2682899

### Examples

```
data(aboutrsq2)
head(aboutrsq2)
```

---

airnyc	<i>Airquality data in New York city.</i>
--------	--

---

### Description

Daily air quality measurements in New York, May to September 1973.

### Usage

```
data(airnyc)
```

### Format

Contains 6 variables, as follows:

**ozone** numeric Ozone (ppb).  
**solar** numeric Solar R (lang).  
**wind** numeric Wind (mph).  
**temp** numeric Temperature (degrees F).  
**month** numeric Month (1–12).  
**day** numeric Day of month (1–31).

### Source

The data were obtained from the library *datasets*.

### References

Chambers J, Cleveland W, Kleiner B, Tukey P. 1983. Graphical Methods for Data Analysis. Belmont. CA: Wadsworth.

### Examples

```
data(airnyc)
head(airnyc)
```

---

`airnyc2`*Calidad del aire en la ciudad de Nueva York.*

---

**Description**

Calidad del aire diario medido en New York, de Mayo a Septiembre de 1973.

**Usage**

```
data(airnyc2)
```

**Format**

Contiene 6 variables:

**ozone** Ozono (ppb).

**solar** Solar R (largo).

**wind** Viento (mph).

**temp** Temperatura (grados F).

**month** Mes del año (1–12).

**day** Dia del mes (1–31).

**Source**

Los datos fueron obtenidos desde la librería 'datasets'.

**References**

Chambers J, Cleveland W, Kleiner B, Tukey P. 1983. Graphical Methods for Data Analysis. Belmont. CA: Wadsworth.

**Examples**

```
data(airnyc2)
head(airnyc2)
```

---

annualppCities	<i>Time series of annual precipitations in cities of Chile.</i>
----------------	---

---

**Description**

Data contains annual precipitations in six cities in Chile (Santiago, Talca, Chillán, Temuco, Valdivia, and Puerto Montt) at different years.

**Usage**

```
data(annualppCities)
```

**Format**

The dataframe contains three variables as follows:

**city** Name of city.

**year** Year of registry.

**annual** Value of the annual precipitation of a given year (mm).

**Source**

The data were obtained from <https://explorador.cr2.cl/>.

**Examples**

```
data(annualppCities)
head(annualppCities)
```

---

annualppCities2	<i>Serie de tiempo de precipitaciones anuales en Chile.</i>
-----------------	---

---

**Description**

Data contains annual precipitations in six cities in Chile (Santiago, Talca, Chillan, Temuco, Valdivia, and Puerto Montt) at different years.

**Usage**

```
data(annualppCities2)
```

**Format**

The dataframe contains three variables as follows:

**ciudad** Name of city.

**anho** Year of registry.

**pp.anual** Value of the annual precipitation of a given year (mm).



**Source**

Los datos fueron obtenidos desde <https://explorador.cr2.cl/>.

**Examples**

```
data(annualppCities2)
head(annualppCities2)
```

---

assigncl

---

*Function to assign classes based upon a variable of interest.*


---

**Description**

Assigns class of each observation in a dataframe

**Usage**

```
assigncl(
  data = data,
  variable = variable,
  num.class = 4,
  breaks = NULL,
  wclass = NULL,
  mincl = NULL,
  name.class = NULL
)
```

**Arguments**

data	a dataframe having the variable of interest for each observation.
variable	a character giving the column name of the numeric variable to be used for defining the limits of each class.
num.class	the number of classes to be build. The default is set to 4. Regardless, the percentiles are used to set the limits of each class.
breaks	is a vector having the numbers to be used as breakpoints, by default is set to NULL, therefore the breakpoints will be determined by the num.class.
wclass	a number defining the width or amplitud of the classes. By default is set to NULL, otherwise, the width is determined by the previous explained options, such as, breaks or num.classes.
mincl	the number of the minimum class to be used. By default is set to NULL, otherwise, this option is used to define the breaks.
name.class	a character giving the column name of the new class variable. By default is set to NULL, then, the column name will be a composite-name merging the character provided in variable followed by ".class". Otherwise, will be name.class.

## Details

The function assign a class or category to a random variable of interest. Several alternatives are implemented to define the way on which the allocation to a respective class is carried out.

## Value

The main output is the data including a new column having the created class variable.

## Author(s)

Christian Salas-Eljatib and Marcos Marivil.

## References

- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de roble-laurel-lingue. Bosque 23(2):81–92. doi:[10.4067/S071792002002000200009](https://doi.org/10.4067/S071792002002000200009).

## Examples

```
# The data
library(datana)
maple
# Example 1
graphics::boxplot(maple$dbh)
df<-assigncl(data=maple,variable="dbh")
head(df)
table(df$dbh.class)
# Example 2, changing the number of classes
df<-assigncl(data=maple,variable="dbh",num.class=5)
table(df$dbh.class)
tapply(df$dbh,df$dbh.class,range)
# Example 3, fixing the breakpoints
df<-assigncl(data=maple,variable="dbh",
             breaks = c(25.60,36.44,40.12,42.3))
table(df$dbh.class)
tapply(df$dbh,df$dbh.class,range)
# Example 4, giving the amplitude
# of the classes
df<-assigncl(data=llancahue,variable="dbh",wclass = 5)
table(df$dbh.class)
tapply(df$dbh,df$dbh.class,range)
```

---

bears

*Age and physical measurement data for wild bears*


---

## Description

Wild bears were anesthetized, and their bodies were measured and weighed. One goal of the study was to make a table (or perhaps a set of tables) for people interested in estimating the weight of a bear based on other measurements. Notice that there are missing values for some of the variables.

**Usage**

```
data(bears)
```

**Format**

Contains individual-level variables, as follows:

**id** Bear id

**age** Age in total number of months.

**month** Month number within a given year.

**sex** 1 =male, 2 = female.

**headL** Length of head, in cm.

**headW** Width of head, in cm.

**neckG** Girth of neck, in cm.

**length** Body length, in cm.

**chestG** Girth of chest, in cm.

**weight** body weight, in kg.

**obs** Temporal observation number for bear.

**name** Name given to bear.

**sex.name** Sex factor (male or female).

**Source**

According to Prof. Timothy Gregoire at Yale University (New Haven, CT, USA), the data set was supplied by Gary Alt.

**References**

Entertaining references are in Reader's Digest April, 1979, and Sports Afield September, 1981.

**Examples**

```
data(bears)
head(bears)
table(bears$sex.name)
boxplot(headL~sex.name, data=bears)
```

---

bears2*Edad y características biométricas de osos salvajes*

---

**Description**

Los osos salvajes fueron anestesiados y sus cuerpos medidos. Uno de los objetivos del estudio fue hacer una tabla (o quizás un conjunto de tablas) para las personas interesadas en estimar el peso de un oso basandose en otras medidas. Observe que faltan valores para algunas de las variables.

**Usage**

```
data(bears2)
```

**Format**

Contiene variables de nivel individual, como se describen a continuación:

**id** Identificador del oso.

**edad** edad en meses

**mes** identificador del mes, dentro del año.

**sexo** Indicador del sexo del animal (1 = macho, 2 = hembra)

**cabezaL** longitud de la cabeza, en cm

**cabezaA** ancho de la cabeza, en cm

**cuelloP** circunferencia del cuello, en cm

**largo** longitud del cuerpo, en cm

**pechoG** circunferencia del pecho, en cm

**peso** peso corporal, en kg

**obs** número de observación temporal para el oso

**nombre** nombre dado al oso

**sexo.nombre** factor del sexo del animal (macho o hembra)

**Source**

Segun el Prof. Timothy Gregoire de Yale University (New Haven, CT, USA), los datos fueron cedidos por Gary Alt. Minitab, Inc. La descripcion de los datos fue dada por él.

**References**

Algunas referencias generales estan en el Reader's Digest de Abril, 1979, y Sports Afield de Septiembre, 1981.

**Examples**

```
data(bears2)
head(bears2)
table(bears2$sexo.nombre)
boxplot(cabezaL~sexo.nombre, data=bears2)
```

---

bearscomp	<i>Age and physical measurement data for wild bears (complete cases)</i>
-----------	--

---

## Description

Wild bears were anesthetized, and their bodies were measured and weighed. One goal of the study was to make a table (or perhaps a set of tables) for people interested in estimating the weight of a bear based on other measurements. The current version of without missing values

## Usage

```
data(bearscomp)
```

## Format

Individual-level variables, as follows:

**id** Bear identifier.  
**age** Age in total number of months.  
**month** Month number within a given year.  
**sex** Sex code: 1 =male, 2 = female.  
**headL** Length of head, in cm.  
**headW** Width of head, in cm.  
**neckG** Girth of neck, in cm.  
**length** Body length, in cm.  
**chestG** Girth of chest, in cm.  
**weight** Body weight, in kg.  
**obs** Temporal observation number for bear.  
**name** name given to bear  
**sex.name** Sex factor (male or female).

## Source

According to Prof. Timothy Gregoire at Yale University (New Haven, CT, USA), the data set was supplied by Gary Alt.

## References

Entertaining references are in Reader's Digest April, 1979, and Sports Afield September, 1981.

## Examples

```
data(bearscomp)
head(bearscomp)
table(bearscomp$sex.name)
boxplot(headL~sex.name, data=bearscomp)
```

bearscomp2

*Edad y características biométricas de osos salvajes (solo datos completos, i.e., sin valores vacíos)***Description**

Los osos salvajes fueron anestesiados y sus cuerpos medidos. Uno de los objetivos del estudio fue hacer una tabla (o quizás un conjunto de tablas) para las personas interesadas en estimar el peso de un oso basándose en otras medidas. Esta dataframe es igual que "bears" pero sin valores perdidos.

**Usage**

```
data(bearscomp2)
```

**Format**

Contiene variables de nivel individual, como se describen a continuación:

**id** Identificador del oso.

**edad** edad en meses.

**mes** identificador del mes, dentro del año.

**sexo** 1 = macho, 2 = hembra.

**cabezaL** longitud de la cabeza, en cm.

**cabezaA** ancho de la cabeza, en cm.

**cuelloP** circunferencia del cuello, en cm.

**largo** longitud del cuerpo, en cm.

**pechoG** circunferencia del pecho, en cm.

**peso** peso corporal, en kg.

**obs** número de observación temporal para el oso.

**nombre** nombre dado al oso.

**sexo.nombre** factor del sexo del animal (macho o hembra)

**Source**

Según el Prof. Timothy Gregoire de Yale University (New Haven, CT, USA), los datos fueron cedidos por Gary Alt. Minitab, Inc. La descripción de los datos fue dada por él.

**References**

Algunas referencias generales están en el Reader's Digest de Abril, 1979, y Sports Afield de Septiembre, 1981.

**Examples**

```
data(bearscomp2)
head(bearscomp2)
table(bearscomp2$sexo.nombre)
boxplot(cabezaL~sexo.nombre, data=bearscomp2)
```

beetles

*Population density growth of beetles***Description**

Temporal measurements of density of beetles (*Tribolium confusum*) growing in different controlled environments.

**Usage**

```
beetles
```

**Format**

**days** Number of days.

**diet** The quantities of flour (in grams) of the environments where the beetles were growing. Six levels of the factor diet.

**type** The various stage of beetles, *i.e.*, eggs, larvae, pupae, and adults.

**density** The number of insects per environment.

**Source**

Data from Table No. 1, page 116, of Chapman (1928). Series of experiments under controlled conditions in which flour beetles (*Tribolium confusum*) are kept in environments of known size. The period from egg to adult is approximately forty days at 27C degrees. The data were entered by Miss Yamara Arancibia, a former student of Prof. Christian Salas-Eljatib.

**References**

- Chapman RN. 1928. The quantitative analysis of environmental factors. Ecology 9(2):111-122. doi:10.2307/1929348

**Examples**

```
data(beetles)
table(beetles$type)
name.diet<-unique(beetles$diet)
num.diet<-length(name.diet)
##Time series plot
#first, some computation
alys<-with(beetles,tapply(density,list(as.factor(days),as.factor(diet)),sum))
```

```

out<-as.data.frame(als)
out$time<-row.names(out)
head(out)
#Figure 1 of the paper
matplot(out[, "time"], out[, 1:num.diet], las=1, type=c("b"), pch=1,
        xlab="Time in days", ylab="Total individuals")
legend("topleft", legend = name.diet, title = "Diet (gr)",
       col = 1:6, lty = 1:6, pch = 1)

```

beetles2

*Crecimiento poblacional de escarabajos*

## Description

Mediciones temporales de densidad de escarabajos (*Tribolium confusum*) creciendo en diferentes ambientes controlados.

## Usage

```
beetles2
```

## Format

**días** Número de días.

**dieta** La cantidad de harina (en gramos) de ambientes donde crecen los escarabajos. Seis niveles del factor dieta.

**tipo** Estados de desarrollo de los escarabajos, *i.e.*, huevos, larvas, pupas, y adultos.

**densidad** Número total de individuos por ambiente de crecimiento.

## Source

Datos del Cuadro No. 1, page 116, de Chapman (1928). Serie de experimentos bajo condiciones controladas donde escarabajos (*Tribolium confusum*) se mantienen en ambientes de tamaño conocido. El periodo desde huevo a adulto es de aproximadamente de cuarenta días a 27 grados Celsius. Los datos fueron digitados por la Srta. Yamara Arancibia, una estudiante del Prof. Christian Salas-Eljatib.

## References

- Chapman RN. 1928. The quantitative analysis of environmental factors. Ecology 9(2):111-122. doi:10.2307/1929348



### Examples

```
data(beetles2)
table(beetles2$tipo)
nom.dieta<-unique(beetles2$dieta)
num.dieta<-length(nom.dieta)
##Grafico de serie de tiempo
#primero algunos calculos
alys<-with(beetles2,tapply(
  densidad,list(as.factor(dias),as.factor(dieta)),sum)
)
out<-as.data.frame(alys)
out$tiempo<-row.names(out)
head(out)
##Figura 1 del paper
matplot(out[, "tiempo"], out[,1:num.dieta], las=1, type=c("b"),pch=1,
  xlab="Tiempo en dias",ylab="Densidad de individuos")
legend("topleft", legend = nom.dieta, title = "Dieta (gr)",
  col = 1:6, lty = 1:6, pch = 1)
```

---

cameratrapp

---

*Camera trap data on mammals in Ruaha National Park, southern Tanzania.*


---

### Description

Dataset contains 14604 observations and sampling was carried out for two months during the dry season of 2013 and two months during the wet season of 2014. Each camera station is associated with a randomly placed camera and a trail-based camera, with the aim of comparing communities resulting from the two camera trap placement strategies.

### Usage

```
data(cameratrapp)
```

### Format

Contains 6 variables, as follows:

- reference** Number of observation od datasets.
- placement** Type of "placement" placed in each station (random or trail).
- season** Season where were made the samplings.
- station** Station where were collected the data.
- specie** Name of specie medium to large terrestrial mammals.
- date.time** The date and time of each photographic event is also given.

### Source

The data were provided by Dr Jeremy Cusack.

## References

- Cusack J, Dickman A, Rowcliffe M, Carbone C, Macdonald D, Coulson T. 2016. Random versus game trail-based camera trap placement strategy for monitoring terrestrial mammal communities. PLoS ONE 10(5): e0126373.

## Examples

```
data(cameratrapp)
head(cameratrapp)
```

---

cameratrapp2	<i>Cameras trampa de mamiferos en el parque nacional Ruaha, en el sur de Tanzania</i>
--------------	---

---

## Description

Contains information of Camera trap data on medium to large terrestrial mammals collected at 54 camera stations in Ruaha National Park, southern Tanzania. Dataset contains 14604 observations and sampling was carried out for two months during the dry season of 2013 and two months during the wet season of 2014. Each camera station is associated with a randomly placed camera and a trail-based camera, with the aim of comparing communities resulting from the two camera trap placement strategies.

## Usage

```
data(cameratrapp2)
```

## Format

Contiene 6 variables, como sigue:

- referencia** Number of observation of datasets.
- posicion** Type of "placement" placed in each station (random or trail).
- temporada** Season where were made the samplings.
- estacion** Station where were collected the data.
- especie** Name of specie medium to large terrestrial mammals.
- fecha.hora** The date and time of each photographic event is also given.

## Source

Los datos fueron cedidos por el Dr Jeremy Cusack.

## References

- Cusack J, Dickman A, Rowcliffe M, Carbone C, Macdonald D, Coulson T.
1. Random versus game trail-based camera trap placement strategy for monitoring terrestrial mammal communities. PLoS ONE 10(5): e0126373.

**Examples**

```
data(cameratrap2)
head(cameratrap2)
```

---

carAccidents	<i>Driver status after car accidents in Greece.</i>
--------------	---

---

**Description**

A data frame showing the use of seat belt and the driver status after a car accident in Greece.

**Usage**

```
data(carAccidents)
```

**Format**

Contains the factor variables:

**record** factor representing the driver status.

**seatBelt** factor indicating whether the driver wore a setbelt.

**Source**

R package 'gginference'

**Examples**

```
data(carAccidents)
head(carAccidents)
table(carAccidents)
```

---

caribou	<i>Caribou survival</i>
---------	-------------------------

---

**Description**

Caribou survival

**Usage**

```
caribou
```

**Format**

Data frame con 91 filas y 3 columnas:

**herd** Herd identifier.

**wolf.density** Wolf density of the herd as wolf / 100 km<sup>2</sup>.

**alive** Caribou survival, 1 survives, 0 don't survive.

**Examples**

```
data(caribou)
table(caribou$alive, caribou$herd)
```

---

caribou2	<i>Sobrevivencia de caribú</i>
----------	--------------------------------

---

**Description**

Sobrevivencia de caribú

**Usage**

```
caribou2
```

**Format**

Data frame con 91 filas y 3 columnas:

**herd** Identificador de la manada.

**wolf.density** Densidad de lobos, en número de lobos / 100 km<sup>2</sup>.

**alive** Sobrevivencia de un caribú, 1 sobrevive, 0 no sobrevive.

**Examples**

```
data(caribou2)
table(caribou2$alive, caribou2$herd)
```

---

casen*Datos encuesta CASEN del 2022*

---

**Description**

Encuesta de Caracterización Socioeconómica Nacional (CASEN) de Chile, es realizada por el Ministerio de Desarrollo Social y Familia con el objetivo de disponer de información que permita conocer situación de los hogares y de la población. Estos datos corresponden a los de la encuesta CASEN 2022.

**Usage**

```
data(casen)
```

**Format**

Este set de datos contiene las siguientes columnas:

**id.vivienda** Identificador de la vivienda.

**id.persona** Identificador de la persona.

**region** Región administrativa de Chile.

**comuna** Comuna.

**edad** Edad de la persona, en años.

**sexo** Sexo de la persona.

**esc** Años de escolaridad (edad  $\geq$  15).

**educ** Clasificación de educación recibida.

**personas.hogar** Número de personas que habitan en el hogar.

**tipohogar** Nivel de tipo de hogar según encuesta.

**activ** Nivel de actividad actual de la persona según encuesta.

**ytot** Ingreso total.

**ytoth** Ingreso total del hogar.

**ypch** Ingreso total per cápita del hogar.

**ytotcor** Ingreso total corregido.

**ytotcorh** Ingreso total corregido del hogar.

**ypc** Ingreso total corregido per cápita del hogar.

**mayor.nivel.edu** ¿Cuál es el nivel educacional al que asiste o el más alto al cual asistió?

**area.edu.cinef** Clasificación Internacional Normalizada de Educación (CINE-F).

**subarea.edu.cinef** Clasificación Internacional Normalizada de Sub-Area de Educación (CINE-F).

**previ.salud** Sistema de previsión de salud.

## Source

Los datos fueron obtenidos desde el web <https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen>. Note que solo algunas columnas son utilizadas aca, así como el nombre de algunas columnas fueron levemente cambiados.

## Examples

```
data(casen)
head(casen)
table(casen$region)
table(casen$region, casen$sexo)
tapply(casen$ytotcor, casen$sexo, sum)
```

cdf

*Function to compute the cumulative distribution of a variable*

## Description

Builds the cumulative distribution of a vector, using a `step%` of the data as fixed-intervals.

## Usage

```
cdf(y = y, step = 0.05)
```

## Arguments

<code>y</code>	a vector of a random variable
<code>step</code>	a numeric proportion of the data used as increment interval for building the cdf of the random variable. The default value for 'step' is 0.05, representing a 5%.

## Details

By default the cumulative distribution is build using 5% of the data as intervals, that is to say, from 0.05 (i.e., 5%) to 0.95 (i.e., 95%).

## Value

returns a dataframe having two columns: the first contains the random variable values and the second the cumulative distribution for the variable.

## Author(s)

Christian Salas-Eljatib

## References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

### Examples

```
y.var <- rnorm(10)
cdf(y.var)
cdf(y.var, step=0.1)
```

---

chicksw

*Chicken growth data.*

---

### Description

The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

### Usage

```
data(chicksw)
```

### Format

Contains four variables, as follows:

**chick** An ordered factor with levels different giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

**diet** A factor with levels 1,2,3 and 4 indicating which experimental diet the chick received.

**time** A numeric vector giving the number of days since birth when the measurement was made.

**weight** A numeric vector giving the body weight of the chick (gm).

### Source

The data were obtained from the *alr4* library.

### References

Crowder M, Hand D. 1990. Analysis of Repeated Measures. Chapman and Hall

### Examples

```
data(chicksw)
head(chicksw)
```

---

chicksw2

*Crecimiento de pollos.*


---

### Description

El peso de pollos fueron medidos al momento de nacer y cada día por medio hasta el día 20. Ellos también fueron medidos el día 21. Hubo cuatro grupos de pollos en diferentes dietas de proteínas.

### Usage

```
data(chicksw2)
```

### Format

Contiene cuatro variables, como sigue:

**pollo** Un identificador único para cada pollo. La numeración está ordenada según el peso final dentro de cada dieta.

**dieta** Un factor con cuatro niveles: 1,2,3 y 4 indicando qué dieta recibió el pollo.

**tiempo** Número de días desde el nacimiento.

**peso** Peso del pollo (gm).

### Source

Los datos fueron obtenidos desde la librería *alr4*.

### References

Crowder M, Hand D. 1990. Analysis of Repeated Measures. Chapman and Hall

### Examples

```
data(chicksw2)
head(chicksw2)
```

---

co2temp

*CO2 emissions and temperature at country-level.*


---

### Description

Data obtained from the *hockeystick* package, which retrieves annual global carbon dioxide emissions since 1750 from the World Data repository <https://github.com/owid/co2-data>, as well as other climate-related variables.



**Usage**

```
data(co2temp)
```

**Format**

The data contains 75 variables, and the fully description can be reviewed in the references provided [here](#).

**country** Country.

**year** Calendar year.

**iso\_code** TBA.

**population** Population size, in number of people.

**gdp** Gross domestic product, a measure of the value added created through the production of goods and services in a country.

**cement\_co2** TBA.

**cement\_co2\_per\_capita** TBA.

**co2** TBA.

**co2\_growth\_abs** TBA.

**co2\_growth\_prc** TBA.

**co2\_including\_luc** TBA.

**co2\_including\_luc\_growth\_abs** TBA.

**co2\_including\_luc\_growth\_prc** TBA.

**co2\_including\_luc\_per\_capita** TBA.

**co2\_including\_luc\_per\_gdp** TBA.

**co2\_including\_luc\_per\_unit\_energy** TBA.

**co2\_per\_capita** TBA.

**co2\_per\_gdp** TBA.

**co2\_per\_unit\_energy** TBA.

**coal\_co2** TBA.

**coal\_co2\_per\_capita** TBA.

**consumption\_co2** TBA.

**consumption\_co2\_per\_capita** TBA.

**consumption\_co2\_per\_gdp** TBA.

**cumulative\_cement\_co2** TBA.

**cumulative\_co2** TBA.

**cumulative\_co2\_including\_luc** TBA.

**cumulative\_coal\_co2** TBA.

**cumulative\_flaring\_co2** TBA.

**cumulative\_gas\_co2** TBA.

**cumulative\_luc\_co2** TBA.  
**cumulative\_oil\_co2** TBA.  
**cumulative\_other\_co2** TBA.  
**energy\_per\_capita** TBA.  
**energy\_per\_gdp** TBA.  
**flaring\_co2** TBA.  
**flaring\_co2\_per\_capita** TBA.  
**gas\_co2** TBA.  
**gas\_co2\_per\_capita** TBA.  
**ghg\_excluding\_lucf\_per\_capita** TBA.  
**ghg\_per\_capita** TBA.  
**land\_use\_change\_co2** TBA.  
**land\_use\_change\_co2\_per\_capita** TBA.  
**methane** TBA.  
**methane\_per\_capita** TBA.  
**nitrous\_oxide** TBA.  
**nitrous\_oxide\_per\_capita** TBA.  
**oil\_co2** TBA.  
**oil\_co2\_per\_capita** TBA.  
**primary\_energy\_consumption** TBA.  
**share\_global\_cement\_co2** TBA.  
**share\_global\_co2** TBA.  
**share\_global\_co2\_including\_luc** TBA.  
**share\_global\_coal\_co2** TBA.  
**share\_global\_cumulative\_cement\_co2** TBA.  
**share\_global\_cumulative\_co2** TBA.  
**share\_global\_cumulative\_co2\_including\_luc** TBA.  
**share\_global\_cumulative\_coal\_co2** TBA.  
**share\_global\_cumulative\_flaring\_co2** TBA.  
**share\_global\_cumulative\_gas\_co2** TBA.  
**share\_global\_cumulative\_luc\_co2** TBA.  
**share\_global\_cumulative\_oil\_co2** TBA.  
**share\_global\_cumulative\_other\_co2** TBA.  
**share\_global\_flaring\_co2** TBA.  
**share\_global\_gas\_co2** TBA.  
**share\_global\_luc\_co2** TBA.  
**share\_global\_oil\_co2** TBA.

```

share_global_other_co2 TBA.
share_of_temperature_change_from_ghg TBA.
temperature_change_from_ch4 TBA.
temperature_change_from_co2 TBA.
temperature_change_from_ghg TBA.
temperature_change_from_n2o TBA.
total_ghg TBA.
total_ghg_excluding_lucf TBA.
trade_co2 TBA.
trade_co2_share TBA.

```

## Source

The data were obtained from the *hockeystick* library of R. Notice that in the dataframe only a portion of countries have been kept.

## References

- <https://www.globalcarbonproject.org/carbonbudget/>
- Friedlingstein P. et al. 2020. Global Carbon Budget 2020, Earth System Science Data 12:3269-3340 [doi:10.5194/essd1232692020](https://doi.org/10.5194/essd1232692020)

## Examples

```

data(co2temp)
names(co2temp)
table(co2temp$country)
lattice::xyplot(co2~year|country,data=co2temp,type="l",as.table=TRUE)

```

---

contrast

*Function to compute the needed statistics for a given contrast*

---

## Description

The function computes the statistics for inference in a given contrast, subject to a given significance level. Those statistics are as follows: estimated contrast, standard error of the contrast, and the confidence interval of the contrast.

## Usage

```

contrast(
  model = model,
  coef.cont = coef.cont,
  grp.m = grp.m,
  grp.n = grp.n,
  alpha = 0.05,
  full = TRUE
)

```

**Arguments**

<code>model</code>	object containing the fitted model
<code>coef.cont</code>	vector with the coefficients to establish the contrasts
<code>grp.m</code>	a vector having the sample mean per each group, or level of the factor under study.
<code>grp.n</code>	a vector having the sample size per each group, or level of the factor under study.
<code>alpha</code>	is the significance level for building the confidence intervals. Default value is 0.05, which is 95% confidence level.
<code>full</code>	FALSE if want short output, TRUE for longer (i.e. more details). Default is TRUE.

**Details**

The contrast is established based upon an already fitted statistical model that describe the relationship among variables. The significance level ('alpha') is defined by the user, although by default has been set to 0.05, that is to say, a 95% of statistical confidence.

**Value**

This function returns the above described statistics for a given contrast.

**Author(s)**

Christian Salas-Eljatib

**References**

- Salas-Eljatib C. 2025. datana: Datasets and Functions to Accompany Análisis de Datos con R. R package version 1.0.7, [doi:10.32614/CRAN.package.datana](https://doi.org/10.32614/CRAN.package.datana), <https://CRAN.R-project.org/package=datana>

**Examples**

```
data(fertiliza)
table(fertiliza$treat)
means.trt <- tapply(fertiliza$volume, fertiliza$treat, mean); means.trt
sds.trt <- tapply(fertiliza$volume, fertiliza$treat, sd); sds.trt
ns.trt <- tapply(fertiliza$volume, fertiliza$treat, length); ns.trt
m1 <- lm(volume ~ treat, data=fertiliza)
anova(m1)
## Coefficients to be used in the contrast
#c1: (tmoA1-A2) - (tmoA3-A4)
C1.coeff <- c(0, 1, 1, -1, -1)
contrast(model=m1, C1.coeff, grp.m=means.trt, grp.n=ns.trt, alpha=0.1, full=TRUE)
contrast(model=m1, C1.coeff, grp.m=means.trt, grp.n=ns.trt, alpha=0.1, full=FALSE)
contrast(m1, C1.coeff, grp.m=means.trt, grp.n=ns.trt, alpha=0.05, full=TRUE)
contrast(m1, C1.coeff, grp.m=means.trt, grp.n=ns.trt)
```

---

corkoak*Tree-level cork biomass data for Oak trees in Portugal*

---

**Description**

Measurements of cork weight in *Quercus suber* (Oak) trees in Portugal.

**Usage**

corkoak

**Format**

**tree** A correlative number for each sample tree.

**csc** is tree circumference at 1.3 m outside bark, in cm.

**cbe** is tree circumference at 1.3 m under bark, in cm.

**bt** bark thickness, in cm.

**hdeb** is debarking height, in m.

**hblc** height to base of live crown, in m.

**nb** number of branches debarked

**cr.diam** crown diameter, in m.

**w** total green weight of the stripped cork, in kg

**stratum** Stratum

**Source**

Data supplied electronically to Prof. Timothy Gregoire (Yale University) by authors accompanied by a note which said "After the article was published we discovered a problem with 2 of the observations so Teresa and I decided it was best just to delete them."

**References**

- Fonseca TJ, Parresol BR. 2001. A new model for cork weight estimation in northern Portugal with methodology for construction of confidence intervals. *Forest Ecology and Management* 152(1):131–139.

**Examples**

```
data(corkoak)
head(corkoak)
```

---

corkoak2*Datos de biomasa de corcho en árboles de Encino en Portugal*

---

**Description**

Mediciones de peso de corcho en árboles muestra de *Quercus suber* en Portugal.

**Usage**

corkoak2

**Format**

**arbol** A correlative number for each sample tree.

**perimetro.cc** is tree circumference at 1.3 m outside bark, in cm.

**perimetro.sc** is tree circumference at 1.3 m under bark, in cm.

**e.corteza** bark thickness, in cm.

**h.desc** is debarking height, in m.

**hcc** height to base of live crown, in m.

**num.ram** number of branches debarked

**diam.copa** crown diameter, in m.

**biomasa** total green weight of the stripped cork, in kg

**estrato** Estrato

**Source**

Datos cedidos por Prof. Timothy Gregoire (Yale University) y los autores originales mencionaron "After the article was published we discovered a problem with 2 of the observations so Teresa and I decided it was best just to delete them."

**References**

- Fonseca TJ, Parresol BR. 2001. A new model for cork weight estimation in northern Portugal with methodology for construction of confidence intervals. *Forest Ecology and Management* 152(1):131–139.

**Examples**

```
data(corkoak2)
head(corkoak2)
```

---

deleteLeft	<i>Deletes the first n-characters of a string</i>
------------	---

---

**Description**

Function to delete the last n-characters of a string from the left-hand side.

**Usage**

```
deleteLeft(fac, n)
```

**Arguments**

fac	is an object of class string or factor
n	is the number of characters to be deleted of a the string given in 'fac'.

**Details**

It is specially set to arrange data vector having alphanumeric format.

**Value**

This function returns an object having n-less characters from the left-hand side.

**Author(s)**

Christian Salas-Eljatib

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
plot.id <- c("BNE1", "BNE2", "PLE1")
deleteLeft(plot.id, 1)
deleteLeft(plot.id, 2)
deleteLeft(plot.id, 3)
```

---

deleteRight	<i>Deletes the last n-characters of a string</i>
-------------	--

---

### Description

Function to delete the last n-characters of a string from the right-hand side.

### Usage

```
deleteRight(fac, n)
```

### Arguments

fac	is an object of class string or factor
n	is the number of characters to be deleted of a the string given in 'fac'.

### Details

It is specially set to arrange data vector having alphanumeric format.

### Value

This function returns an object having n-less characters from the right-hand side.

### Author(s)

Christian Salas-Eljatib

### References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

### Examples

```
last.names.id <- c("Stage-1924", "Gregoire-1958", "Robinson-1967")
deleteRight(last.names.id, 5)
deleteRight(last.names.id, 4)
```



---

descstat	<i>Creates a descriptive statistics table for continuous variables.</i>
----------	---

---

## Description

Function to create a descriptive statistics table for continuous variables from a dataframe.

## Usage

```
descstat(
  data = data,
  decnum = 3,
  eng = TRUE,
  full = FALSE,
  reduced = FALSE,
  all.outputs = FALSE,
  landscape = FALSE,
  short.names = FALSE,
  segregated = FALSE,
  factor = NA,
  ...
)
```

## Arguments

<code>data</code>	a dataframe containing numeric variables as columns.
<code>decnum</code>	the number of decimals to be used in the output. The default is set to 3.
<code>eng</code>	logical; if TRUE (by default), the language of the statistics will be in English; if "FALSE" will be in Spanish. descriptive statistics. The default is to FALSE.
<code>full</code>	logical; if TRUE, the output includes some extra descriptive statistics. The default is to FALSE.
<code>reduced</code>	logical; if TRUE, the output includes the same descriptive statistics as using the <code>summary()</code> basis R function.
<code>all.outputs</code>	logical; if TRUE, the returns several elements as results of the function, which can be of importance for further analyses later on. The default is to FALSE.
<code>landscape</code>	logical; the default is set to FALSE, thus the output table will have the statistics as rows, and in each column the variables. Otherwise, if TRUE the variables will be the rows, and each statistics the columns. Therefore this last option is only advisable when <code>full=FALSE</code> .
<code>short.names</code>	logical; if TRUE, the names of the computed statistics are in lower cases and are short compared to the more formal ones. For instance, "sd" is used instead of "Std. Dev.", furthermore, no space is used among letters. The default is to FALSE.
<code>segregated</code>	logical, if TRUE performs a call to <code>datana::descstat</code> by a factor given in the parameter <code>factor</code> .

factor	a vector of type factor. Can be an external vector (not within the data given in the data parameter).
...	additional options for basic stats functions.

### Details

The resulting table offers the main central and dispersion statistics.

### Value

This function wraps descriptive statistics into a summarize table having the following statistics: sample size, minimum, maximum, mean, median, SD, and coefficient of variation. If the `full` option is set to `TRUE`, the following statistics will be added to the table: 25th and 75th percentiles, the interquartile range, skewness, and kurtosis.

### Author(s)

Christian Salas-Eljatib and Tomas Cayul.

### References

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. <https://eljatib.com>

### Examples

```
df <- datana::idahohd
##create a factor dbh.class
df <- datana::assigncl(data = df, variable = "dbh")
## subset of numeric data
df.h<-df[,c("dbh","toth")]
## using the function
descstat(data=df.h)
## trying with a factor, note that the factor is a vector of the
## dataframe `df` and is of type `factor`
descstat(df.h, segregated = TRUE, factor = df$spp, eng = FALSE)
## trying with two factors
descstat(df.h, segregated = TRUE, factor = list(df$spp, df$dbh.class), eng = FALSE)

## exploring other parameters
descstat(data=df.h,decnum=1,eng=FALSE)
descstat(df.h,2)
descstat(df.h,2,full=TRUE)
descstat(df.h,2,reduced=TRUE)
descstat(df.h,2,reduced=TRUE,eng=FALSE)
descstat(data=df.h[, "dbh"],decnum=1,eng=FALSE,landscape = FALSE)
descstat(data = df.h, eng = FALSE, full = TRUE)
descstat(data = df.h, eng = TRUE, reduced = TRUE)
```

---

education*Stock and Watson CPS Education Earnings*

---

## Description

Stock and Watson (2007) provide several subsets created from March Current Population Surveys (CPS) with data on the relationship of earnings and education over several years. This data corresponds to the CPSSWEducation dataset.

## Usage

```
data(education)
```

## Format

A data frame containing 2,950 observations on 4 variables.

**age** Age in years.

**gender** Factor indicating gender.

**earnings** Average hourly earnings (sum of annual pretax wages, salaries, tips, and bonuses, divided by the number of hours worked annually).

**education** Number of years of education.

## Source

Data corresponds to dataset CPSSWEducation from the package AER. Online complements to Stock and Watson (2007).

## References

- Stock, J.H. and Watson, M.W. (2007). *Introduction to Econometrics*, 2nd ed. Boston: Addison Wesley.

## Examples

```
data(education)

## Stock and Watson, p. 165
plot(earnings ~ education, data = education)
fm <- lm(earnings ~ education, data = education)
abline(fm)
```

---

election

*Presidential election data of Florida (USA) in 2000.*

---

### Description

County-by-county vote for president in Florida in 2000 for Bush, Gore and Buchanan.

### Usage

```
data(election)
```

### Format

Contains three variables, as follows:

**gore** Vote for Gore.

**bush** Vote for Bush.

**buchanan** Vote for Pat Buchanan.

### Source

The data were obtained from the *alr4* library.

### References

Weisberg S. 2014. Applied Linear Regression. 4th edition. Hoboken NJ: Wiley

### Examples

```
data(election)
head(election)
```

---

election2

*Elección presidencial en el estado de Florida (USA) en el 2000.*

---

### Description

Conteo de votos a nivel de condado en el estado de Florida, año 2000.

### Usage

```
data(election2)
```

## Format

Contiene las siguientes tres columnas:

**gore** Votos para Gore. Número de votos para Al Gore.

**bush** Votos para Bush. Número de votos para George W. Bush.

**buchanan** Votos para Buchaman. Número de votos para Pat Buchanan.

## Source

Los datos se obtuvieron desde el paquete *alr4* de R.

## References

Weisberg S. 2014. Applied Linear Regression. 4th edition. Hoboken NJ: Wiley

## Examples

```
data(election2)
head(election2)
```

---

endfid2

---

*Puntaje ENDFID 2021 por carrera*


---

## Description

Puntaje promedio por carrera de la Evaluación Nacional Diagnóstica de la Formación Inicial Docente (ENDFID), enfocado en matemática. Se tienen 79 observaciones. Se incluyen dos variables binarias: *cuech* (pertenece 1 o no 0 al CUECH) y *pace* (tiene cupos PACE 1 o no 0).

## Usage

```
data(endfid2)
```

## Format

Variables se describen a continuación:

**programa** Nombre de la carrera dictada

**universidad** Universidad correspondiente al programa

**zona** Ubicación de la sede de la carrera

**region** Región de la sede de la carrera

**tipo.programa** Tipo de carrera (1 Ped. En Matemáticas, 2 Enseñanza General Básica, 3 Programa formación pedagógica)

**cuech** Universidad pertenece al Consejo de Universidades del Estado (1 si, 0 no)

**pace** Carrera incluye cupos PACE (1 si, 0 no)

**end.pcp**g Puntaje promedio de la carrera en la Prueba de Conocimientos Pedagógicos Generales

**end.pcdd** Puntaje promedio de la carrera en la Prueba de Conocimientos Disciplinarios y Didácticos

**matricula** Cantidad de estudiantes matriculados en la carrera el 2022

### Source

Datos obtenidos desde el Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP) del Mineduc y desde los sitios web respectivo de cada universidad. Los datos fueron digitados por Diego Fernández, estudiante del Prof. Christian Salas-Eljatib.

### Examples

```
data(endfid2)
head(endfid2)
```

---

eucaleaf

*Leaf measurements for Eucalyptus nitens trees in Tasmania, Australia.*

---

### Description

The length, width, and area of Eucalyptus nitens leaves were measured.

### Usage

```
data(eucaleaf)
```

### Format

Contains leaf-level variables, as follows:

**time** Time factor, in two levels: early or Late.

**tree** Sample tree code identifier.

**shoot** Shoot description factor, in three levels.

**l** Length of the leaf, in mm.

**w** Width of the leaf, in mm.

**la** leaf area, in cm<sup>2</sup>.

### Source

Although the original source of the measurements is the Dissertation of Dr Candy (1999), the data file used here was courtesy of Prof. Timothy Gregoire at Yale University (New Haven, CT, USA). Furthermore, these data were used by Gregoire and Salas (2009).

## References

- Candy SG. 1999. Predictive models for integrated pest management of the leaf beetle *Chrysophtharta bimaculata* in *Eucalyptus nitens* in Tasmania. Doctoral dissertation, University of Tasmania, Hobart, Australia.
- Gregoire TG, and Salas C. 2009. Ratio estimation with measurement error in the auxiliary variate. Biometrics 65(2):590-598 doi:10.1111/j.15410420.2008.01110.x

## Examples

```
data(eucaleaf)
head(eucaleaf)
```

---

eucaleaf2	<i>Mediciones foliares para árboles de Eucalyptus nitens en Tasmania, Australia.</i>
-----------	--

---

## Description

Mediciones de largo, ancho y area de hojas de Eucalyptus nitens.

## Usage

```
data(eucaleaf2)
```

## Format

Contiene variables a nivel de hoja, como sigue:

**tiempo** Factor a dos niveles: Temprano o Tardío.

**arbol** Identificador del árbol muestra.

**meristema** Factor de la descripción del meristema, en tres niveles.

**largo** Largo de la hoja, en mm.

**ancho** Ancho de la hoja, en mm.

**area** Área foliar, en cm<sup>2</sup>.

## Source

Aunque la fuente original de estas mediciones proviene de la tesis del Dr. Candy (1999), el archivo de datos fue cortesía del Prof. Timothy Gregoire de Yale University (New Haven, CT, USA). Además, estos datos fueron ocupados en el estudio de Gregoire y Salas (2009).

## References

- Candy SG. 1999. Predictive models for integrated pest management of the leaf beetle *Chrysophtharta bimaculata* in *Eucalyptus nitens* in Tasmania. Doctoral dissertation, University of Tasmania, Hobart, Australia.
- Gregoire TG, and Salas C. 2009. Ratio estimation with measurement error in the auxiliary variate. *Biometrics* 65(2):590-598 doi:[10.1111/j.15410420.2008.01110.x](https://doi.org/10.1111/j.15410420.2008.01110.x)

## Examples

```
data(eucaleaf2)
head(eucaleaf2)
hist(eucaleaf2$area)
```

---

eucaleafAll	<i>Leaf measurements (all, n=744) for Eucalyptus nitens trees in Tasmania, Australia.</i>
-------------	---

---

## Description

The length, width, and area of *Eucalyptus nitens* leaves were measured for all the samples of Candy (1999).

## Usage

```
data(eucaleafAll)
```

## Format

Contains leaf-level variables, as follows:

**time** Time factor, in two levels: early or Late.

**tree** Sample tree code identifier.

**shoot** Shoot description factor, in three levels.

**l** Length of the leaf, in mm.

**w** Width of the leaf, in mm.

**la** leaf area, in cm<sup>2</sup>.

## Source

Although the original source of the measurements is the Dissertation of Dr Candy (1999), the data file used here was courtesy of Prof. Timothy Gregoire at Yale University (New Haven, CT, USA). Furthermore, these data were used by Gregoire and Salas (2009).



## References

- Candy SG. 1999. Predictive models for integrated pest management of the leaf beetle *Chrysophtharta bimaculata* in *Eucalyptus nitens* in Tasmania. Doctoral dissertation, University of Tasmania, Hobart, Australia.

## Examples

```
data(eucaleafA11)
head(eucaleafA11)
```

---

eucaleafA112	<i>Mediciones foliares (todas, n=744) para árboles de Eucalyptus nitens en Tasmania, Australia.</i>
--------------	---

---

## Description

Mediciones de largo, ancho y área de hojas de *Eucalyptus nitens* para toda la muestra de Candy (1999).

## Usage

```
data(eucaleafA112)
```

## Format

Contiene variables a nivel de hoja, como sigue:

**tiempo** Factor a dos niveles: Temprano o Tardío

**arbol** Identificador del árbol muestra

**meristema** Factor de la descripción del meristema, en tres niveles.

**largo** Largo de la hoja, en mm

**ancho** Ancho de la hoja, en mm

**area** Área foliar, en cm<sup>2</sup>

## Source

Aunque la fuente original de estas mediciones proviene de la tesis del Dr. Candy (1999), el archivo de datos fue cortesía del Prof. Timothy Gregoire de Yale University (New Haven, CT, USA).

## References

- Candy SG. 1999. Predictive models for integrated pest management of the leaf beetle *Chrysophtharta bimaculata* in *Eucalyptus nitens* in Tasmania. Doctoral dissertation, University of Tasmania, Hobart, Australia.

## Examples

```
data(eucaleafA112)
head(eucaleafA112)
```

---

extractLeft	<i>Extracts the last n-characters of a string</i>
-------------	---

---

### Description

Function to extract the first n-characters of a string from the left-hand side.

### Usage

```
extractLeft(fac, n)
```

### Arguments

fac	is an object of class string or factor
n	is the number of characters to be deleted of a the string given in 'fac'.

### Details

It is specially set to arrange data vector having alphanumeric format.

### Value

This function returns an object having the first n-characters from the left-hand side.

### Author(s)

Christian Salas-Eljatib

### References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

### Examples

```
plot.id <- c("BNE1", "BNE2", "PLE1")
extractLeft(plot.id, 1)
extractLeft(plot.id, 2)
extractLeft(plot.id, 3)
```

---

extractRight	<i>Extracts the last n-characters of a string</i>
--------------	---

---

**Description**

Function to extract the last n-characters of a string from the right-hand side.

**Usage**

```
extractRight(fac, n)
```

**Arguments**

fac	is an object of class string or factor
n	is the number of characters to be deleted of a the string given in 'fac'.

**Details**

It is specially set to arrange data vector having alphanumeric format.

**Value**

This function returns an object having the last n characters from the right-hand side.

**Author(s)**

Christian Salas-Eljatib

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
last.names.id <- c("Stage-1924", "Gregoire-1958", "Robinson-1967")
extractRight(last.names.id, 4)
extractRight(last.names.id, 2)
```

---

fdamage	<i>Foliar damage by Ozone</i>
---------	-------------------------------

---

**Description**

Foliar damage by Ozone

**Usage**

fdamage

**Format**

Data frame con 52 filas y 2 columnas:

**damage** Foliar decoloration, 1 with decoloration, 0 without decoloration.

**ozone** Maximum charge of Ozone concentration.

**Examples**

```
data(fdamage)
table(fdamage$damage)
```

---

fdamage2	<i>Daño foliar por Ozono</i>
----------	------------------------------

---

**Description**

Daño foliar por Ozono

**Usage**

fdamage2

**Format**

Data frame con 52 filas y 2 columnas:

**damage** Decoloración foliar, 1 con decoloración, 0 sin decoloración.

**ozone** Máxima carga de concentración de Ozono.

**Examples**

```
data(fdamage2)
table(fdamage2$damage)
```

---

fertiliza	<i>Fertilization experiment data.</i>
-----------	---------------------------------------

---

**Description**

Data contains volume data at plot-level for a fertilization experiment.

**Usage**

```
data(fertiliza)
```

**Format**

Contains two variables, as follows:

**treat** Treatment level.

**volume** Plot-level volume, in m<sup>3</sup>.

**Source**

The data were provided by Dr. Christian Salas-Eljatib (Universidad de Chile, Santiago, Chile).

**Examples**

```
data(fertiliza)
head(fertiliza)
class(fertiliza$treat)
unique(fertiliza$treat)
means.g <- tapply(fertiliza$volume,fertiliza$treat,mean);means.g
sds.g <- tapply(fertiliza$volume,fertiliza$treat,sd);sds.g
ns.g <- tapply(fertiliza$volume,fertiliza$treat,length);ns.g
```

---

fertiliza2	<i>Experimento de fertilización</i>
------------	-------------------------------------

---

**Description**

Datos a nivel de parcela de un experimento de fertilización con tratamientos y replicas.

**Usage**

```
data(fertiliza2)
```

**Format**

Contiene tres columnas como sigue:

**tmo** Tratamiento.Factor medido en diferentes niveles.

**vol** Volumen de madera en la parcela experimental, en m<sup>3</sup>.

**Source**

Datos cedidos por el Prof. Christian Salas.

**References**

not yet

**Examples**

```
data(fertiliza2)
head(fertiliza2)
class(fertiliza2$tmo)
unique(fertiliza2$tmo)
media.g <- tapply(fertiliza2$vol,fertiliza2$tmo,mean);media.g
desvst.g <- tapply(fertiliza2$vol,fertiliza2$tmo,sd);desvst.g
n.g <- tapply(fertiliza2$vol,fertiliza2$tmo,length);n.g
```

---

ficdiamgr

*Diameter growth of trees*

---

**Description**

The 'ficdiamgr' is a fictitious dataframe built to show the structure of longitudinal data. The dataframe has records of tree diameter growth of five sample trees, spanning three species.

**Usage**

```
data(ficdiamgr)
```

**Format**

A time series data containing the following columns:

**tree.id** an ordered factor indicating the tree on which the measurement is made. The ordering is according to increasing maximum diameter.

**time** a numeric vector giving the numbers of days since establishment.

**dbh** a numeric vector of diameter at breast height, in cm.

**site** a factor variable, representing site conditions with two levels.

**spp** a factor variable, representing tree species with three levels.

## Source

This dataframe was built from the 'Orange' data of the *datasets* package, by Christian Salas-Eljatib.

## Examples

```
data(ficdiamgr)

coplot(dbh ~ time | tree, data = ficdiamgr, show.given = FALSE)
```

---

ficdiamgr2	<i>Crecimiento diametral de árboles</i>
------------	---

---

## Description

Los datos 'ficdiamgr2' son ficticios, y fue construida para mostrar la estructura de datos longitudinales. Los datos tienen registro de crecimiento en cinco árboles muestra, representando a tres especies.

## Usage

```
data(ficdiamgr2)
```

## Format

Una serie de tiempo conteniendo las siguientes columnas:

**arbol** indica el identificador del árbol.

**tiempo** número de días desde el inicio de las mediciones.

**dap** diámetro a la altura del pecho, en cm.

**sitio** un factor, representando condiciones de sitio, en dos niveles.

**espe** un factor, representando especie del árbol, en tres niveles.

## Source

Estos datos fueron modificados desde la dataframe 'Orange' de la librería 'datasets', por Christian Salas-Eljatib.

## Examples

```
data(ficdiamgr2)

coplot(dap ~ tiempo | arbol, data = ficdiamgr2, show.given = FALSE)
```

---

findColumn.byname	<i>Finds the position of a specific variable.</i>
-------------------	---

---

**Description**

Sometimes in data manipulation we face the task of locating the position of a specific variable within a dataframe. The function finds the position in which a column name is within an object.

**Usage**

```
findColumn.byname(data = data, col.name = col.name)
```

**Arguments**

data	is a dataframe
col.name	is a string specifying the name of the variable

**Details**

Although the function finds the position of a specific variable, can also be used for more than one variable.

**Value**

This function returns the number of a specific column-name.

**Note**

It can be used for a vector of specified column-names as well.

**Author(s)**

Christian Salas-Eljatib

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
df <- data.frame(varX=1:5, varY=letters[1:5], varZ=rep("a",5),  
varK=rep("b",5))  
df  
#using the function  
findColumn.byname(df, c("varY","varZ"))  
findColumn.byname(df, "varK")  
#Creating an example vector  
vector <- letters
```



```
vector  
findColumn.byname(vector, c("h", "z"))
```

---

fishgrowth	<i>Fish growth variables.</i>
------------	-------------------------------

---

### Description

Variables of small mouth bass (i.e, a fish) collected in West Bearskin Lake, Minnesota, in 1991.

### Usage

```
data(fishgrowth)
```

### Format

Contains three variables, as follows:

**years** Year at capture.

**length** Length at capture (mm).

**scale** radius of a key scale (mm).

### Source

The data were obtained from the *alr4* library of R, specifically from the dataframe *wblake* that includes only fish of ages 8 or younger.

### References

Weisberg S. 2014. Applied Linear Regression. 4th edition. Hoboken NJ: Wiley

### Examples

```
data(fishgrowth)  
head(fishgrowth)  
plot(length~age, data=fishgrowth)
```

---

fishgrowth2	<i>Crecimiento de peces</i>
-------------	-----------------------------

---

**Description**

Variables de crecimiento de peces en el lago West Bearskin del estado de Minnesota, en 1991.

**Usage**

```
data(fishgrowth2)
```

**Format**

Contiene tres variables, como sigue:

**edad** Edad de captura.

**largo** Largo al momento de la captura, en mm.

**escama** Radio de una escama clave, en mm.

**Source**

Datos obtenidos desde el paquete *alr4* de R, de la dataframe *wblake* que incluye peces de hasta 8 años.

**References**

Weisberg S. 2014. Applied Linear Regression. 4th edition. Hoboken NJ: Wiley

**Examples**

```
data(fishgrowth2)
head(fishgrowth2)
plot(largo ~ edad, data = fishgrowth2)
```

---

forestfire	<i>Forest fire occurrence in central Chile</i>
------------	--

---

**Description**

Data of forest fire occurrence in central Chile having 7210 observations, with 890 cases of fire occurrence and 6320 cases of non-occurrence. The binary variable ( $Y$ ) is the occurrence of forest fire, where  $Y = 1$  to denotes occurrence and  $Y = 0$ , otherwise.

**Usage**

```
data(forestfire)
```

## Format

The data frame contains four variables as follows:

**fire** Occurrence of forest fire (1 yes, 0 no)  
**xcoord** Geographic coordinate x.utm  
**ycoord** Geographic coordinate y.utm  
**aspect** Exposure (degrees from north)  
**eleva** Elevation (m)  
**slope** Slope (degrees)  
**distr** Distance to dirt roads  
**distcity** Distance to cities  
**distriver** Distance to paved roads  
**covera** Land use classifications according to a polygon  
**coverb** Land use classifications according to a polygon  
**tempe** Minimum temperature of the coldest month  
**ppan** Annual precipitation  
**ndii** Normalized difference infrared index  
**nvgi** Normalized difference vegetation index  
**tempe2** Minimum temperature of the warmest month  
**ppan2** Precipitation of the driest month  
**frec.fire** Frequency of fires  
**perc.fire** Percentage of fire frequency  
**fireClass** Class for frequency fire  
**asp.class** Class of variable exposure  
**eleva.class** Class of numerical variable elevation  
**slope.class** Class of numerical variable slope  
**ndii.class** Normalized difference infrared index class  
**nvgi.class** Normalized difference vegetation index class

## Source

Data were provided by Dr Adison Altamirano at the Universidad de La Frontera (Temuco, Chile).

## References

- Salas-Eljatib C, Fuentes-Ramírez A, Gregoire TG, Altamirano A, Yaitul V. 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators* 85:502-508. doi:10.1016/j.ecolind.2017.10.030
- Altamirano A, Salas C, Yaitul V, Smith-Ramirez C, Avila A. 2013. Influencia de la heterogeneidad del paisaje en la ocurrencia de incendios forestales en Chile Central. *Revista de Geografia del Norte Grande*, 55:157-170.

**Examples**

```
data(forestfire)
head(forestfire)
```

---

forestfire2

*Ocurrencia de incendios forestales*


---

**Description**

Datos de ocurrencia de incendios forestales en la zona central de Chile. Se tienen 7210 observaciones, de las cuales 890 tienen ocurrencia de incendios y 6320 casos de no ocurrencia. La variable binaria ( $Y$ ) es la ocurrencia de un incendio forestal, donde  $Y = 1$  denota ocurrencia y  $Y = 0$ , lo contrario.

**Usage**

```
data(forestfire2)
```

**Format**

Variables se describen a continuacion:

**fire** Presencia de incendio forestal (1 si, 0 no)  
**xcoord** Coordenada geografica x.utm  
**ycoord** Coordenada geografica y.utm  
**aspect** Exposicion (grados desde el norte)  
**eleva** Elevacion (m)  
**slope** Pendiente (grados)  
**distr** Distancia a caminos de tierra  
**distcity** Distancia a ciudades  
**distriver** Distancia a caminos pavimentados  
**covera** Clasificaciones de uso del suelo segun un poligono  
**coverb** Clasificaciones de uso del suelo segun un poligono  
**tempe** Temperatura m?nima del mes m?s frio  
**ppan** Precipitacion anual  
**ndii** Indice infrarrojo de diferencia normalizado  
**nvd** Indice de vegetacion de diferencia normalizado  
**tempe2** Temperatura m?nima del mes mas calido  
**ppan2** Precipitacion del mes mas seco  
**frec.fire** Frecuencia de incendios  
**perc.fire** Porcentajede la frecuencia de incendios

**fireClass** Clase para variable frecuencia de incendio  
**asp.class** Clase de variable exposicion  
**eleva.class** Clase de variable numerica elevacion  
**slope.class** Clase de variable numerica pendiente  
**ndii.class** Clase de indice infrarrojo de diferencia normalizado  
**nvgi.class** Clase de indice de vegetacion de diferencia normalizado

## Source

Datos fueron cedidos por el Dr. Adison Altamirano, Universidad de La Frontera, Temuco, Chile.

## References

- Salas-Eljatib C, Fuentes-Ramírez A, Gregoire TG, Altamirano A, Yaitul V. 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators* 85:502-508. doi:[10.1016/j.ecolind.2017.10.030](https://doi.org/10.1016/j.ecolind.2017.10.030)
- Altamirano A, Salas C, Yaitul V, Smith-Ramirez C, Avila A. 2013. Influencia de la heterogeneidad del paisaje en la ocurrencia de incendios forestales en Chile Central. *Revista de Geografia del Norte Grande*, 55:157-170.

## Examples

```
data(forestfire2)
head(forestfire2)
```

---

gasoline	<i>Prices of gasoline and crude oil</i>
----------	---

---

## Description

Prices of gasoline and crude oil

## Usage

```
gasoline
```

## Format

Data frame of 14 rows and 3 columns:

**year** Year of data  
**gasoline** Price of gasoline for year in cents / gallon  
**crude.oil** Price of crude oil for year in \$ / bbl

**Source**

McClave, James T. Benson, P.G. 1991. Statistics for Business and Economics, Fifth Edition. Dellen and Macmillan.

**References**

Statistical Abstract of the United States: 1989, pp476, 480.

**Examples**

```
data(gasoline)
plot(gasoline~year, data = gasoline, type = "b",
     ylab = "Gasoline price (cents/gallon)",
     xlab = "Year")
```

---

gasoline2

*Precios de gasolina y petróleo*


---

**Description**

Precios de gasolina y petróleo

**Usage**

```
gasoline2
```

**Format**

Data frame que contiene 14 filas y 3 columnas:

**año** Año del precio

**gasolina** Precio de la gasolina para el año en centavos / galón

**petroleo** Precio del petróleo para el año en \$ / bbl

**Source**

McClave, James T. Benson, P.G. 1991. Statistics for Business and Economics, Fifth Edition. Dellen and Macmillan.

**References**

Statistical Abstract of the United States: 1989, pp476, 480.

**Examples**

```
data(gasoline2)
plot(gasolina~año, data = gasoline2, type = "b",
     ylab = "Precio de la gasolina (centavos/galón)",
     xlab = "Año")
```

gdpcap

*Datos GDP-per capita***Description**

Datos del producto interno bruto per capita, por país.

**Usage**

```
data(gdpcap)
```

**Format**

Este set de datos contiene las siguientes columnas:

**país** Nombre del país.

**país.cod** Codificación del país.

**gdp.pc** GDP per capita, en US dollars.

**y** GDP per capita, en miles de US dollars.

**Source**

Los datos fueron obtenidos desde la web <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>

**Examples**

```
data(gdpcap)
head(gdpcap)
unique(gdpcap$país)
hist(gdpcap$y, breaks=20,xlab='PIB per capita (miles de US$)', col='orange', las=1)
```

gmean

*Function to compute the geometric mean of a numeric vector***Description**

Computes the geometric mean of a numeric vector. It is the n-th root of the product of n numbers, as follows.

$$y_g = \left( \prod_{i=1}^n y_i \right)^{1/n}$$

for  $y_i > 0$ . The geometric mean can be used a central position statistics of a random variable.

**Usage**

```
gmean(v)
```

**Arguments**

`v` is a numeric vector

**Details**

Notice that can only be computed for positive values. For negative values, there are alternatives, but not covered here.

**Value**

This function returns the geometric mean, a numeric scalar.

**Author(s)**

Christian Salas-Eljatib.

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
y.var <- runif(10, min=10, max=45)
gmean(y.var)
```

---

hgrdfir	<i>Tree height growth of Douglas-fir sample trees in the Northwest of the United States</i>
---------	---

---

**Description**

Data contains 148 observations on the height growth of dominant trees of *Pseudotsuga mensiezzi* in the Northwest of the United States.

**Usage**

```
data(hgrdfir)
```



## Format

The data frame contains seven variables as follows:

**natfor.id** Code identifier.  
**plot.code** Plot number identification  
**tree.code** Tree number identification.  
**dbh** Diameter at breast height at sampling, in in.  
**toth** Total height at sampling, in ft.  
**age** Age of tree, yr.  
**height** Height at a given age, in ft.

## Source

The data were provided by Dr Christian Salas.

## References

- Monserud RA. 1984. Height growth and site index curves for Inland Douglas-fir based on stem analysis data and forest habitat type. *Forest Science* 30(4):943-965.
- Salas C, Stage AR, and Robinson AP. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. *Forest Science* 54(1):107-122. doi:10.1093/forestscience/54.1.107

## Examples

```
data(hgrdfir)
head(hgrdfir)
unique(hgrdfir$tree.code)
table(hgrdfir$plot.code, hgrdfir$tree.code)
tapply(hgrdfir$dbh, hgrdfir$tree.code, mean)
tapply(hgrdfir$dbh, hgrdfir$tree.code, mean) #dbh of each sample tree
tapply(hgrdfir$toth, hgrdfir$tree.code, mean) #toth of each sample tree
```

---

hgrdfir2

*Crecimiento en altura de una muestra de árboles en los Estados Unidos*

---

## Description

Data contiene 148 observaciones sobre el crecimiento en altura de árboles dominantes de *Pseudotsuga menziesii* en el Nor-Oeste de los Estados Unidos

## Usage

```
data(hgrdfir2)
```

## Format

La data frame contiene siete variables:

**bosque.id** Código identificador del bosque.

**parcela** Código identificador de la parcela.

**arbol** Número de identificación árbol.

**dap** Diámetro a la altura del pecho, en pulgadas.

**atot** Altura total, en pies

**edad** Edad, en os

**altura** Altura para cada edad del árbol, en pies

## Source

La data fue cedida por el Dr Christian Salas-Eljatib.

## References

Monserud RA. 1984. Height growth and site index curves for Inland Douglas-fir based on stem analysis data and forest habitat type. *Forest Science* 30(4):943-965.

Salas C, Stage AR, and Robinson AP. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. *Forest Science* 54(1):107-122. doi:[10.1093/forestscience/54.1.107](https://doi.org/10.1093/forestscience/54.1.107)

## Examples

```
data(hgrdfir2)
head(hgrdfir2)
unique(hgrdfir2$arbol.id)
table(hgrdfir2$parcela,hgrdfir2$arbol.id)
tapply(hgrdfir2$dap, hgrdfir2$arbol.id, mean) #dap de cada arbol muestra
tapply(hgrdfir2$atot, hgrdfir2$arbol.id, mean) #atot de cada arbol muestra
```

---

histbxp

*Function for building a figure having both an histogram and a boxplot for a single random variable*

---

## Description

The function creates a figure having both an histogram and a boxplot for a random variable, as a way to help understanding its distribution.

**Usage**

```

histbxp(
  y = y,
  freq = NULL,
  freqlab = "Frequency",
  varlab = "Variable",
  eng = TRUE,
  refval = NA,
  print.refval = FALSE,
  col.hist = "gray",
  col.bxp = "gray",
  portrait = TRUE,
  oma = c(3, 0.5, 2, 0),
  mar = c(1, 4, 0.2, 1),
  cex.varlab = 1.2,
  refval.symbol = expression(bar(y)),
  col.refval = "blue",
  varlim = NA,
  freqlim = NA
)

```

**Arguments**

<code>y</code>	A numeric vector representing the random variable.
<code>freq</code>	A logical option for plotting the histogram. By default it is set to <code>NULL</code> , thus uses the actual frequencies. Meanwhile, when is <code>TRUE</code> the percentual frequencies are plot, and if <code>FALSE</code> a density is is used. Further details can be found in the function <code>hist()</code> .
<code>freqlab</code>	(optional) A string specifying the frequency label. The default is set to "Frequency".
<code>varlab</code>	(optional) A string specifying the random variable label. The default is set to "Variable".
<code>eng</code>	logical; if "TRUE" (by default), the language of some default text will be in English; if "FALSE" will be in Spanish. The default is to "TRUE".
<code>refval</code>	A numeric value to be used for printing as reference for the random variable. By default is set to the mean of the variable <code>y</code> .
<code>print.refval</code>	A logical statement to define whether a reference value should be printed, if set to <code>TRUE</code> , the mean of the <code>y</code> vector will be plotted. The default is <code>FALSE</code> .
<code>col.hist</code>	A string specifying the histogram color. The default is "gray".
<code>col.bxp</code>	A string specifying the boxplot color. The default is "gray".
<code>portrait</code>	A logical statement, if set to <code>TRUE</code> , the boxplot will be located under the histogram (2 rows, 1 column). If is set to <code>FALSE</code> , the boxplot will be located next to the histogram (1 row, 2 columns). The default is <code>TRUE</code> .
<code>oma</code>	As in the plot environment. The default is <code>c(3, .5, 2, 0)</code> .
<code>mar</code>	As in the plot environment. The default is <code>c(1, 4.0, 0.2, 1)</code> .

<code>cex.varlab</code>	A numeric value for the <code>cex</code> option of plotting to the assigned <code>varlab</code> element. The default value is set to 1.2.
<code>refval.symbol</code>	A string of type expression with name of the <code>refval</code> being printed, if <code>print.refval</code> is set to TRUE. The default is <code>expression(bar(y))</code> .
<code>col.refval</code>	A string specifying the <code>refval.symbol</code> color, if <code>print.refval</code> is set to TRUE. The default is "blue"
<code>varlim</code>	(optional) A numeric vector having the minimum and maximum, respectively for the random variable.
<code>freqlim</code>	(optional) A numeric vector having the minimum and maximum, respectively for the frequency axis.

### Details

The variable must be numeric.

### Value

The function returns the above described graph.

### Author(s)

Christian Salas-Eljatib

### References

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. 170 p. <https://eljatib.com>
- Salas C, Stage AR, and Robinson AP. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. Forest Science 54(1):107-122. [doi:10.1093/forestscience/54.1.107](https://doi.org/10.1093/forestscience/54.1.107)

### Examples

```
df <- datana::fishgrowth
histbxp(y=df$length)
histbxp(y=df$length,freq = TRUE)
histbxp(y=df$length,freq = FALSE)

## Now in Spanish
histbxp(y=df$length,eng=FALSE)
histbxp(y=df$length,freq = TRUE,eng=FALSE)
histbxp(y=df$length,freq = FALSE,eng=FALSE)

### distribution of 'length'
## with mean refval
histbxp(y=df$length, print.refval = TRUE)

## with given refval
histbxp(y=df$length, print.refval = TRUE, refval = 250)
```

```

## changing labels
histbxp(y=df$length, print.refval = TRUE, refval = 250,
        freqlab = "Freq", varlab = "Length")

## changing colors
histbxp(y=df$length, print.refval = TRUE, refval = 250,
        freqlab = "Freq", varlab = "Length",
        col.hist = "blue",
        col.bxp = "green",
        col.refval = "red")

### distribution of 'scale'
## with mean refval
histbxp(y=df$scale, print.refval = TRUE)

## landscape mode
histbxp(y=df$scale, print.refval = TRUE, portrait = FALSE)

## with limits
histbxp(y=df$scale, print.refval = TRUE, portrait = FALSE,
        freqlim = c(0,100),
        varlim = c(0, max(df$scale)))

```

---

hmean

---

*Function to compute the harmonic mean of a numeric vector*


---

## Description

Computes the harmonic mean of a numeric vector. It is the inverse of the mean of the reciprocals of  $n$  numbers, as follows.

$$y_h = \frac{n}{\left(\sum_{i=1}^n \frac{1}{y_i}\right)}$$

for  $y_i \neq 0$ . The harmonic mean can be used a central position statistics of a random variable.

## Usage

```
hmean(v)
```

## Arguments

`v` is a numeric vector

## Details

Notice that can only be computed for values different from zero.

**Value**

This function returns the harmonic mean, a numeric scalar.

**Author(s)**

Christian Salas-Eljatib.

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
y.var <- runif(10, min=10, max=45)
hmean(y.var)
```

---

idahohd

*Tree height-diameter data from Idaho (USA)*

---

**Description**

These data are forest inventory measures from the Upper Flat Creek stand of the University of Idaho Experimental Forest, dated 1991.

**Usage**

```
data(idahohd)
```

**Format**

Contains five variables, as follows:

**plot** Plot number.

**tree** Tree within plot.

**spp** Tree species. A factor variable having the following levels: "DF" is Douglas-fir (*Pseudotsuga menziesii*), "GF" is Grand fir (*Abies grandis*), "SF" is Subalpine fir (*Abies lasiocarpa*), "WL" is Western larch (*Larix occidentalis*), "WC" is Western red cedar (*Thuja plicata*), and "WP" is White pine (*Pinus strobus*).

**dbh** Diameter 137 cm perpendicular to the bole, cm.

**toth** Height of the tree, in m.

**Source**

The data were assembled from the 'ufc' dataframe from the *alr4* library.

## References

Weisberg S. 2014. Applied Linear Regression. 4th edition. New York: Wiley.

## Examples

```
data(idahohd)
head(idahohd)
plot(toth~dbh, data=idahohd)
```

---

idahohd2	<i>Altura-diámetro de árboles en el estado de Idaho (USA)</i>
----------	---

---

## Description

Estos datos provienen de un muestreo en el bosque experimental de la University of Idaho, en Upper Flat Creek, Idaho, USA. Medido en 1991.

## Usage

```
data(idahohd2)
```

## Format

Contiene cinco variables detalladas a continuación:

**parce** Número de la parcela de muestreo.

**arbol** Número del árbol dentro de la parcela.

**spp** Especie del árbol. Una variable factor con los siguientes niveles: "DF" es Douglas-fir (*Pseudotsuga menziesii*), "GF" es Grand fir (*Abies grandis*), "SF" es Subalpine fir (*Abies lasiocarpa*), "WL" es Western larch (*Larix occidentalis*), "WC" es Western red cedar (*Thuja plicata*), y "WP" es White pine (*Pinus strobus*).

**dap** Diámetro del fuste a los 1.3 m sobre el suelo, en cm.

**atot** Altura total del árbol, en m.

## Source

Los datos fueron obtenidos desde la dataframe ufc de la librería alr4.

## References

Weisberg S. 2014. Applied Linear Regression. 4th edition. New York: Wiley.

## Examples

```
data(idahohd2)
head(idahohd2)
plot(atot~dap, data=idahohd2)
```

---

imacec2*Índice Mensual de Actividad Económica (IMACEC)*

---

**Description**

Base de datos con el Índice Mensual de Actividad Económica (IMACEC) de Chile, que incluye información desde enero de 1997 en adelante. La base cuenta con 340 observaciones, que representan meses, e incorpora diversas desagregaciones sectoriales. La variable principal es el IMACEC mensual, que representa una estimación de la evolución de la actividad económica del país respecto al mismo mes del año anterior.

**Usage**

```
data(imacec2)
```

**Format**

Variables se describen a continuación:

**fecha** Fecha de la observación (formato Date, primer día del mes)

**anho** Año de la observación

**mes** Mes de la observación

**imacec** Índice mensual de actividad económica total

**crec.prod** Crecimiento del sector producción de bienes

**crec.min** Crecimiento del sector minería

**crec.ind** Crecimiento del sector industrial

**crec.rest** Crecimiento del resto de bienes no mineros ni industriales

**crec.com** Crecimiento del sector comercio

**crec.serv** Crecimiento del sector servicios

**imacec.fac** IMACEC ajustado por costo de factores

**crec.imp** Crecimiento de los impuestos sobre los productos

**imacec.nomin** Índice de actividad económica excluyendo minería

**Source**

Banco Central de Chile. Datos extraídos de la serie histórica de indicadores mensuales. Los datos fueron digitados por Saúl Ketterer, estudiante del Prof. Christian Salas-Eljatib.

**References**

- Banco Central de Chile. “Serie IMACEC”, disponible en <https://si3.bcentral.cl/siete>

**Examples**

```
data(imacec2)
head(imacec2)
```



---

interp	<i>Interpolation function</i>
--------	-------------------------------

---

**Description**

Interpolation function

**Usage**

```
interp(
  x,
  y,
  xlu = NA,
  ylu = NA,
  arrange = y,
  asc = TRUE,
  completename.x = "xlu",
  completename.y = "ylu",
  overwrite = FALSE,
  method = "spline",
  failsafe = FALSE,
  failsafe.condition = "< 0",
  failsafe.method = "lm"
)

interp.l(data)

interp.m(data)
```

**Arguments**

x	vector of x values, should have same length as y.
y	vector of y values, should have same length as x.
xlu	vector of new x values given to interpolate corresponding y values.
ylu	vector of new y values given to interpolate corresponding x values.
arrange	sort data based on x or y values
asc	wether to sort ascending (TRUE, default) or descending (FALSE).
completename.x	name to use for the completevals xlu generated columns.
completename.y	name to use for the completevals ylu generated columns.
overwrite	wether to overwrite original values (TRUE) or not (FALSE, default) if given interpolation points exists in the original data.
method	what method use for interpolation, any of "spline", "lm" (linear model prediction) or "linear" (standalone rule-of-3).
failsafe	logical, wheter to apply a failsafe method.

`failsafe.condition`  
 string indicating the condition to trigger a failsafe method, should be the right part of the logical expression, *e.g.* for the condition " $x > 0$ " the string "> 0" should be provided.

`failsafe.method`  
 method to be used in failsafe operation, posibles values are the same of the parameter method.

`data`  
 data.frame to interpolate on. It has to have two columns and just one NA value

## Details

This function interpolate via spline missing values in a two dimensional array, where one column ascends while the other descends in value.

Function that interpolates via linear interpolation. This function is used by `datana::interp`.

Function that interpolates via fitting a linear model to the data. This function is used by `datana::interp`.

## Author(s)

Christian Salas-Eljatib and Nicolás Campos

## Examples

```
##- example data

my.x <- seq(40, 0, -4)
my.x

my.y <- seq(0, 20, 2)
my.y

myData <- data.frame(x = my.x, y = my.y)
myData

##- example `xlu`
my.xlu <- c(11, 15, 25)

##- example `ylu`
my.ylu <- c(15, 5, 9) # note that values can be unordered

##- interpolation
interp(x = my.x, y = my.y, ylu = 15)
interp(x = my.x, y = my.y, xlu = my.xlu)
new.y <- interp(x = my.x, y = my.y, xlu = my.xlu) # interp missing ylu
new.y$intvalues # interpolated rows
new.y$datares # interpolated rows appended to original dataframe
new.y$completevals

new.x <- interp(x = my.x, y = my.y, ylu = my.ylu) # interp missing xlu
new.x$intvalues # interpolated rows
new.x$datares # interpolated rows appended to original dataframe
new.x$completevals
```

```
##- both interpolation at the same time
interp(x = my.x, y = my.y, xlu = my.xlu, ylu = my.ylu,
       arrange = my.y, asc = TRUE)

interp(x = my.x, y = my.y, xlu = my.xlu, ylu = my.ylu,
       arrange = my.x, asc = TRUE, completename.x = "dlu")

##- when overwrite = TRUE a warning is noted
interp(x = my.x, y = my.y, ylu = c(14,11), overwrite = TRUE)
interp(x = my.x, y = my.y, xlu = c(28, 15), overwrite = TRUE)
interp(x = my.x, y = my.y, xlu = c(28, 15), ylu = c(14,11), overwrite = TRUE)

##- testing the failsafe condition
to.failsafe <- interp(x = my.x, y = my.y, ylu = 14.54)
## the resulting value is 10.92 via spline
to.failsafe$completevals

## adding a condition to be different from 10.92, now using lm method
failsafed.lm <- interp(x = my.x, y = my.y, ylu = 15, failsafe = TRUE,
                     failsafe.condition = "!= 10.92", failsafe.method = "lm")
failsafed.lm$completevals

## what about using linear method
failsafed.linear <- interp(x = my.x, y = my.y, ylu = 15, failsafe = TRUE,
                        failsafe.condition = "!= 10.92", failsafe.method = "linear")
failsafed.linear$completevals

x <- c(40, 30, 20, 10, 0)
y <- c(0, 5, 10, NA, 20)
df <- data.frame(x, y)
df
interp.l(df)

x <- c(40, 30, 20, 10, 0)
y <- c(0, 5, 10, NA, 20)
df <- data.frame(x, y)
interp.m(df)
```

---

kurto

---

*Computes the sample kurtosis of a distribution*


---

## Description

The kurtosis is about the tailedness, or the degree of heaviness of the tails, in the frequency distribution. The function computes an estimator of the kurtosis.

## Usage

```
kurto(x, na.rm = TRUE)
```

**Arguments**

`x` a numeric vector of a random variable.

`na.rm` logical operator to remove NA values. The default is set to TRUE.

**Details**

The kurtosis of a random variable is the fourth moment of the standardized variable. There are several ways of parameterizing a kurtosis estimator, such as depending on the fourth moment and the standard deviation of the random variable.

**Value**

An estimator of the kurtosis.

**Author(s)**

Christian Salas-Eljatib

**Examples**

```
y.var<-rnorm(100);x.var<-rbeta(100,.2,2)
kurto(y.var)
kurto(x.var)
```

---

landcover

*Land-cover, environmental and sociodemographic data for the 34 municipalities composing the Greater Santiago area, Santiago, Chile.*

---

**Description**

dataset contains 476 observations, 34 categorical and 442 numerical. Land-cover data was generated through remote sensing classification techniques using Sentinel-2 satellite images from year 2016. Temperatures were obtained from TIRS band 10 of Landsat 8 satellites images. Particulate matter concentrations were estimated using spatial modelling techniques from 10 pollution stations distributed in the city. Altitude was generated from a Digital Elevation Model. Population and poverty were gathered from Casen 2017 survey.

**Usage**

```
data(landcover)
```

## Format

The data frame contains four variables as follows:

**county** Name of Municipality  
**built.p** Percentage of surface covered by built-up area  
**vegeta.p** Percentage of surface covered by vegetation  
**naked.p** Percentage of surface covered by bare soil  
**grass.p** Percentage of surface covered by deciduous vegetation  
**p.Deciduo** Percentage of surface covered by evergreen vegetation  
**p.Siempreverde** Percentage of surface covered by evergreen vegetation  
**temp.winter** Land surface temperature in celsius degrees at 2pm on a winter 0% cloud day  
**temp.summer** Land surface temperature in celsius degrees at 2pm on a summer 0% cloud day  
**pm10.winter** Average particulate matter 10 micron during winter months  
**pm10.summer** Average particulate matter 10 micron during summer months  
**poor.p** Percentage of people under poverty line year 2017.  
**eleva** Average altitude of municipal area.  
**pop** Total population of municipality

## Source

Data were provided by Dr Ignacio Fernandez at Universidad Adolfo Ibañez (Santiago, Chile).

## References

Not yet

## Examples

```
data(landcover)
head(landcover)
```

---

landcover2	<i>Cobertura territorial, ambiental y sociodemografica de los 34 municipios que componen el area del Gran Santiago, Santiago, Chile..</i>
------------	---

---

## Description

El conjunto de datos contiene 476 observaciones, 34 categoricas y 442 numericas. Los datos de cobertura terrestre se generaron mediante tecnicas de clasificacion de teledeteccion utilizando imagenes de satelite Sentinel-2 del año 2016. Las temperaturas se obtuvieron de la banda TIRS 10 de las imagenes de los satelites Landsat 8. Las concentraciones de material particulado se estimaron mediante tecnicas de modelado espacial de 10 estaciones de contaminacion distribuidas en la ciudad. La altitud se genero a partir de un modelo de elevacion digital. La poblacion y la pobreza se obtuvieron de la encuesta Casen 2017.

**Usage**

```
data(landcover2)
```

**Format**

Variables se describen a continuacion:

**comuna** Name of Municipality

**const.p** Porcentaje de superficie cubierta por area construida

**vegeta.p** Porcentaje de superficie cubierta por vegetacion

**desnu.p** Porcentaje de superficie cubierta por suelo desnudo

**pasto.p** Porcentaje de superficie cubierta por cesped

**deci.p** Porcentaje de superficie cubierta por vegetacion de hoja caduca

**sverde.p** Porcentaje de superficie cubierta por vegetacion siempre verde

**temp.inv** Temperatura de la superficie terrestre en grados celsius a las 2 p.m.en un dia de invierno con 0% de nubes

**temp.ver** Temperatura de la superficie de la tierra en grados celsius a las 2 p.m.en un dia de verano con 0% de nubes

**pm10.inv** Material particulado promedio de 10 micrones durante los meses de invierno

**pm10.ver** Material particulado promedio de 10 micrones durante los meses de verano

**pobreza.p** Porcentaje de personas por debajo de la linea de pobreza año 2017

**altitud** Altitud media del termino municipal

**pob** Poblacion total del municipio

**Source**

Los datos fueron cedidos por el Dr Ignacio Fernandez de la Universidad Adolfo Ibañez (Santiago, Chile).

**References**

Not yet

**Examples**

```
data(landcover2)
head(landcover2)
```

---

largetrees*Large trees in forests near Tolga, in Eastern Norway.*

---

### Description

The study area is situated in the municipality of Tolga, located in Hedmark County, Eastern Norway. Field plots 32 m × 32 m in size were established in forests. A total of 1109 plots were sampled. In each plot, Scots pines (*Pinus sylvestris* L.). trees with a stem diameter larger than 35 cm were measured and counted.

### Usage

```
data(largetrees)
```

### Format

Contains two variables, as follows:

**plot** Plot code.

**y** Number of large-diameter trees in a given sample plot.

### Source

Although Christian Salas was part of the study, he just reproduced the needed data to mimic the distribution of the random variable of interest, as shown in the study of Korkhonen et al (2016).

### References

- Korkhonen L, Salas C, Ostgard T, Lien V, Gobakken T, Naesset E. 2016. Predicting the occurrence of large-diameter trees using airborne laser scanning. Canadian Journal of Forest Research 46:461–469. doi:[10.1139/cjfr20150384](https://doi.org/10.1139/cjfr20150384)

### Examples

```
data(largetrees)
head(largetrees)
hist(largetrees$y)
```

---

`largetrees2`*Árboles grandes en bosques cercanos a Tolga, en el Este de Noruega.*

---

### Description

El área de estudio esta ubicada en la municipalidad de Tolga, en la comuna de Hedmark, al Este de Noruega. 1109 parcelas de muestreo de 32 m × 32 m se establecieron en los bosques. En cada parcela, los árboles de pino escoses (*Pinus sylvestris* L.). que tuvieran un diámetro mayor a 35 cm fueron medidos y contados.

### Usage

```
data(largetrees2)
```

### Format

Los datos poseen las siguientes dos columnas:

**parc** Identificador de la parcela de muestreo.

y Número de árboles de gran diámetro encontrados en una parcela de muestreo.

### Source

Aunque el Prof. Christian Salas fue parte del estudio, acá se han reproducido los datos necesarios que imitan la distribución de la variable aleatoria de interés, tal como se muestra en el estudio de Korkhonen et al (2016).

### References

- Korkhonen L, Salas C, Ostgard T, Lien V, Gobakken T, Naesset E. 2016. Predicting the occurrence of large-diameter trees using airborne laser scanning. Canadian Journal of Forest Research 46:461–469. doi:[10.1139/cjfr20150384](https://doi.org/10.1139/cjfr20150384)

### Examples

```
data(largetrees2)
head(largetrees2)
hist(largetrees2$y)
```



---

leafw2	<i>Area y peso para hojas de árboles.</i>
--------	---

---

### Description

Mediciones de área y peso de hojas.

### Usage

```
data(leafw2)
```

### Format

Contiene variables a nivel de hoja, como sigue:

**peso** Peso de la hoja, en gramos.

**area** Área foliar, en  $\text{cm}^2$ .

### Source

El archivo de datos fue cortesía del Prof. Timothy Gregoire de Yale University (New Haven, CT, USA).

### References

- Gove JH, Barrett JP, and Gregoire TG. 1982. When is n sufficiently large for regression estimation? Journal of Environmental Management 15:229-237.

### Examples

```
library(datana)
data(leafw2)
head(leafw2)
plot(peso~area, data=leafw2)
```

---

lifexpect	<i>Esperanza de vida de países</i>
-----------	------------------------------------

---

### Description

El repositorio del Observatorio Mundial de la Salud (GHO) de la Organización Mundial de la Salud (WHO) mantiene un registro del estado de salud como también otros factores relacionados, para todos los países. Las bases de datos son publicadas con el objetivo de analizarlos. La base de datos de esperanza de vida ha sido compilada en conjunto con datos económicos de las Naciones Unidas.

## Usage

```
data(lifexpect)
```

## Format

Este set de datos contiene 22 columnas:

**country** País de origen

**year** Año

**status** Categoría del país Desarrollado/En desarrollo

**life.expectancy** Esperanza de vida en años

**adult.mortality** Mortalidad en adultos expresado como la probabilidad de morir entre 15 y 60 años de edad por cada 1000 habitantes

**infant.deaths** Mortalidad en infantes cada 1000 habitantes

**alcohol** Consumo de alcohol percapita en mayores de 15 años

**percentage.expenditure** Porcentaje de vacunación

**hepatitis.b** Porcentaje de vacunación contra hepatitis b

**measles** Casos de sarampión cada 1000 habitantes

**bmi** Índice de masa corporal (BMI) promedio

**under.five.deaths** Muertes de menores de 5 años cada 1000 habitantes

**polio** Porcentaje de vacunación contra polio

**total.expenditure** Inversión en salud como porcentaje del GDP per cápita

**diphtheria** Porcentaje de vacunación contra diphtheria

**hiv.aids** Porcentaje casos de VIH, ETS

**gdp** GDP per cápita en USD

**population** Población total

**thinness10.19** Desnutrición entre 10 y 19 años de edad

**thinness5.9** Desnutrición entre 5 y 9 años de edad

**icr** Índice de desarrollo humano en términos de composición de ingresos

**schooling** Promedio de años de educación

## Source

Los datos fueron obtenidos desde la web <https://rpubs.com/Alvian2022/LifeExpectancy>.  
Note que solo los datos del año 2014 son utilizados acá.

## Examples

```
data(lifexpect)
head(lifexpect)
table(lifexpect$status)
tapply(lifexpect$life.expectancy, lifexpect$status, mean)
```

llancahue

*Tree locations for a sample plot in the Llancahue experimental forest*

## Description

The Cartesian position, species, and diameter of trees within a plot were measured. The sample plot is rectangular of 130 m by 70 m. Further details can be # reviewed in the reference.

## Usage

```
data(llancahue)
```

## Format

Contains tree-level variables, as follows:

**tree.code** Tree identifier

**spp** Tree species abbreviation as follows: "AP" is *Aextoxicon punctatum*, "EC" is *Eucryphia cordifolia*, "GA" is *Gevuina avellana*, "LP" is *Laureliopsis philippiana*, "LS" is *Laurelia semper-virens*, "ND" is *Nothofagus dombeyi*, "PS" is *Podocarpus saligna*, and "Ot" represents other species different from the above described.

**dbh** diameter at breast height, in cm.

**x.coord** Cartesian position in the X-axis, in m.

**y.coord** Cartesian position in the Y-axis, in m.

## Source

The data are provided courtesy of Prof. Daniel Soto at Universidad de Aysen (Coyhaique, Chile).

## References

- Soto DP, Salas C, Donoso PJ, Uteau D. 2010. Heterogeneidad estructural y espacial de un bosque mixto dominado por *Nothofagus dombeyi* después de un disturbio parcial. Revista Chilena de Historia Natural 83(3): 335-347.

## Examples

```
data(llancahue)
head(llancahue)
descstat(llancahue$dbh)
boxplot(dbh~spp, data=llancahue)
```

---

llancahue2Ubicación cartesiana de árboles en el bosque de Llancahue

---

**Description**

Corresponde a la posición cartesiana, especie, y diámetro de árboles en una parcela de muestreo en el bosque de Llancahue, cerca de Valdivia, Chile. La parcela es rectangular con dimensiones de 130 m por 70 m. Mayores antecedentes aparecen en las referencias.

**Usage**

```
data(llancahue2)
```

**Format**

Contains tree-level variables, as follows:

**arb.id** Identificador del árbol.

**spp** Codificación de la especie como sigue: "AP" es *Aextoxicon punctatum*, "EC" es *Eucryphia cordifolia*, "GA" es *Gevuina avellana*, "LP" es *Laureliopsis philippiana*, "LS" es *Laurelia sempervirens*, "ND" es *Nothofagus dombeyi*, "PS" es *Podocarpus saligna*, y "Ot" representa a cualquier especie distinta a cualquiera de las ya definidas.

**dap** Diámetro a la altura del pecho, en cm.

**coord.x** Posición cartesiana en el eje-X, en m.

**coord.y** Posición cartesiana en el eje-Y, en m.

**Source**

Los datos fueron cedidos por el Prof. Daniel Soto de Universidad de Aysen (Coyhaique, Chile).

**References**

- Soto DP, Salas C, Donoso PJ, Uteau D. 2010. Heterogeneidad estructural y espacial de un bosque mixto dominado por *Nothofagus dombeyi* después de un disturbio parcial. Revista Chilena de Historia Natural 83(3): 335-347.

**Examples**

```
data(llancahue2)
head(llancahue2)
descstat(llancahue2$dap)
boxplot(dap~spp, data=llancahue2)
```

---

lrt	<i>Performs a likelihood ratio test between two models being fitted by maximum likelihood.</i>
-----	--

---

## Description

Function to perform a likelihood ratio test (LRT) between a reduced model (modA) versus a more complex model (modB), provided both models were fitted by maximum likelihood. The function requires to be filled with the needed values used to perform a LRT.

## Usage

```
lrt(
  llma = llma,
  llmb = llmb,
  qa = qa,
  qb = qb,
  nfit = nfit,
  modA = "modA",
  modB = "modB",
  alpha = 0.05
)
```

## Arguments

llma	maximized log-likelihood of the reduced model (or modA).
llmb	maximized log-likelihood of the more-complex model (or modB).
qa	the number of parameters of the reduced model.
qb	the number of parameters of the more-complex model.
nfit	the sample size used for fitted both models.
modA	is a character with a name to be assigned to object modA.
modB	is a character with a name to be assigned to object modB.
alpha	is the level of significance to used for computing as a reference only, the tabulated value of the respective Chi-Squared statistic. By the default is set to 0.05.

## Details

The resulting output offers statistical inference estimates of the LRT, as well as other maximum likelihood-based statistics. Notice that the function only works if the number of parameters for modA is lower than the ones of modB.

## Value

This function wraps two outputs: (i) a table that computes the AIC, BIC and AICc goodness-of-fit statistics for both models, and (ii) the result of the likelihood ratio test, such as the value of the statistic being computed, its respective p-value, and the tabulated value of the statistics using the a defined alpha significance of level.

**Author(s)**

Christian Salas-Eljatib.

**References**

Salas-Eljatib, C. 2025. Estadística Aplicada e Inferencial. Borrador de libro, Universidad de Chile, Santiago, Chile. <https://eljatib.com/rlibro>

**Examples**

```
#Maximized values for two probability mass functions
max.ll.pois<- -39.86337; max.ll.bneg<--33.823003
c(max.ll.pois,max.ll.bneg)
sample.size<-26
#Number of parameters
num.para.pois<- 1; num.para.bneg<- 3
c(num.para.pois, num.para.bneg)
#Names to be used for each model
modA="Poisson"; modB="hiper"
outall<-lrt(llma=max.ll.pois,llmb=max.ll.bneg,qa=num.para.pois,
qb=num.para.bneg,nfit = sample.size,modA = "Poisson",
modB = "Hipergeometrico")
#Output1: A comparative table
tab.out<-outall$tab.models
tab.out
#Output2: the results of the LRT
out<-outall$lrt.out
out$r.tab
out$Ldif
```

---

lrt.glm	<i>Computes a likelihood ratio test between a reduced model and a full model. Both models must be already fitted using and R function.</i>
---------	--

---

**Description**

Computes a likelihood ratio test between a reduced model (modr) and a full model (modf). Both models must be previously fitted by maximum likelihood using an R function such as nlme() and such, that are part of the generalized lineal models.

**Usage**

```
lrt.glm(modr, modf)
```

**Arguments**

- modr is the object containing a previously fitted reduced model, using a glm-type of function, having less parameters than modf.
- modf is the object containing a previously fitted full model, using a glm-type of function, having more parameters than modr.

**Details**

Double-check the order of the reduced and full model, before of using the model

**Value**

This function returns an object having the following elements: "loglik.Modr" maximized log-likelihood of modr; "loglik.Modf" maximized log-likelihood of modf; "dif.loglik" difference in log-likelihood between both models, and "dif.df" difference in degrees of freedong of both models, and "p-value" is the p-value for the LRT.

**Author(s)**

Christian Salas-Eljatib.

**References**

Pinheiro JC, and Bates DM. 2000. Mixed-effects models in S and Splus. Springer-Verlag, New York, NY. 528 p.

**Examples**

```
#not yet implemented
```

---

maple

*Tree biomass of sugar maple (Acer saccharum) trees.*

---

**Description**

These are tree-level measurement data of sample trees in the US.

**Usage**

```
data(maple)
```

**Format**

Contains tree-level variables, as follows:

**tree** Sample tree identification number.

**dbh** Diameter at breast height, in cm

**leaf** Leaf biomass, in kg.

**branch** Branches biomass, in kg.

**bole** Stem, or bole, biomass, in kg.

**bark** Bark biomass, in kg.

**total** Total biomass, in kg.

**Source**

The data were provided courtesy of Dr Timothy Gregoire, Yale University, in New Haven, CT, USA.

**References**

- Prof. Christian Salas-Eljatib at Universidad de Chile, Santiago, Chile.

**Examples**

```
data(maple)
head(maple)
plot(total~dbh,data=maple)
```

---

moda

*Computes the mode*

---

**Description**

Computes the mode of a random variable.

**Usage**

```
moda(y = y)
```

**Arguments**

y is a numeric vector.

**Details**

The mode is an statistics representing the most "used" value of a random variable as a measurement of central position. We use the Spanish name of mode, i.e., "moda", to avoid any confution with the mode function of R, wich was programmed for a different use.

**Value**

The function returns the mode, a numeric scalar.

**Author(s)**

Christian Salas-Eljatib.



### Examples

```
library(datana)
data(casen)
head(casen)
df<-casen
#Compare
mean(df$edad)
median(df$edad)
# Using the function
moda(df$edad)
```

---

modresults	<i>Creates an object having the main fitting statistics from a regression model.</i>
------------	--

---

### Description

Function to save the main statistics results from a fitted regression model.

### Usage

```
modresults(mod = mod, print.output = FALSE, conf.lev = 0.95, eng = TRUE)
```

### Arguments

mod	An object containing the fitted model by using the <code>lm()</code> function.
print.output	A logical option for printing, or displaying, the saved outputs at the console. The default is set to <code>TRUE</code> , meanwhile if <code>print.output=FALSE</code> , nothing is printed.
conf.lev	A numeric value (between 0.0001 and 0.9999) representing the confidence level to be used for some components of the output. The default value is 0.95.
eng	The language to be used in the output. English is the default, meanwhile if <code>eng=FALSE</code> , Spanish is used.

### Details

The resulting object contains several outputs derived from a regression model.

### Value

This function returns a list having several components of a fitted regression model.

### Author(s)

A somehow related version of this function was first created by Prof. Timothy Gregoire (Yale University), but the current version is due to Christian Salas-Eljatib.

## References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

## Examples

```
library(datana)
df <- datana::maple
head(df)
datana::descstat(df[,c("total", "dbh")])
graphics::plot(total ~ dbh, data=df)
slr.m1<-stats::lm(total ~ dbh, data=df)
## Example 1 -- store all the results to an object
out<-modresults(mod = slr.m1)
out$modsumm
out$sigma.e
out$press
out$tc.al.coef
out$vp.tc.al.coef
## Example 2
modresults(mod = slr.m1, print.output=TRUE)
```

---

obspredplot

*Observed versus predicted values plot.*

---

## Description

Creates a scatterplot between the observed values and the predicted ones from a fitted model.

## Usage

```
obspredplot(
  obs = obs,
  pred = pred,
  col = "black",
  linecol = "red",
  eng = TRUE,
  xlab = NULL,
  ylab = NULL,
  xlim = NULL,
  ylim = NULL,
  coef.max.val = 1.1,
  ...
)
```

## Arguments

<code>obs</code>	observed values of the variable of interest
<code>pred</code>	predicted values of the variable of interest
<code>col</code>	A string specifying the color of the data points. The default is "black".
<code>linecol</code>	A string specifying the straight line color. The default is set to "red".
<code>eng</code>	logical; if TRUE (by default), the language of the statistics will be in English; if "FALSE" will be in Spanish.
<code>xlab</code>	(optional) A string specifying X-axis label.
<code>ylab</code>	(optional) A string specifying Y-axis label.
<code>xlim</code>	(optional) A range (min,max) for the limits of the X-axis. By default is set to be 0 and the maximum value of the both observed and predicted values, multiplied by the option <code>coef.max.val</code> .
<code>ylim</code>	(optional) A range (min,max) for the limits of the Y-axis. By default is set to be 0 and the maximum value of the both observed and predicted values, multiplied by the option <code>coef.max.val</code> .
<code>coef.max.val</code>	(optional) A number to be used for multiplying the maximum value of the variable (either the observed or the predicted values). By default is set to 1.1.
<code>...</code>	other graphical parameters (see <code>par</code> and section 'Details' below).

## Details

Notice that the straight-line is draw using an intercept=0, and a slope=1.

Commonly used graphical parameters are: `col` The colors for lines and points. Multiple colors can be specified so that each point can be given its own color. If there are fewer colors than points they are recycled in the standard fashion. Lines will all be plotted in the first colour specified. `bg` a vector of background colors for open plot symbols, see `points`. Note: this is not the same setting as `par("bg")`. `pch` a vector of plotting characters or symbols: see `points`. `cex` a numerical vector giving the amount by which plotting characters and symbols should be scaled relative to the default. This works as a multiple of `par("cex")`. NULL and NA are equivalent to 1.0. Note that this does not affect annotation: see below. `lty` a vector of line types, see `par`. `cex.main`, `col.lab`, `font.sub`, etc settings for main- and sub-title and axis annotation, see `title` and `par`. `lwd` a vector of line widths, see `par`.

## Value

The function returns the above described graph.

## Author(s)

Christian Salas-Eljatib

## References

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. 170 p. <https://eljatib.com>

- Piñeiro G, Perelman S, Guerschman JP, Paruelo JM. 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecological Modelling 216(3-4):316-322 doi:[10.1016/j.ecolmodel.2008.05.006](https://doi.org/10.1016/j.ecolmodel.2008.05.006)

## Examples

```
df <- datana::maple
head(df)
m1 <- lm(leaf ~ dbh, data = df)
# Example 1, a residual plot
obspredplot(obs = df$leaf, pred = fitted(m1))
```

---

papersdocstu

*Productividad científica de estudiantes de postgrado*

---

## Description

Corresponde a un estudio realizado en la Universidad de Indiana, sobre el número de papers publicados por estudiantes egresados de programas de doctorado en bioquímica luego de 3 años.

## Usage

```
data(papersdocstu)
```

## Format

Este set de datos contiene las siguientes columnas:

**papers** Es el número de artículos científicos publicados luego de 3 años de egresado.

**genero** Hombre/mujer.

**est.civil** Estado civil del egresado.

**nin.men5** Número de hijos menores a 6 años que dependen del egresado.

**prog.prest** Puntaje asignado al prestigio del programa de postgrado.

**papers.guia** Número de papers publicados por el profesor(a) guía del egresado, en el mismo periodo de tiempo.

## Source

Los datos fueron obtenidos desde el paquete 'AER'.

## References

Long JS. 1997. The Origin of Sex Differences in Science.

**Examples**

```
data(papersdocstu)
df<-papersdocstu
head(df)
barplot(table(df$papers),xlab="Numero de papers publicados",
        ylab="Frecuencia (num. de estudiantes)")
table(df$genero)
table(df$est.civil,df$genero)
tapply(df$papers,df$est.civil,summary)
```

pesohojas

*Peso de hojas***Description**

Peso de hojas

**Usage**

pesohojas

**Format**

Data frame con 64 filas y 2 columnas:

**peso** peso foliar en gramos (g)

**area** área foliar en centímetros cuadrados (cm<sup>2</sup>)

**Examples**

```
data(pesohojas)
plot(peso~area, data = pesohojas)
```

plotrend

*Function for building a scatterplot with a superposing smoothed line***Description**

The function creates a scatterplot with a superposing smoothed line as a way to reveal any potential pattern between the variables.

**Usage**

```
plotrend(
  x = x,
  y = y,
  col = "black",
  linecol = "red",
  lwd = 2,
  xlab = NULL,
  ylab = NULL,
  ...
)
```

**Arguments**

<code>x</code>	A numeric vector representing the X-axis variable.
<code>y</code>	A numeric vector representing the Y-axis variable (response).
<code>col</code>	A string specifying the color of the data points. The default is "black".
<code>linecol</code>	A string specifying the smooth line color. The default is set to "red".
<code>lwd</code>	the width of the smooth line to be drawn. The default is set to 2.
<code>xlab</code>	(optional) A string specifying X-axis label.
<code>ylab</code>	(optional) A string specifying Y-axis label.
<code>...</code>	other graphical parameters (see <code>par</code> and section 'Details' below).

**Details**

Notice that the smoothed-line is derived from a rather standard algorithm (i.e., loess), implemented in the function `smoothfit`, thus it is only an approximation.

Commonly used graphical parameters are: `col` The colors for lines and points. Multiple colors can be specified so that each point can be given its own color. If there are fewer colors than points they are recycled in the standard fashion. Lines will all be plotted in the first colour specified. `bg` a vector of background colors for open plot symbols, see `points`. Note: this is not the same setting as `par("bg")`. `pch` a vector of plotting characters or symbols: see `points`. `cex` a numerical vector giving the amount by which plotting characters and symbols should be scaled relative to the default. This works as a multiple of `par("cex")`. `NULL` and `NA` are equivalent to 1.0. Note that this does not affect annotation: see below. `lty` a vector of line types, see `par`. `cex.main`, `col.lab`, `font.sub`, etc settings for main- and sub-title and axis annotation, see `title` and `par`. `lwd` a vector of line widths, see `par`.

**Value**

The function returns the above described graph.

**Author(s)**

Christian Salas-Eljatib

## References

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. 170 p. <https://eljatib.com>
- Salas C, Stage AR, and Robinson AP. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. Forest Science 54(1):107-122. doi:10.1093/forestscience/54.1.107

## Examples

```
df <- datana::maple
head(df)
m1<-lm(leaf~dbh,data=df)
# Example 1, a residual plot
plotrend(x=df$dbh,y=residuals(m1))
abline(h=0)
```

---

predstat

*Function to compute prediction statistics based on observed values*

---

## Description

Computes three statistics as a way to assess the prediction capabilities of a model, based on comparing observed versus predicted values of a response variable of interest. The statistics are: the root mean square differences (*RMSD*), the aggregated difference (*AD*), and the absolute aggregated differences (*AAD*). All of them are based on

$$r_i = y_i - \hat{y}_i$$

where  $y_i$  and  $\hat{y}_i$  are the observed and the predicted value of the response variable  $y$  for the  $i$ -th observation, respectively. Both the observed and predicted values must be expressed in the same units.

## Usage

```
predstat(
  obs = obs,
  pre = pre,
  decnum = 6,
  want.percent = FALSE,
  want.n = FALSE,
  ...
)
```

## Arguments

<code>obs</code>	observed values of the variable of interest
<code>pre</code>	predicted values of the variable of interest. That is to say, if the response variable of a fitted model has a transformation of the variable of interest, <code>pre</code> has to be back-transformed to the original units of the variable of interest.
<code>decnum</code>	the number of decimals to be used in the output. The default is set to 6.
<code>want.percent</code>	A logic option for requesting to also computed the prediction statistics as a percentage of the sample mean of <code>obs</code> . By default is set to FALSE.
<code>want.n</code>	A logic option to add the sample size <code>n</code> to the output. Bu default is set to FALSE
<code>...</code>	Passing all other potential options.

## Details

The function computes the three aforementioned statistics expressed in both (a) the units of the response variable and (b) the percentage. Notice that to represent each statistic in percentual terms, we divided them by the mean observed value of the response variable.

## Value

The main output depends on the `want.percent`; if TRUE, then it has the following six prediction statistics as a vector: (`rmsd`, `rmsd.p`, `ad`, `ad.p`, `aad`, `aad.p`); where `rmsd.p` stands for RMSD expressed as a percentage, and the same applies to `AD.p` and `AAD.p`. Meanwhile, if `want.percent=FALSE`, then it has the following three prediction statistics as a vector: (`rmsd`, `ad`, `aad`)

## Author(s)

Christian Salas-Eljatib.

## References

- Salas C, Ene L, Gregoire TG, Nasset E, Gobakken T. 2010. Modelling tree diameter from airborne laser scanning derived variables: a comparison of spatial statistical models. *Remote Sensing of Environment* 114(6):1277-1285. doi:[10.1016/j.rse.2010.01.020](https://doi.org/10.1016/j.rse.2010.01.020)
- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de roble-laurel-lingue. *Bosque* 23(2):81–92. doi:[10.4067/S071792002002000200009](https://doi.org/10.4067/S071792002002000200009).

## Examples

```
# Creates a fake dataframe
set.seed(1234)
df <- as.data.frame(cbind(Y=rnorm(30, 30,9), X=rexp(30,rate=0.9)))
head(df)
descstat(df)
# Fitting a candidate model
mod1 <- lm(Y~X, data=df)
#Using the predstat function
predstat(obs=df$Y,pre=fitted(mod1))
# note the units of these statistics are the same of the Y variable.
```



```
# If you want to add the statistics in percentual units.
predstat(obs=df$Y,pre=fitted(mod1),want.percent = TRUE)
# If some of the predicted values is missing (e.g. because of a
# missing predictor variable) the number of observations can be exported
df2 <- data.frame(yobs=df$Y,ypre=fitted(mod1))
df2[c(14,26), 2] <- NA
descstat(df2)
#Notice the different sample size
predstat(obs = df2$yobs, pre = df2$ypre, want.n = TRUE)
#Thus, only 28 observations are used, as it should be.
```

---

predstatmod

---

*Function to create a table with the prediction statistics by model.*


---

## Description

Creates a table with the prediction statistics for previously fitted models, based on the observed data.

## Usage

```
predstatmod(
  data = data,
  obs = "obs",
  pre = "pre",
  model = "model",
  want.by.valcl = FALSE,
  val.class = NA,
  want.all.outputs = FALSE,
  ...
)
```

## Arguments

<code>data</code>	a dataframe having the predicted and observed values of the response variable for a set of models.
<code>obs</code>	a character giving the column name of the response variable for the data.
<code>pre</code>	a character giving the column name of the predicted value for the response variable giving the predictor(s) variable(s) values for the data and the respective fitted model.
<code>model</code>	a character giving the column name for the name of previously fitted model(s).
<code>want.by.valcl</code>	A logical option for requesting to also computed the prediction statistics by validation classes, which are stored in the column defined in <code>val.class</code> . By default is set to FALSE.
<code>val.class</code>	If validation classes were assigned to each observation, this option corresponds to character giving the column name of the validation class. By default this option is set to NA, meaning is not available.

want.all.outputs

A logical option to save a full set of result elements in the output, thus the output is class list. By default is set to FALSE.

...

Passing all other potential options.

## Details

The function computes prediction statistics for a previously fitted model, and prepare an output summarizing the results to facilitate the comparison among models.

## Value

The main output is a table having as number of rows the total number of fitted models, and number of columns the statistics being computed. By default the statistics implemented in the predstat() function are computed.

## Author(s)

Christian Salas-Eljatib and Marcos Marivil.

## References

- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de roble-laurel-lingue. Bosque 23(2):81–92. doi:10.4067/S0717920020020000200009.

## Examples

```
#Creates a fake dataframe
set.seed(1234);
Y=rnorm(30, 30,9);X=rnorm(30, 450,133); Z=rbeta(30, .1,2)
df <- as.data.frame(cbind(Y, X,Z))
## Fitting some models
mod1 <- lm(Y~X, data=df)
mod2 <- lm(Y~X+I(X^2), data=df)
mod3 <- lm(Y~Z+I(X^2), data=df)
## Preparing the format of the input-data for the function
df.m1<-df.m2<-df.m3<-df
df.m1$model<-"mod1";df.m1$y.aju=fitted(mod1)
df.m2$model<-"mod2";df.m2$y.aju=fitted(mod2)
df.m3$model<-"mod3";df.m3$y.aju=fitted(mod3)
dfypredmod<-rbind(df.m1,df.m2,df.m3)
head(dfypredmod)
table(dfypredmod$model)
# Example
predstatmod(data=dfypredmod,obs="Y",pre="y.aju")
# Example but not including to report the percentage of the statistics
predstatmod(data=dfypredmod,obs="Y",pre="y.aju", want.percent=TRUE)
```

---

presenceIce	<i>Presence or absence of sea ice from logbook records of annual cruises</i>
-------------	--

---

**Description**

Data containing 52717 observations about presence of sea ice from logbook records of annual cruises to the B-C-B in an unbroken record between years 1850 to 1910.

**Usage**

```
data(presenceIce)
```

**Format**

The dataframe contains the following columns:

**ship.id** The code number for ships.

**move.type** Type of movement of ships. 0 indicates a sail-powered vessel and 1 indicates an auxiliary-powered vessel.

**year** Year of registry.

**month** Month of registry.

**day** Day of registry.

**lat.dec** Decimal latitude.

**long.dec** Decimal longitude.

**e.w** East or west of the Prime Meridian.

**ice.cov** Sea Ice Observed. 0 no see (Not registered) and 1 presence sea ice (Registered).

**Source**

The data were provided from Sea Ice Group at the Geophysical Institute.

**References**

Mahoney A, Bockstoce J, Botkin D, Eicken H, Nisbet R. 2011. Sea-Ice Distribution in the Bering and Chukchi Seas: Information from Historical Whaleships' Logbooks and Journals ARCTIC. 64(4): 465-477.

**Examples**

```
data(presenceIce)
head(presenceIce)
```

---

 president

*Eleccion presidencial del 2012 en Chile.*


---

### Description

Datos de mesa de la eleccion presidencial del 2012 en Chile. La eleccion se llevo a cabo el 19 de Diciembre del 2021.

### Usage

```
data(president)
```

### Format

Los datos contienen las siguientes columnas:

**region.no** Número de la region adminsitrativa de Chile.

**region** Nombre de la region administrativa de Chile

**provincia** Provincia.

**circu.senatorial** Circunscripcion senatorial.

**distrito** Distrit.

**comuna** County.

**circu.elec** Circunscripcion electoral.

**local** Local de votacion. Generalmente es un colegio.

**no.mesa** Número de mesa.

**tipo.mesa** Tipo de mesa de votacion.

**mesas.fusionadas** Mesa de votacion fucionada.

**electores** Electores.

**nro.en.voto** .

**candidato** Candidato, ya sea Gabriel Boric o Jose A. Kast

**votos.tricel** Número total de votos segun el TRICEL (Tribunal calificador de elecciones).

### Source

Los datos fueron obtenidos desde el sitio web del Servicio Electoral del Gobierno de Chilean (SERVEL) en <https://www.servel.cl>. El archivo de datos descargado el 24 de Octubre del 2022 tenia el nombre Resultados mesa presidencial TRICEL 2v 2021-1.xlsx.

### Examples

```
data(president)
head(president)
```

---

pressind	<i>Function to compute the PRESS statistics of a regression model .</i>
----------	---

---

**Description**

Computes the PRESS (Predicted residual sum of squares) statistics of a regression model.

**Usage**

```
pressind(model = model)
```

**Arguments**

model                    an object having the fitted model.

**Details**

The function computes the PRESS based on the Hat matrix of a previously fitted model.

**Value**

The main output is the PRESS statistics

**Author(s)**

Christian Salas-Eljatib.

**References**

- Myers RH. 1990. Classical and Modern Regression with Applications. Second Edition. Duxbury Classic Series, Pacific Grove, CA, USA.

**Examples**

```
#Creates a fake dataframe
set.seed(12)
df <- as.data.frame(cbind(Y=rnorm(30, 30,9), X=rnorm(30, 450,133)))
#fitting a candidate model
mod1 <- lm(Y~X, data=df)
#Using the `pressind` function
pressind(mod1)
```

---

primarias*Elección primaria para la presidencia de Chile*

---

**Description**

Datos a nivel de mesa de la votación para elecciones primarias para Presidente de Chile en 2021.

**Usage**

```
data(primarias)
```

**Format**

Este set de datos contiene las siguientes columnas:

**region.no** Región administrativa de Chile.

**region** Nombre de la región.

**provincia** Provincia.

**distrito** Distrito.

**comuna** Comuna.

**circu.elec** Circunscripción electoral.

**local** Local de votación.

**tipo.mesa** tipo de mesa.

**mesa** Código identificador de la mesa.

**mesas.fusionadas** Mesas fusionadas.

**nro.voto** .

**lista** Lista política del candidato.

**pacto** Pacto político del candidato.

**partido** Partido político del candidato.

**candidato** Nombre del candidato.

**votos** Número total de votos.

**Source**

Los datos fueron obtenidos desde el servicio electoral de Chile (SERVEL) en el web <https://www.servel.cl>. El nombre del archivo era Resultados Primarias Presidenciales 2021 CHILE.xlsx, y fue descargado el 4 de octubre del 2022. Los datos fueron ordenados, y solo aquellas filas que contenían información en la columna 'votos' son parte de la dataframe.

**Examples**

```
data(primarias)
head(primarias)
table(primarias$region)
table(primarias$region,primarias$candidato)
tapply(primarias$votos,primarias$candidato,sum)
```

---

pspLlancahue	<i>Ubicación cartesiana de árboles en un bosque (solo como referencia para uso del libro).</i>
--------------	--

---

## Description

Esta dataframe solo se mantiene para ser de utilidad a quienes usan el libro de Salas-eljatib (2021), ya que la nueva versión se encuentra en la dataframe llancahue2.

## Usage

```
data(pspLlancahue)
```

## Format

Contains tree-level variables, as follows:

**arb.id** Identificador del árbol.

**spp** Codificación de la especie. Detalles en llancahue2.

**dap** Diámetro a la altura del pecho, en cm.

**coord.x** Posición cartesiana en el eje-X, en m.

**coord.y** Posición cartesiana en el eje-Y, en m.

## Source

Detalles en llancahue2.

## References

- Detalles en llancahue2.

## Examples

```
data(pspLlancahue)
descstat(pspLlancahue$dap)
boxplot(dap~spp, data=pspLlancahue)
```

pspruca

*Tree spatial coordinates in the Rucamanque forest***Description**

Tree-level variables and spatial coordinates in a permanent sample plot of 1 ha (100 x 100m) in the Rucamanque experimental forest, near Temuco, Chile.

**Usage**

```
data(pspruca)
```

**Format**

The data frame contains four variables for the standing-alive trees as follows:

**tree** Tree number identification.

**spp** Codificación de la especie como sigue: "A. punctatum" es *Aextoxicon punctatum*, "E. cordifolia" es *Eucryphia cordifolia*, "G. avellana" es *Gevuina avellana*, "L. dentata" es *Lomatia dentata*, "L. philippiana" es *Laureliopsis philippiana*, "L. sempervirens" es *Laurelia sempervirens*, "N. obliqua" es *Nothofagus obliqua*, "P. lingue" es *Persea lingue*, y "Other" representa a cualquier especie distinta a cualquiera de las ya definidas.

**crown.class** Crown class (1: superior, 2: intermediate, 3; inferior)

**dbh** diameter at breast-height, in cm

**x.coord** Cartesian position at the X-axis, in m

**y.coord** Cartesian position at the Y-axis, in m

**Source**

Data were provided by Dr Christian Salas-Eljatib (Universidad de Chile, Santiago, Chile).

**References**

Salas C, LeMay V, Nunez P, Pacheco P, and Espinosa A. 2006. Spatial patterns in an old-growth *Nothofagus obliqua* forest in south-central Chile. *Forest Ecology and Management* 231(1-3): 38-46.  
doi:10.1016/j.foreco.2006.04.037

**Examples**

```
data(pspruca)
head(pspruca)
table(pspruca$species)
```



pspruca2

*Ubicación espacial de árboles en el bosque de Rucamanque***Description**

Medidas a nivel de árbol y coordenadas espaciales en un parcela de muestreo permanente de 1 ha (100 x 100m) en el bosque de Rucamanque, cerca de Temuco, Chile. Mayores antecedentes en las referencias.

**Usage**

```
data(pspruca2)
```

**Format**

Las columnas describen características de los árboles vivos en pie, como sigue:

**arbol** Número del árbol

**spp** Codificación de la especie como sigue: "A. punctatum" es *Aextoxicon punctatum*, "E. cordifolia" es *Eucryphia cordifolia*, "G. avellana" es *Gevuina avellana*, "L. dentata" es *Lomatia dentata*, "L. philippiana" es *Laureliopsis philippiana*, "L. sempervirens" es *Laurelia sempervirens*, "N. obliqua" es *Nothofagus obliqua*, "P. lingue" es *Persea lingue*, y "Other" representa a cualquier especie distinta a cualquiera de las ya definidas.

**clase.copa** Clase de copa (1: superior, 2: intermedio, 3; inferior)

**dap** Diámetro a la altura del pecho, en cm

**coord.x** Posicion cartesiana en el eje X, en m

**coord.y** Posicion cartesiana en el eje Y, en m

**Source**

Los datos fueron cedidos por el Dr Christian Salas-Eljatib (Santiago, Chile).

**References**

Salas C, LeMay V, Nunez P, Pacheco P, and Espinosa A. 2006. Spatial patterns in an old-growth *Nothofagus obliqua* forest in south-central Chile. *Forest Ecology and Management* 231(1-3): 38-46.  
doi:10.1016/j.foreco.2006.04.037

**Examples**

```
data(pspruca2)
table(pspruca2$spp)
```

ptaeda

*Height growth of Pinus taeda (Loblolly pine) trees***Description**

The Loblolly data frame has 84 rows and tree columns of records of the tree height growth of Loblolly pine trees. This dataframe is a slight modification to the original dataframe "Loblolly" from the *datasets* R package.

**Usage**

```
data(ptaeda, package="datana")
```

**Format**

A dataframe containing the following columns:

**seed.id** an ordered factor indicating the seed source for the tree. The ordering is according to increasing maximum height.

**age** a numeric vector of tree ages, in yr.

**toth** a numeric vector of tree heights, in m.

**Source**

Pinheiro, J. C. and Bates, D. M. (2000) Mixed-effects Models in S and S-PLUS. Springer.

**Examples**

```
data(ptaeda, package="datana")
head(ptaeda)
plot(toth ~ age, data = subset(ptaeda, seed.id == 329),
     xlab = "Age (yr)", las = 1,
     ylab = "Height (m)")
```

ptaeda2

*Crecimiento en altura de Pinus taeda***Description**

Esta dataframe contiene 84 folas y tres columnas de crecimiento en altura de árboles de Pinus taeda (Loblolly pine). Es una modificación de la dataframe "Loblolly" del paquete 'datasets' de R.

**Usage**

```
data(ptaeda2)
```

**Format**

Los datos contienen las siguientes columnas:

**semilla.id** Un factor indicando el origen de la semilla del árbol.

**edad** Edad del árbol, en años.

**atot** Altura total, en m.

**Source**

Pinheiro, J. C. and Bates, D. M. (2000) Mixed-effects Models in S and S-PLUS. Springer.

**Examples**

```
data(ptaeda2, package="datana")
head(ptaeda2)
plot(atot ~ edad, data = subset(ptaeda2, semilla.id == 329),
     xlab = "Edad (años)", las = 1,
     ylab = "Altura (m)")
```

---

pvalt

*Obtain the P-value for a Standard t-distributed random variable*

---

**Description**

Function to compute the P-value for a Standard t-distributed random variable.

**Usage**

```
pvalt(t.value, df, decnum = 14)
```

**Arguments**

t.value	A numeric random variable following a t-student pdf distribution.
df	degrees of freedom of the random variable following a t-student pdf distribution.
decnum	the number of decimals to be used in the output. The default is set to 5.

**Details**

It is suited to compute the P-value for any random variable following a Standard t probability density function (pdf). For instance, to obtain the p-value in a t-test.

**Value**

The function returns the P-value or probability of getting a value as large as t.value.

**Author(s)**

Christian Salas-Eljatib

**Examples**

```
# Load dataset
df <- datana::fertiliza2
head(df)
## Computes the t-test statistics (from the 'stats' package)
t.value <- stats::t.test(df$vol)
t.value
t.v <- as.numeric(t.value$statistic);t.v
deg.f <- as.numeric(t.value$parameter);deg.f

## Obtaining the p ##  pvalt(t.v,deg.f)
```

---

pvalz

---

*Obtain the P-value for a Standard Gaussian random variable*


---

**Description**

Function to computes the P-value for a Standard Gaussian random variable.

**Usage**

```
pvalz(zval, decnum = 5)
```

**Arguments**

zval	A numeric random variable following a Standard Gaussian distribution.
decnum	the number of decimals to be used in the output. The default is set to 5.

**Details**

It is suited to compute the P-value for any random variable following a Standard Gaussian probability density function.

**Value**

This function returns the P-value or probability of getting a value as large as 'zval'.

**Author(s)**

Christian Salas-Eljatib

**Examples**

```
pvalz(1.96)
```

---

qqgauss	<i>Function for producing a QQ plot for a Gaussian probability density function.</i>
---------	--

---

## Description

The function creates a QQ plot for a given random variable `y` and a Gaussian probability density function. This graph is a scatterplot between the sample quantiles of the data and the theoretical quantiles, following the Gaussian pdf.

## Usage

```
qqgauss(y = y, linecol = "red", xlab = NULL, ylab = NULL, eng = TRUE, ...)
```

## Arguments

<code>y</code>	A numeric vector representing the Y-random variable
<code>linecol</code>	A string specifying the 1:1 straight-line color. The default is set to "red".
<code>xlab</code>	(optional) A string specifying X-axis label. If not provide it, uses the default setting.
<code>ylab</code>	(optional) A string specifying Y-axis label. If not provide it, uses the default setting.
<code>eng</code>	logical; if TRUE (by default), the language of the statistics will be in English; if "FALSE" will be in Spanish.
<code>...</code>	other graphical parameters (see <code>par</code> and section 'Details' below).

## Details

Notice that the reference pdf model is the Gaussian one, i.e., uses the `qnorm()` function.

## Value

The function returns the above described graph.

## Author(s)

Christian Salas-Eljatib

## References

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. 170 p. <https://eljatib.com>

### Examples

```
df <- datana::maple
head(df)
m1<-lm(leaf~dbh,data=df)
# Example 1, a residual plot
qqgauss(residuals(m1))
```

---

rainfallCA

*Datos de precipitación en California*

---

### Description

Datos de precipitación medidos en distintos lugares de california, con las coordenadas de los puntos y su distancia a la costa.

### Usage

```
data(rainfallCA)
```

### Format

Este set de datos contiene las siguientes columnas:

**saple.id** Identificador del punto de muestreo.

**easting** Coordenada este del punto.

**northing** Coordenada norte del punto.

**pp** Precipitación, en pulgadas.

**ele** Elevación, en pies.

**lat** Latitud del punto.

**d.coast** Distancia a la costa, en millas.

### Source

Los datos provienen de mediciones hechas en California

### Examples

```
data(rainfallCA)
head(rainfallCA)
plot(pp~ele, data=rainfallCA)
hist(rainfallCA$pp)
```

---

rankmod*Function to produce a table ranks the models being analyzed.*

---

## Description

Prepare a table that ranks previously fitted models according to a series of statistics

## Usage

```
rankmod(  
  tabstatmod = tabstatmod,  
  all.refstat = TRUE,  
  refstat = refstat,  
  want.all.outputs = FALSE  
)
```

## Arguments

tabstatmod	a data object having the statistics values for each model. The first column must have the name of the model, and the other columns are the statistics.
all.refstat	A logic option to specify the statistics to be used for ranking the models. By default is to TRUE, implying that all the statistics provided in the object (i.e., columns) are going to be used for developing the ranking. If the option <code>all.refstat</code> is set to FALSE, then the option <code>refstat</code> must be specified.
refstat	A vector with the names of the columns (i.e., statistics) to be used. By the default is the name of all the columns of the object <code>tabstatmod</code> but not including the first column (it has the name of the model).
want.all.outputs	A logic option to save a full set of result elements in the output, thus the output is class <code>list</code> . By default is set to FALSE.

## Details

The function assigns scores to models based on an array of statistics already computed. For instance, the function can use the three prediction capabilities statistics RMSD, AD, and AAD, that are computed by the function `predstat()`. Nonetheless, any other statistics can be provided to the function. The only requirement is that the lower the value of the statistics (in absolute value), the better. The current version of the function is based upon the approach proposed by Salas (2002).

## Value

The main result is a table having assigned scores for each of the statistics, as well as the sum of the scores. The column ranking has between the best ranking (number 1) to the lowest ranking (number equal to the quantity of models being analyzed). The actual output is a list with two elements: (1) a dataframe sorted by the number of the model, and (2) a dataframe sorted by the ranking of the model.

**Author(s)**

Christian Salas-Eljatib and Marcos Marivil.

**References**

- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de roble-laurel-lingue. Bosque 23(2):81–92. doi:10.4067/S071792002002000200009.

**Examples**

```
##Creates a fake dataframe
set.seed(1234);y<-rnorm(30, 30,9);x<-rnorm(30, 450,133);z<-rbeta(30, .1,2)
db <- as.data.frame(cbind(y, x,z))
## Fitting some models
mod1 <- lm(y~x, data=db)
mod2 <- lm(y~x+I(x^2), data=db)
mod3 <- lm(y~z+I(x^2), data=db)
## Preparing the format of the input-data for the function
df.m1<-df.m2<-df.m3<-db
df.m1$model<-"mod1";df.m1$y.aju=fitted(mod1)
df.m2$model<-"mod2";df.m2$y.aju=fitted(mod2)
df.m3$model<-"mod3";df.m3$y.aju=fitted(mod3)
df<-rbind(df.m1,df.m2,df.m3)
head(df)
##Assign validation class
df<-assigncl(data=df,variable="y")
table(df$model)
table(df$y.class)
head(df)
##Computes prediction capabilities statistics
df.torank<-predstatmod(data=df,obs = "y", pre="y.aju",
  want.by.valcl = TRUE,val.class = "y.class")
df.torank
##Example 1: getting the main output, sorted by the ranking
rankmod(tabstatmod = df.torank)
##Example 2: only consider a portion of the available statistics
rankmod(tabstatmod = df.torank, all.refstat=FALSE,
  refstat=c("rmsd","ad","aad"))
```

---

raulihg

*Height growth of Nothofagus alpina trees in Chile.*

---

**Description**

Time series data of height for rauli (*Nothofagus alpina*) trees in south-central Chile. These sampled trees are part of the ones used in Salas-Eljatib (2021, Ecological Applications). The full citation is provided below.



**Usage**

```
data(raulihg)
```

**Format**

The data frame contains four variables as follows:

**tree.code** tree id code

**spp** species common name

**bha.t** breast-height age, in yrs.

**h.t** total height, in m.

**Source**

Data were provided by Dr Christian Salas-Eljatib (Santiago, Chile).

**References**

- Salas-Eljatib C. 2021. An approach to quantify climate-productivity relationships: an example from a widespread Nothofagus forest. Ecological Applications 31(4): e02285. [doi:10.1002/eap.2285](https://doi.org/10.1002/eap.2285)
- Salas-Eljatib, C. 2021. Time series height-data for Nothofagus alpina trees. [doi:10.6084/m9.figshare.13521602.v5](https://doi.org/10.6084/m9.figshare.13521602.v5)

**Examples**

```
data(raulihg)
head(raulihg)
```

---

raulihg2

*Crecimiento en altura de árboles de Nothofagus alpina.*

---

**Description**

Datos de series de tiempo de altura para árboles muestreados de Nothofagus alpina (raulí) en el centro-sur de Chile. Estos árboles son parte de los usados en Salas-Eljatib (2021, Ecological Applications). La cita completa se da en referencias.

**Usage**

```
data(raulihg2)
```

**Format**

Contiene variables de nivel individual, como se describen a continuacion::

- tree.code** Código del árbol
- spp** Nombre comun especie
- bha.t** Edad a la altura del pecho, en años.
- h.t** Altura total, en m.

**Source**

Datos cedidos por el Prof. Christian Salas-Eljatib.

**References**

- Salas-Eljatib C. 2021. An approach to quantify climate-productivity relationships: an example from a widespread Nothofagus forest. Ecological Applications 31(4): e02285. [doi:10.1002/eap.2285](#)
- Salas-Eljatib C. 2021. Time series height-data for Nothofagus alpina trees. [doi:10.6084/m9.figshare.13521602.v5](#)

**Examples**

```
data(raulihg2)
head(raulihg2)
```

---

rendesc2	<i>Rendimiento escolar por estudiante en Chile 2024</i>
----------	---

---

**Description**

Base de datos con información anónima de rendimiento escolar por estudiante, correspondiente al año 2024. Contiene 687033 observaciones de estudiantes de Enseñanza Media Humanístico Científica modalidad Jóvenes, pertenecientes a establecimientos municipales, particulares subvencionados y particulares pagados. Cada fila representa un estudiante y sus características básicas, incluyendo su promedio general, asistencia y situación final del curso.

**Usage**

```
data(rendesc2)
```

## Format

Variables se describen a continuación:

**region** Región de Chile del registro

**comuna** Comuna de la region correspondiente

**mrn** Identificador anónimo del estudiante

**cod.depe** Código de dependencia administrativa del establecimiento (1 = municipal, 2 = particular subvencionado, 3 = particular pagado)

**gen.alu** Género del estudiante (1 = hombre, 2 = mujer)

**edad.alu** Edad del estudiante

**prom.gral** Promedio general de notas (escala de 1.0 a 7.0)

**asistencia** Porcentaje de asistencia anual del estudiante

**sit.fin** Situación final del estudiante (P = promovido, R = reprobado)

## Source

Ministerio de Educación de Chile (MINEDUC), portal de datos abiertos: <https://datosabiertos.mineduc.cl/>. Los datos fueron digitados por Saúl Ketterer, estudiante del Prof. Christian Salas-Eljatib.

## References

- MINEDUC (2024). Datos de rendimiento por estudiante. Subsecretaría de Educación.

## Examples

```
data(rendesc2)
head(rendesc2)
```

---

simce2

*Puntaje SIMCE 2023 en matemática 4to Básico por RBD*

---

## Description

Puntaje promedio por establecimiento del SIMCE 2023 en matemática de 4to Básico. Se tienen 6534 observaciones. La variable binaria ( $Y$ ) es la presencia de convenio PIE en el establecimiento, donde  $Y = 1$  denota presencia y  $Y = 0$ , lo contrario.

## Usage

```
data(simce2)
```

**Format**

Variables se describen a continuación:

**rbd** Rol Base de Datos del establecimiento

**region** Región del establecimiento

**comuna** Comuna del establecimiento

**dependencia** Dependencia administrativa del establecimiento

**prom.mate4b** Puntaje promedio del establecimiento en la prueba de matemática del SIMCE de 4to básico en 2023

**mat.total** Cantidad de estudiantes matriculados en el establecimiento

**convenio.pie** Establecimiento tiene convenio PIE (1 si, 0 no)

**Source**

Datos obtenidos desde la Agencia de Calidad de la Educación del Mineduc y desde el portal de DatosAbiertos del Mineduc ([datosabiertos.mineduc.cl](https://datosabiertos.mineduc.cl)). Los datos fueron digitados por Diego Fernández, estudiante del Prof. Christian Salas-Eljatib.

**Examples**

```
data(simce2)
head(simce2)
```

---

simmeind

*Computes the simmetry index*

---

**Description**

Computes the simmetry index of a random variable.

**Usage**

```
simmeind(y = y, lt = lt)
```

**Arguments**

**y** is a numeric vector.

**lt** is the lower treshold used to collect the sample data represented by the vector y.

## Details

A more sensitive indicator of skewness is the the symmetry index, defined as the ratio between the mode and the 95th percentile of the observed distribution, as follows.

$$SimmI = \frac{y_{Mode} - y_{LT}}{y_{.95} - y_{LT}}$$

where  $y_{Mode}$ ,  $y_{LT}$  and  $y_{.95}$  are the mode of the distribution, the lower treshold of the variable, and the 95th percentile of the distribution.

According to Lorimer and Krug (1983) helps to distinguish between descending monotonic, skewed unimodal and symmetric unimodal curves. Negative exponential distributions have  $SimmI$  close to 0, Gaussian distribution have  $SimmI$  close to 0.5, and positively skewed unimodal curves have values intermediate between the two. Negatively skewed distributions have values  $> 0.5$ , with a theoretical maximum of 1.0.

## Value

The function returns the simmetry index, a numeric scalar.

## Author(s)

Christian Salas-Eljatib.

## References

- Lorimer CG. and Krug AG. 1983. Diameter Distributions in Even-aged Stands of Shade-tolerant and Midtolerant Tree Species. American Midland Naturalist 109 (2):331–345.

## Examples

```
library(datana)
data(casen)
head(casen)
df<-casen
#Compare
summary(df$edad)
mean(df$edad)
median(df$edad)
moda(df$edad)
simmeind(y=df$edad,lt = 0)
```

---

singleupp

Convert

---

## Description

Function to get the first letter only as upper case of a string.

**Usage**

```
singleupp(fac)
```

**Arguments**

fac                      is an object of class string or factor

**Details**

It is specially set to arrange an data vector having alphanumeric format.

**Value**

This function returns an object having the resulting string.

**Author(s)**

Christian Salas-Eljatib

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
singleupp("PETER")
cities <- c("bogota", "lima", "new york", "santiago", "madrid")
singleupp(cities)
```

---

skewn

*Computes the skewness of a numeric vector*

---

**Description**

The skewness is about the departure from symmetry of a frequency distribution. Therefore, It is about asymmetry. One way to assess asymmetry of a random variable is to compute an statistics representing its skewness. The current function an dimensionless statistics of the skewness of given vector.

**Usage**

```
skewn(x, na.rm = TRUE)
```

**Arguments**

x                      A numeric vector representing a random variable.  
na.rm                  Logical value to remove NA values. The default is set to TRUE.

**Details**

The skewness of a random variable is the third moment of the standardized variable. There are several ways of parameterizing an skewness estimator, such as depending on the third moment and the standard deviation of the random variable.

**Value**

The value of the the skewness of given vector

**Author(s)**

Christian Salas-Eljatib.

**Examples**

```
y.var<-rnorm(100);x.var<-rbeta(100,.2,2)
skewn(y.var)
skewn(x.var)
```

---

sludge

*Sludge data are at different cities, with a value of concentration zinc.*

---

**Description**

Dataset contains 36 observations

**Usage**

```
data(sludge)
```

**Format**

Contains four variables, as follows:

**city** Name of city.

**rate** Concentration rate of sludge.

**zinc** Value of concentration (in ppm).

**trt.comb** Combination between city and rate factors.

**Source**

The data were provided from.. still remember.

**References**

not yet

**Examples**

```
data(sludge2)
table(sludge$city, sludge$rate)
levels(sludge$city)
tapply(sludge$zinc, list(sludge$city, sludge$rate), mean)
```

---

sludge2

*Sludge data are at different cities, with a value of concentration zinc.*

---

**Description**

Datos de contenido de Zinc en el tratamiento de lodos

**Usage**

```
data(sludge2)
```

**Format**

Contiene las siguientes cuatro variables:

**ciudad** Nombre de la ciudad.

**tasa** Tasa de concentracion de lodo.

**zinc** Concentracion de Zinc, en ppm.

**trt.comb** Identificador de la combinacion de niveles entre los factores ciudad y tasa.

**Source**

The data were provided from.. still remember.

**References**

not yet

**Examples**

```
data(sludge2)
table(sludge2$ciudad, sludge2$tasa)
levels(sludge2$ciudad)
tapply(sludge2$zinc, list(sludge2$ciudad, sludge2$tasa), mean)
```



smoothfit

*Function to produce a smooth curve over a scatterplot***Description**

The function estimates a simple locally weighted regression. The main aim of the function is to be used to describe a graphical pattern between two variables in a scatterplot.

**Usage**

```
smoothfit(x, y, linecol = "red", lty = 2, ...)
```

**Arguments**

x	A numeric vector representing the X-axis variable.
y	A numeric vector representing the Y-axis variable (response).
linecol	A string specifying the smooth line color. The default is set to "red".
lty	The type of line to be draw according to the R plot parameters. By default is set to 2 which is a dashed line.
...	other graphical parameters (see par and section 'Details' below).

**Details**

Notice that both variables must be numeric.

**Value**

The function returns the above described curve, but in order to be used, you must first to create the scatterplot.

**Author(s)**

A somehow related version of this function was first created by Prof. Timothy Gregoire (Yale University), but the current version is due to Christian Salas-Eljatib.

**References**

- Weisberg S. Applied Linear Regression. 3rd edition. Wiley, New York, NY, USA. 310 p.

**Examples**

```
df <- datana::annualppCities
# Example 1
plot(annual.pp~year, data=df, col="gray")
smoothfit(x=df$year, y=df$annual.pp)
# Example 2
df2<-subset(df,city=="Chillan")
plot(annual.pp~year, data=df2, col="gray")
smoothfit(x=df2$year, y=df2$annual.pp, linecol="red",lwd=2)
```

---

snaspe	<i>On the National System of State Protected Wild Areas (SNASPE) of Chile.</i>
--------	--

---

## Description

Units of the National System of State Protected Wild Areas (SNASPE).

## Usage

```
data(snaspe)
```

## Format

Contains the following variables:

**unit.id** Number for the unit.

**unit** Name of the protected area.

**category** Category of the unit. It can be either a National Park, a National Reserve or a Natural Monument.

**county** Name of the county where the unit is located.

**province** Province where the unit is located.

**region** Region where the unit is located.

**perim.km** Perimeter, in km.

**area.ha** Area, in hectares.

**area.m2** Area, in m<sup>2</sup>.

## Source

These data are freely available at <https://ide.minagri.gob.cl>

## References

The Chilean SNASPE is under the direction of the Chilean Forest Service (CONAF). Further information and documentation can be found at <https://www.conaf.cl>

## Examples

```
data(snaspe)
head(snaspe)
table(snaspe$category)
tapply(snaspe$area.ha, snaspe$category, mean)
```

---

snaspe2*Sistema nacional de áreas protegidas del estado (SNASPE) de Chile*

---

**Description**

Contiene variables general de las unidades del sistema de áreas protegidas por el estado de Chile (SNASPE).

**Usage**

```
data(snaspe2)
```

**Format**

Contiene las siguientes variables para cada unidad del SNASPE:

**uni.id** Número indentificador de la unidad.

**unidad** Nombre de la unidad.

**categoria** Categoría de la unidad. Puede ser Parque Nacional, Reserva Nacional, o Monumento Natural.

**comuna** Nombre de la comuna donde esta la unidad.

**province** Nombre de la provincia donde esta la unidad.

**region** Nombre de la región.

**perim.km** Perímetro, en km.

**area.ha** Área, en hectáreas.

**area.m2** Área, en m<sup>2</sup>.

**Source**

Estos datos fueron obtenidos desde <https://ide.minagri.gob.cl>

**References**

EL SNASPE esta bajo la administración de la Corporación Nacional Forestal (CONAF) de Chile. Mayor información se puede encontrar en <https://www.conaf.cl>

**Examples**

```
data(snaspe2)
head(snaspe2)
table(snaspe2$categoria)
tapply(snaspe2$area.ha, snaspe2$categoria, mean)
```

---

soiltreat

*Soil treatment experiment in tree seedlings*


---

## Description

A test was made of the effect of three soil treatments on the height growth of 2-year-old seedlings. Treatments were assigned at random to the three plots within each of 11 blocks. Each plot was made up of 50 seedlings. Average 5-year height growth was the criterion for evaluating treatments.

## Usage

```
data(soiltreat)
```

## Format

Contains the four following columns, at the plot-level,

**block** Block unit.

**treat** Treatment level.

**ini.h** Initial height, in m.

**inc.h** Increment in height during 5-year, in m.

## Source

Table in page 71 of Freese (1967). The data were entered by Miss Nayeli Ramirez, a former student of Prof. Christian Salas-Eljatib.

## References

- Freese, F 1967. Elementary statistical methods for foresters. Agriculture Handbook 3171, USDA Forest Service.

## Examples

```
data(soiltreat)
head(soiltreat)
tapply(soiltreat$inc.h,soiltreat$treat,summary)
tapply(soiltreat$inc.h,soiltreat$treat,sd)
```

---

soiltreat2*Tratamientos del suelo en el crecimiento de plantulas.*

---

### Description

Un experimento sobre el efecto de tres tratamientos del suelo en el crecimiento en altura de plantulas de 2-años de edad. Los tratamientos fueron asignados aleatoriamente a tres parcelas dentro de cada uno de 11 bloques. Cada parcela esta constituida por hasta 50 plantulas. El promedio del incremento en altura de los últimos 5 años fue la variable de interes para evaluar los tratamientos.

### Usage

```
data(soiltreat2)
```

### Format

Los datos, a nivel de parcela, tienen las siguientes columnas,

**bloque** Bloque del experimento.

**tmo** Factor tratamiento, medido en tres niveles.

**alt.ini** Altura initial, rn m.

**alt.inc** Incremento en altura durante los últimos cinco años, en m.

### Source

Cuadro de la página 71 de Freese (1967). Los datos fueron digitados por la Srta. Nayeli Ramirez, una estudiante del Prof. Christian Salas-Eljatib.

### References

- Freese, F 1967. Elementary statistical methods for foresters. Agriculture Handbook 3171, USDA Forest Service.

### Examples

```
data(soiltreat2)
head(soiltreat2)
tapply(soiltreat2$alt.inc,soiltreat2$tmo,summary)
tapply(soiltreat2$alt.inc,soiltreat2$tmo,sd)
```

---

spataustria	<i>Tree locations for several plots of Norway spruce (Picea abies) in Austria</i>
-------------	---

---

## Description

The Austrian Research Center for Forests established a spacing experiment with Norway spruce (*Picea abies*) in the Vienna Woods. In the 'Hauersteig' experiment, several tree-level variables were measured within four sample plots over time. The current dataframe has only the measurements carried out in 1944.

## Usage

```
data(spataustria)
```

## Format

Contains cartesian position of trees, and covariates, in sample plots, as follows:

**plot** Plot number.

**tree** Tree number.

**species** Species code as follows: PCAB=*Picea abies*, LADC=*Larix decidua*, PNSY=*Pinus sylvestris*, FASY=*Fagus Sylvatica*, QCPE=*Quercus petraea*, BTPE=*Betula pendula*.

**x.coord** Cartesian position in the X-axis, in m.

**y.coord** Cartesian position in the Y-axis, in m.

**year** Measurement year.

**dbh** diameter at breast-height, in cm.

## References

- Kindermann G, Kristofel F, Neumann M, Rossler G, Ledermann T & Schueler.
1. 109 years of forest growth measurements from individual Norway spruce trees. Sci. Data 5:180077 doi:[10.1038/sdata.2018.77](https://doi.org/10.1038/sdata.2018.77)

## Examples

```
data(spataustria)
head(spataustria)
df<-spataustria
oldpar<-par(mar=c(4,4,0,0))
bord<-data.frame(
  x=c(min(df$x.coord),max(df$x.coord),min(df$x.coord),max(df$x.coord)),
  y=c(min(df$y.coord),min(df$y.coord),max(df$y.coord),min(df$y.coord))
)
plot(bord,type="n", xlab="x (m)", ylab="y (m)", asp=1, bty='n')
points(df$x.coord,df$y.coord,col=df$plot,cex=0.5)
par(oldpar)
```

---

tabtexanova	<i>Creates a LaTeX file having an ANOVA table for a previously fitted linear regression model</i>
-------------	---

---

## Description

Function to create a LaTeX file of an ANOVA table.

Function to create a LaTeX file for a table with the main fitting statistics from a fitted regression model.

## Usage

```
tabtexanova(
  mod = mod,
  nametab = nametab,
  cap = cap,
  save.file = FALSE,
  filename = "tabregre.tex",
  eng = TRUE,
  rowlab = "Source of variation",
  decnum = 3,
  font.size.tab = "normalsize",
  font.type.tab = "normalfont",
  ...
)
```

```
tabtexregre(
  mod = mod,
  nametab = nametab,
  cap = cap,
  save.file = FALSE,
  filename = "tabregre.tex",
  eng = TRUE,
  rowlab = "Parameter",
  decnum = 3,
  font.size.tab = "normalsize",
  font.type.tab = "normalfont",
  ...
)
```

## Arguments

mod	an object containing the fitted model by using the <code>lm()</code> function.
nametab	a string having a brief name to be used in both the label of the table and the file name. For instance, if "mod1", the table can be referred in your LaTeX document by using <code>\ref{tab:mod1}</code>

cap	a string having the caption of the LaTeX table.
save.file	The defaults is set to "FALSE", if is set to TRUE, then the option filename must be provided.
filename	A string having the name of the resulting LaTeX file having the table. The default is set to "tabdescdata.tex".
eng	The language to be used in the output. English is the default, meanwhile if eng=FALSE, Spanish is used.
rowlab	a character with the name to be used as label for the column where the variables will be printed. The default is set to "Parameter".
decnum	the number of decimals to be used in the output. The default is set to 3.
font.size.tab	The defaults is set to "normalsize". You could also try with "footnotesize".
font.type.tab	The defaults is set to "normalfont".
...	Other options of the main functions being used here.

### Details

The resulting file is a LaTeX table, that can be added to your main LaTeX document by using `\input{filename}`.

The resulting file is a LaTeX table, that can be added to your main LaTeX document by using `\input{filename}`.

### Value

This function creates a LaTeX file having an ANOVA table, from a fitted regression model.

This function creates a LaTeX file having the main fitting statistics of a linear regression model.

### Author(s)

Christian Salas-Eljatib.

### References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

### Examples

```
df <- datana::fishgrowth2
head(df)
descstat(df[,c("largo", "edad")])
plot(largo ~ edad, data=df)
mod1<-lm(largo ~ edad, data=df)
##example 1
tabtexanova(mod=mod1,nametab="anovatab",
cap="ANOVA-style table of the fitted regression model")
```



```
##example 2
tabtexanova(mod=mod1,nametab="anovatab",
cap="Cuadro estilo ANOVA para modelo de regresion ajustado",
eng=FALSE)

df <- datana::fishgrowth2
head(df)
datana::descstat(df[,c("largo","edad")])
graphics::plot(largo ~ edad, data=df)
mod1<-stats::lm(largo ~ edad, data=df)
## Example 1
tabtexregre(mod=mod1,nametab="basicmodel",
cap="Parameter estimates of the fitted regression model")
## Example 2
tabtexregre(mod=mod1,nametab="basicmodel",
cap="Cuadro con parametros estimados del modelo de regresion",
eng=FALSE)
```

---

tabtexdescstat

---

*Creates a LaTeX file having a descriptive statistics table for continuous variables*


---

## Description

Function to create a LaTeX file for a table of descriptive statistics of continuous variables from a dataframe.

## Usage

```
tabtexdescstat(
  data = data,
  colnames = colnames,
  varnames = varnames,
  cap = cap,
  nametab = nametab,
  save.file = FALSE,
  filename = "tabdescdata.tex",
  eng = TRUE,
  rowlab = "Variable",
  decnum = 3,
  font.size.tab = "normalsize",
  font.type.tab = "normalfont",
  landscape = FALSE,
  ...
)
```

**Arguments**

<code>data</code>	a dataframe containing numeric variables as columns.
<code>colnames</code>	a string having the column names of the dataframe to which the descriptive statistics will be computed.
<code>varnames</code>	a string having the name of each of the variables to be used in the LaTeX table.
<code>cap</code>	a string having the caption of the LaTeX table.
<code>nametab</code>	a string having a brief name to be used in both the label of the table and the file name. For instance, if <code>"=descdata"</code> , the table can be referred in your LaTeX document by using <code>\ref{tab:descdata}</code>
<code>save.file</code>	The defaults is set to <code>"FALSE"</code> , if is set to <code>TRUE</code> , then the option <code>filename</code> must be provided.
<code>filename</code>	A string having the name of the resulting LaTeX file having the table. The default is set to <code>"tabdescdata.tex"</code> .
<code>eng</code>	The language to be used in the output. English is the default, meanwhile if <code>eng=FALSE</code> , Spanish is used.
<code>rowlab</code>	a character with the name to be used as label for the column where the variables will be printed. The default is set to <code>"Variables"</code> .
<code>decnum</code>	the number of decimals to be used in the output. The default is set to 3.
<code>font.size.tab</code>	The defaults is set to <code>"normalsize"</code> . You could also try with <code>"footnotesize"</code> .
<code>font.type.tab</code>	The defaults is set to <code>"normalfont"</code> .
<code>landscape</code>	logical; this option is passed to the function <code>descstat()</code> that is call from this current function. By default is set to <code>FALSE</code> , thus the output table will have the statistics as rows, and in each column the variables. Otherwise, if <code>TRUE</code> the variables will be the rows, and each statistics the columns. Therefore this last option is only advisable when <code>full=FALSE</code> .
<code>...</code>	Other options of the main functions being used here.

**Details**

The resulting file is a LaTeX table, that can be added to your main LaTeX document by using `\input{filename}`.

**Value**

This function creates a LaTeX file having the following descriptive statistics: sample size, minimum, maximum, mean, median, SD, and coefficient of variation. If the full option is set to `TRUE`, the following statistics are added to the table: 25th and 75th percentiles, the interquartile range, skewness, and kurtosis.

**Author(s)**

Christian Salas-Eljatib.

## References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

## Examples

```
df <- datana::idahohd
head(df)
##! Example 1
tabtexdescstat(data=df,nametab="idaho",
  cap="Descriptive statistics table",
  colnames=c("dbh","toth"),varnames = c("Diameter","Height"))
##! Example 2
tabtexdescstat(data=df,nametab="idaho",
  cap="Cuadro con estadística descriptiva",
  colnames=c("dbh","toth"),varnames = c("Diametro","Altura"),
  eng=FALSE)
##! Example 3: variables as columns
tabtexdescstat(data=df,nametab="idaho",
  cap="Descriptive statistics table",colnames=c("dbh","toth"),
  varnames = c("Diameter","Height"),landscape=TRUE)
```

---

timeserplot

*Produces a time series plot*

---

## Description

Produces a time series plot, of variable 'y' as a function of 'x' by an observational unit factor.

## Usage

```
timeserplot(
  data = data,
  y = y,
  x = x,
  obs.unit = obs.unit,
  factor1 = NA,
  factor2 = NA,
  only.lines = FALSE,
  ylab = NA,
  xlab = NA,
  linetype.lab = NA,
  factor2.line = TRUE,
  factor2.col = FALSE,
  col.lines = "black",
  max.y.all = NA,
  levels.i.want = FALSE,
  col.lev.i.want = FALSE
)
```

**Arguments**

<code>data</code>	a dataframe with at least three columns representing the response variable ("y"), the main predictor variable ("x"), and a variable indicating the observational unit ("obs.unit").
<code>y</code>	a character giving the column name of the response variable or variable of interest.
<code>x</code>	a character giving the column name of the main predictor variable. Generally this variable is time.
<code>obs.unit</code>	a character giving the column name containing the info of the observational unit.
<code>factor1</code>	an optional character having the name of a column having a factor variable (e.g., treatment). The default value is set to NULL.
<code>factor2</code>	an optional character having the name of a column having another factor variable (e.g., species). The default value is set to NULL.
<code>only.lines</code>	a logic value if only lines, but not including dots, are going to be drawn in the plot. The default value is set to FALSE.
<code>ylab</code>	Label for the Y-axis
<code>xlab</code>	Label for the X-axis
<code>linetype.lab</code>	is an optional string to be used as the title of the factor being represented by lines. It is only needed if factor1 and factor2 are defined. See example.
<code>factor2.line</code>	a logic value if the second factor, factor2, is going to be segregated according to the type of lines. The default value is set to TRUE.
<code>factor2.col</code>	a logic value if the second factor, factor2, is going to be segregated according to the color of the lines only. The default value is set to FALSE.
<code>col.lines</code>	A string specifying the single color to be used for the lines of the timeseries
<code>max.y.all</code>	A number representing the maximum level of Y-axis for all classes
<code>levels.i.want</code>	A vector having the levels for the factor under study
<code>col.lev.i.want</code>	A vector having the colors to be used for the factor under study

**Details**

Both 'y' and 'x' must be numeric variables, and the column representing the observational unit, must be a factor. This factor identifies the longitudinal context of the data, for instance, a student being measured on time. Besides, two more factors can be added to the plotting details, in order to represent the potential variability among them.

**Value**

This function returns a time series plot

**Note**

Please, use the function with caution, and run first the examples to understand it better.

**Author(s)**

Christian Salas-Eljatib

**References**

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

**Examples**

```
data(ficdiamgr, package="datana")
df <- ficdiamgr
head(df)
str(df)
df$site<-as.factor(df$site)
df$species<-as.factor(df$species)
table(df$tree,df$species)
table(df$species,df$site)
#
timeserplot(df, y="dbh", x="time", obs.unit = "tree")
timeserplot(df, y="dbh", x="time", obs.unit = "tree", only.lines = TRUE)
#
## Otros ejemplos de uso de la funcion
timeserplot(df, y="dbh", x="time", obs.unit = "tree", col.lines = "blue",
only.lines = TRUE)
timeserplot(df, y="dbh", x="time", obs.unit = "tree", only.lines = FALSE)
#
timeserplot(df, y="dbh", x="time", obs.unit = "tree", factor1="site")
timeserplot(df, y="dbh", x="time", obs.unit = "tree", factor1="site",
factor2= "species")
timeserplot(df, y="dbh", x="time", obs.unit = "tree", factor1="site",
factor2= "species", factor2.col = TRUE, only.lines = TRUE)
```

treevol

*Diameter, height and volume for Black Cherry Trees***Description**

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. The records are a slight modification to the original dataframe "trees" from the *datasets* R package.

**Usage**

```
data(treevol)
```

**Format**

A data frame with 31 observations and three variables

**dbh** Diameter at breast height, in cm.

**toth** Total height, in m.

**vtot** Timber volume, in cubic meters.

**Source**

Ryan TA, Joiner BL, and Ryan BF. 1976. The Minitab Student Handbook. Duxbury Press.

**Examples**

```
pairs(treevol, panel = panel.smooth, main = "treevol dataframe")
plot(vtot ~ dbh, data = treevol, log = "xy")
coplot(log(vtot) ~ log(dbh) | toth, data = treevol,
       panel = panel.smooth)
summary(m1 <- lm(log(vtot) ~ log(dbh), data = treevol))
summary(m2 <- update(m1, ~ . + log(toth), data = treevol))
anova(m1, m2)
```

---

treevol2

*Volumen, altura, y diámetro para árboles de Black Cherry*


---

**Description**

Estos datos provienen de mediciones de volumen, altura y diámetro en 31 árboles volteados de black cherry (*Prunus serotina*). Son una modificación la dataframe 'trees' del paquete datasets de R.

**Usage**

```
data(treevol2)
```

**Format**

Datos con 31 observaciones y tres variables

**dap** diámetro a la altura del pecho, en cm

**atot** altural total, en m

**vtot** volumen total, en m<sup>3</sup>

**Source**

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) The Minitab Student Handbook. Duxbury Press.

## Examples

```
pairs(treevol2, panel = panel.smooth, main = "treevol dataframe")
plot(vtot ~ dap, data = treevol2, log = "xy")
coplot(log(vtot) ~ log(dap) | atot, data = treevol2,
       panel = panel.smooth)
summary(m1 <- lm(log(vtot) ~ log(dap), data = treevol2))
summary(m2 <- update(m1, ~ . + log(atot), data = treevol2))
anova(m1, m2)
```

---

treevolroble

*Tree volume of roble (Nothofagus obliqua) in the Rucamanque forest*


---

## Description

These are tree-level measurement data of sample trees in the Rucamanque experimental forest, near Temuco, in the Araucania region in south-central Chile, measured in 1999. The data are the same as in the dataframe "treevolruca", but only having observations for the species *Nothofagus obliqua* (roble).

## Usage

```
data(treevolroble)
```

## Format

Contains tree-level variables, as follows:

**tree.no** Tree id  
**dbh** Diameter at breast height, in cm  
**toth** Total height, in m.  
**d6** Upper-stem diameter at 6 m, in cm  
**totv** Tree gross volume, in m<sup>3</sup> with bark.

## Source

The data are provided courtesy of Dr Christian Salas at the Universidad de Chile (Santiago, Chile).

## References

- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de Roble-Laurel-Lingue. Bosque 23(2): 81-92. doi:10.4067/S0717920020020000200009 [https://eljatib.com/publication/2002-07-01\\_ajuste\\_y\\_validacion/](https://eljatib.com/publication/2002-07-01_ajuste_y_validacion/)

## Examples

```
data(treevolroble)
head(treevolroble)
```

---

treevolroble2	<i>Volumen a nivel de árbol para roble (Nothofagus obliqua) especie en el bosque de Rucamanque</i>
---------------	--

---

## Description

Volumen, altura y diámetro, entre otras para árboles muestra de *Nothofagus obliqua* (roble) en el bosque de Rucamanque, cerca de Temuco, en la región de la Araucanía, en el sur de Chile.

## Usage

```
data(treevolroble2)
```

## Format

Las siguientes columnas son parte de la dataframe:

**arbol** Número del árbol.

**especie** Especie.

**dap** Diámetro a la altura del pecho, en cm.

**atot** Altura total, en m.

**d6** Diámetro fustal a los 6 m, en cm.

**vtot** Volumen bruto total, en m<sup>3</sup> with bark.

## Source

Los datos son proporcionados por el Prof. Christian Salas (Universidad de Chile).

## References

- Salas C. 2002. Ajuste y validación de ecuaciones de volumen para un relicto del bosque de Roble-Laurel-Lingue. *Bosque* 23(2): 81-92. doi:10.4067/S0717920020020000200009 [https://eljatib.com/publication/2002-07-01\\_ajuste\\_y\\_validacion/](https://eljatib.com/publication/2002-07-01_ajuste_y_validacion/)

## Examples

```
data(treevolroble2)
head(treevolroble2)
```



---

upperleft	<i>convert the first n-characters of a string to upper-case letters.</i>
-----------	--

---

## Description

Function to upper-case the first n-characters of a string from the left-hand side.

## Usage

```
upperleft(fac, n = 1)
```

## Arguments

fac	is an object of class string or factor
n	is the number of characters to be converted of a the string given in fac.

## Details

It is specially set to arrange data vector having alphanumeric (i.e., letters) format.

## Value

This function returns an object having the first n-characters from the left-hand side in upper-case.

## Author(s)

Christian Salas-Eljatib

## References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

## Examples

```
fac.x<-"willkommen"
upperleft(fac.x)
upperleft(fac.x,n = 2)
upperleft(fac.x,2)
upperleft(fac.x,3)
#A longer vector of characters
fac.x<-c("willkommen","welcome","bem-vindo","bienvenido")
upperleft(fac.x,1)
```

---

vifx	<i>Computes the variance inflation factor (VIF) for a multiple linear regression (MLR) model.</i>
------	---

---

## Description

Function to compute the variance inflation factor (VIF) for a multiple linear regression model.

## Usage

```
vifx(mod = mod)
```

## Arguments

`mod` an object containing the fitted MLR model by using the `lm()` function.

## Details

The resulting out is a dataframe having the VIF for each of the predictor variables.

## Value

This function creates a LaTeX file having the main fitting statistics of a linear regression model.

## Author(s)

Christian Salas-Eljatib.

## References

Salas-Eljatib, C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor, Santiago, Chile. 170 p. <https://eljatib.com/rlibro>

## Examples

```
##Two fitted models
mod1 <- stats::lm(mpg ~ cyl+disp + hp + wt + drat, data = mtcars)
mod2 <- stats::lm(mpg ~ disp + hp + wt + drat, data = mtcars)
##The VIF values for each regression model
vifx(mod=mod1)
vifx(mod=mod2)
```

---

xyboxplot

---

*Function for building a scatterplot with superposing boxplots*


---

## Description

The function creates a scatterplot with superposing boxplots for the Y-axis variable segregated by classes (i.e., groups) of the X-axis variable. For a scatterplot between a response variable  $Y$  and a predictor variable  $X$ , this function superposes boxplots of the response by groups of the predictor variable. The main aim of the above described graph is to get a sense of the distribution of the response variable depending upon the predictor variable.

## Usage

```
xyboxplot(
  x = x,
  y = y,
  col.dots = "blue",
  transp.dots = 0.1,
  xlab = NULL,
  ylab = NULL,
  num.classes = 10,
  segre.type = "percentile",
  limi.classes = NA,
  x.category = FALSE,
  pch.dots = 19,
  col.box = "red",
  transp.bbox = 0.07,
  xlim = NA,
  ylim = NA,
  class.ticks.lwd = 1,
  class.ticks.col = "red",
  class.marks.col = "black",
  cex.dots = 0.7,
  class.marks = FALSE,
  class.ticks = TRUE
)
```

## Arguments

<code>x</code>	A numeric vector representing the X-axis variable.
<code>y</code>	A numeric vector representing the Y-axis variable (response).
<code>col.dots</code>	A string specifying the dot colors. The default value is "blue".
<code>transp.dots</code>	A numeric value to be used as transparency for the dots of the figure to be produced. The defaults is set to 0.2
<code>xlab</code>	(optional) A string specifying X-axis label.

<code>ylab</code>	(optional) A string specifying Y-axis label.
<code>num.classes</code>	The number of classes to be used for computing the prediction capabilities. The default is set to 10.
<code>segre.type</code>	A string specifying the type of segregation to build the classes. The types are: (a) <code>percentile</code> implies to segregate with the same amount, or close, of observations to each of the defined <code>num.classes</code> . (b) <code>user.defined</code> implies that the user must provided the limits of the <code>num.classes-1</code> . The default is set to <code>percentile</code> . Notice if <code>user.defined</code> is specified, the option
<code>limi.classes</code>	A vector of size <code>num.classes-1</code> containing the limits to be used for defining the classes.
<code>x.category</code>	A logical statement, if set to <code>TRUE</code> , the X-axis variable will be treated as categorical for the drawing of the boxplots. The default is set to <code>FALSE</code> .
<code>pch.dots</code>	A numeric factor altering the shape of the dots.
<code>col.box</code>	A string specifying the boxplot color. The default is "red"
<code>transp.boxp</code>	A numeric value to be used as transparency for the boxpot of the figure to be produced. The defaults is set to 0.1
<code>xlim</code>	(optional) A numeric vector having the minimum and maximum, respectively for the X-axis variable.
<code>ylim</code>	(optional) A numeric vector having the minimum and maximum, respectively for the Y-axis variable.
<code>class.ticks.lwd</code>	The numeric width of the tick line for each of the X-axis variable classes. By default is set to 1.
<code>class.ticks.col</code>	A string with the color of the tick line for each of the X-axis variable classes. By default is set to "red".
<code>class.marks.col</code>	A string with the color of the mark value for each of the X-axis variable classes. By default is set to "black".
<code>cex.dots</code>	A numeric factor altering the size of the dots. The default value is 0.7.
<code>class.marks</code>	Whether (logic: <code>TRUE</code> or <code>FALSE</code> ) the number value of each of the X-axis variable classes should be printed. By default is set to <code>FALSE</code> .
<code>class.ticks</code>	Whether (logic: <code>TRUE</code> or <code>FALSE</code> ) the number tick of each of the X-axis variable classes should be printed. By default is set to <code>TRUE</code> .

### Details

Notice that the superposing boxplots for the Y-axis variable are computed by grouping the X-axis variable in 10 classes. Those classes are set by computing the 0.1, 0.2, ..., 0.9-percentiles of the X-axis variable, therefore each group has the same number of observations. The wide of the boxplot represent the extend of the respective X-axis variable used for drawwing each boxplot.

### Value

The function returns the above described graph.

**Author(s)**

Christian Salas-Eljatib

**References**

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. 170 p. <https://eljatib.com>
- Salas C, Stage AR, and Robinson AP. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. Forest Science 54(1):107-122. doi:10.1093/forestscience/54.1.107

**Examples**

```
df <- datana::fishgrowth
xyboxplot(x=df$length,y=df$scale)
xyboxplot(x=df$length,y=df$scale,col.dots = "red",
xlab="Variable X")
xyboxplot(x=df$length,y=df$scale,xlab="Variable X")

## dots with alpha channel
xyboxplot(x=df$length,y=df$scale,xlab="Variable X",
transp.dots = 0.4)

## with categorical x
xyboxplot(x=df$age,y=df$length,x.category = TRUE)

## fixed x axis limits
xyboxplot(x=df$age,y=df$length,x.category = TRUE, xlim = c(0,10))

## x marks width to .5
xyboxplot(x=df$age,y=df$length,x.category = TRUE, xlim = c(0,10),
class.ticks.lwd = .5)

## x marks red and width 2
xyboxplot(x=df$age,y=df$length,x.category = TRUE, xlim = c(0,10),
class.ticks.lwd = 2, class.ticks.col = "red")

## larger dots
xyboxplot(x=df$age,y=df$length,x.category = TRUE, xlim = c(0,10),
cex.dots = 1.5)

## print classes ticks
xyboxplot(x=df$age,y=df$length,x.category = TRUE, xlim = c(0,10),
class.marks = FALSE, class.ticks.col = "green")

### the x-variable not recorded such as a categorical variable
df <- datana::fishgrowth
## print classes ticks, by default with red color
xyboxplot(x=df$length, y=df$scale)

## don't print ticks
```

```

xyboxplot(x=df$length, y=df$scale, class.ticks=FALSE)

## print classes marks values
xyboxplot(x=df$length, y=df$scale, class.marks=TRUE)

## print classes marks values without ticks
xyboxplot(x=df$length, y=df$scale, class.marks=TRUE, class.ticks=FALSE)

## change class marks and ticks colors
xyboxplot(x=df$length, y=df$scale, class.marks=TRUE,
          class.marks.col = "red",
          class.ticks.col = "blue")

## bigger ticks
xyboxplot(x=df$length, y=df$scale, class.marks=TRUE,
          class.marks.col = "red",
          class.ticks.col = "blue", class.ticks.lwd=3)

## Changing the number of the X-variable classes
xyboxplot(x=df$length,y=df$scale,num.classes=5)

## Defining the classes not by percentiles, but by fixed values
xyboxplot(x=df$length,y=df$scale,xlim=c(0,410),
          ylim=c(0,20),num.classes=4,
          segre.type="fixed",limi.classes=c(140,195,250))

## Note that the limits must be in agreement with the num.classes
xyboxplot(x=df$length,y=df$scale,xlim=c(0,410),ylim=c(0,20),
          num.classes=5,segre.type="fixed",limi.classes=c(100,160,200,250))

```

---

xyhist

*A scatterplot with marginal histograms*


---

## Description

The function produces a scatterplot between the 'y'-axis variable and the 'x'-axis variable, but also adding the marginal histograms for both variables.

## Usage

```

xyhist(
  x = x,
  y = y,
  col.x = "blue",
  col.y = "red",
  xlab = NULL,
  ylab = NULL,
  x.lim = NULL,
  y.lim = NULL
)

```

**Arguments**

<code>x</code>	A numeric vector representing the X-axis variable
<code>y</code>	A numeric vector representing the Y-axis variable
<code>col.x</code>	(optional) A string specifying the color of the histogram of the X-variable. Default is "blue".
<code>col.y</code>	(optional) A string specifying the color of the histogram of the Y-variable. Default is "red".
<code>xlab</code>	(optional) A string specifying X-axis label. Default is "xvar".
<code>ylab</code>	(optional) A string specifying Y-axis label. Default is "yvar".
<code>x.lim</code>	(optional) A vector of two elements with the limits of the Y-axis. Default is the range of the X-variable.
<code>y.lim</code>	(optional) A vector of two elements with the limits of the Y-axis. Default is the range of the Y-variable.

**Details**

Both the response variable (Y-axis) and the predictor variable (X-axis) must be numeric.

**Value**

The function returns the above described graph.

**Author(s)**

Christian Salas-Eljatib

**References**

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. <https://eljatib.com>

**Examples**

```
data(treevolroble)
df <- datana::treevolroble
head(df)
xyhist(x=df$dbh,y=df$toth)
xyhist(x=df$dbh,y=df$toth, xlab="Variable X", ylab="Variable Y")
xyhist(x=df$dbh,y=df$toth, xlab="Variable X", ylab="Variable Y",
  col.x = "gray",col.y="white")
```

---

xymultiplot

*Figure of a matrix of scatterplots and histograms for several variables.*


---

### Description

The function produces a panel of multiple scatterplots and histograms, showing the correlation coefficient among all pairs of variables. Notice that the data must contain only numeric variables.

### Usage

```
xymultiplot(
  x,
  smooth = TRUE,
  scale = FALSE,
  density = TRUE,
  digits = 2,
  method = "pearson",
  pch = 20,
  lm = FALSE,
  cor = TRUE,
  jiggle = FALSE,
  factor = 2,
  col.hist = "cyan",
  col.densi.curve = "black",
  show.points = TRUE,
  col.points = "gray",
  smoother = FALSE,
  col.smooth = "red",
  ellipses = FALSE,
  col.ellip = "blue",
  col.cent.point = "green",
  rug = TRUE,
  breaks = "Sturges",
  cex.cor = 1,
  ci = FALSE,
  alpha = 0.05,
  ...
)
```

### Arguments

x	is a dataframe containing all the numeric variables to be used for drawing the panel plot
smooth	a logical value for drawing smooth curves. The default is set to TRUE.
scale	scales the correlation font by the size of the absolute correlation. The default is set to FALSE.



density	a logical value for drawing a density curve. The default is set to TRUE.
digits	an optional numeric value for the digits to be used for drawing the correlation coefficient in the panel. Defaults is set to 2.
method	a string giving the method to be used for computing the correlation coefficient. Default is set to "pearson".
pch	The plot character (The default is 20, which looks like '.').
lm	Plot the linear fit rather than the LOESS smoothed fits. The default is FALSE.
cor	If plotting regressions, should correlations be reported? The default is TRUE.
jiggle	Should the points be jittered before plotting? The default is FALSE.
factor	factor for jittering (1-5), therefore only needed if "jiggle" is set to TRUE.
col.hist	a string giving the color to be used for the histograms of the panel. Default is set to "cyan".
col.densi.curve	a string with the name of the color to be used for the density curve. The default is set to "black".
show.points	a logical value for drawing the points in the scatter-plots. Defaults is set to TRUE.
col.points	a string giving the color to be used for the data points. Default is set to "gray".
smoother	If TRUE, then smooth.scatter the data points-slow but pretty with lots of subjects
col.smooth	a string giving the color to be used for the smoothed curve of the scatterplot. Default is set to "red".
ellipses	an optional logical value for drawing an ellipse for the scatter-plots. The default is set to FALSE.
col.ellip	a string giving the color to be used for the ellipse of the scatterplot. The default is set to "blue".
col.cent.point	a string giving the color to be used for the centroid point of the ellipse of the scatterplot. The default is set to "blue".
rug	a logical value for drawing the rugs in the histograms. Defaults is set to TRUE.
breaks	a string giving the method to be used for obtaining the breaks of the histogram. Defaults is set to "Sturges".
cex.cor	If this is specified, this will change the size of the text in the correlations. this allows one to also change the size of the points in the plot by specifying the normal cex values. If just specifying cex, it will change the character size, if cex.cor is specified, then cex will function to change the point size.
ci	Draw confidence intervals for the linear model or for the loess fit, defaults to ci=FALSE. If confidence intervals are not drawn, the fitting function is lowess.
alpha	an optional numeric value for the significance level. Defaults is set to 0.05.
...	other <a href="#">graphical parameters</a> (see <a href="#">par</a> and section 'Details' below).

## Details

Generates a multipanel (matrix) of scatterplots and histograms to explore potential relationships among variables.

**Value**

This function returns a multipanel of scatterplots and histograms

**Author(s)**

A modification of Christian Salas-Eljatib of the function `pairs.panels` of the package *psych*.

**References**

- Salas-Eljatib C. 2021. Análisis de datos con el programa estadístico R: una introducción aplicada. Ediciones Universidad Mayor. Santiago, Chile. <https://eljatib.com>

**Examples**

```
##First example
data(bears2)
head(bears2)
df <- bears2[,c('peso', 'edad', 'cabezaL', 'cabezaA', 'largo', 'pechoP')]
descstat(df)
xymultiplot(df)
xymultiplot(df, ellipse=TRUE)
xymultiplot(df, ellipses=TRUE, col.cent.point = "yellow",
  col.densi.curve = "dark green", col.hist = "white")
```