

Package ‘PASWR2’

January 20, 2025

Type Package

Title Probability and Statistics with R, Second Edition

Version 1.0.5

Date 2021-09-04

Maintainer Alan T. Arnholt <arnholtat@appstate.edu>

Description Functions and data sets for the text Probability and Statistics with R, Second Edition.

Depends R (>= 2.10), lattice, ggplot2

Imports e1071

LazyData TRUE

License GPL-2

URL <https://github.com/alanarnholt/PASWR2>,
<https://alanarnholt.github.io/PASWR2/>

BugReports <https://github.com/alanarnholt/PASWR2/issues>

RoxygenNote 7.1.1

NeedsCompilation no

Author Alan T. Arnholt [aut, cre]

Repository CRAN

Date/Publication 2021-09-04 20:10:02 UTC

Contents

PASWR2-package	4
AGGRESSION	4
APPLE	5
APTSIZE	6
BABERUTH	7
BAC	8
BATTERY	9

bino.gen	10
BIOMASS	10
BODYFAT	11
CALCULUS	12
CARS2004	13
checking.plots	14
CHIPS	15
CIRCUIT	16
cisim	16
COSAMA	18
COWS	19
DEPEND	20
DROSOPHILA	20
eda	21
ENGINEER	22
EPIDURAL	23
EPIDURALF	24
EURD	25
FAGUS	25
FCD	26
FERTILIZE	27
FOOD	28
FORMULA1	28
GD	29
GLUCOSE	30
GRADES	30
GROCERY	31
HARDWATER	32
HOUSE	33
HSWRESTLER	34
HUBBLE	35
INSURQUOTES	36
interval.plot	36
JANKA	37
KINDER	38
ksdist	39
ksldist	39
LEDDIODE	40
LOSTR	41
MILKCARTON	41
multiplot	42
NC2010DMG	43
normarea	44
nsize	45
ntester	46
oneway.plots	47
PAMTEMP	48
PHENYL	49

PHONE	50
RAT	50
RATBP	51
REFRIGERATOR	52
ROACHEGGS	52
SALINITY	53
SATFRUIT	54
SBIQ	55
SCHIZO	56
SCORE	56
SDS4	57
SIGN.test	58
SIMDATAST	60
SIMDATAXT	61
SOCCER	61
srs	62
STATTEMPS	63
STSCHOOL	64
SUNDIG	65
SUNFLOWER	65
SURFACESPAIN	66
SWIMTIMES	67
TENNIS	68
TESTSCORES	69
TIRE	69
TIREWEAR	70
TITANIC3	71
TOE	72
TOP20	73
tsum.test	73
twoway.plots	76
URLADDRESS	77
VIT2005	78
WAIT	79
WASHER	80
WATER	80
WCST	81
WEIGHTGAIN	82
WHEATSPAIN	83
WHEATUSA2004	83
wilcoxe.test	84
WOOL	86
z.test	87
zsum.test	90

PASWR2-package	<i>The PASWR2 Package</i>
----------------	---------------------------

Description

The PASWR2 Package

Details

Data sets and functions for *Probability and Statistics with R*, Second Edition.

AGGRESSION	<i>TV and Behavior</i>
------------	------------------------

Description

Data regarding the aggressive behavior in relation to exposure to violent television programs.

Usage

AGGRESSION

Format

A data frame with 16 observations on the following two variables:

- violence (an integer vector)
- noviolence (an integer vector)

Details

This is data regarding aggressive behavior in relation to exposure to violent television programs from Gibbons (1977) with the following exposition: "... a group of children are matched as well as possible as regards home environment, genetic factors, intelligence, parental attitudes, and so forth, in an effort to minimize factors other than TV that might influence a tendency for aggressive behavior. In each of the resulting 16 pairs, one child is randomly selected to view the most violent shows on TV, while the other watches cartoons, situation comedies, and the like. The children are then subjected to a series of tests designed to produce an ordinal measure of their aggression factors." (pages 143-144)

Source

Gibbons, J. D. (1977) *Nonparametric Methods for Quantitative Analysis*. American Science Press.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
AL <- reshape(AGGRESSION, varying = c("violence", "noviolence"),
v.names = "aggression", direction = "long")
ggplot(data = AL, aes(x = factor(time), y = aggression, fill = factor(time))) +
geom_boxplot() + labs(x = "") + scale_x_discrete(breaks = c(1, 2),
labels = c("Violence", "No Violence")) + guides(fill = "none") + scale_fill_brewer()
rm(AL)
with(data = AGGRESSION,
wilcox.test(violence, noviolence, paired = TRUE, alternative = "greater"))
```

APPLE

Apple Hardness

Description

An experiment was undertaken where seventeen apples were randomly selected from an orchard (fresh) and measured for hardness. Seventeen apples were also randomly selected from a warehouse (warehouse) where the apples had been stored for one week and measured for hardness.

Usage

APPLE

Format

A data frame with 34 observations on the following two variables:

- hardness (hardness rating measured in kg/meter² for both the fresh and warehouse apples)
- location (factor with two levels fresh and warehouse)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# ggplot2 approach
ggplot(data = APPLE, aes(sample = hardness)) + stat_qq() + facet_grid(. ~ location)
ggplot(data = APPLE, aes(sample = hardness, color = location)) + stat_qq()
ggplot(data = APPLE, aes(x = hardness, fill = location)) + geom_density(alpha = 0.4) +
scale_fill_brewer()
# lattice approach
qqmath(~hardness|location, data = APPLE)
qqmath(~hardness, group = location, type = c('p', 'r'), auto.key = TRUE, data = APPLE)
```

 APTSIZ

*Apartment Size***Description**

Size of apartments in Mendebaldea, Spain, and San Jorge, Spain

Usage

APTSIZE

Format

A data frame with 15 observations on the following two variables:

- size (apartment size in square meters)
- location (factor with two levels SanJorge and Mendebaldea)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
p <- ggplot(data = APTSIZ, aes(x = location, y = size, fill = location)) +
  labs(x = "", y = "Apartment size (square meters)") +
  scale_x_discrete(breaks = c("Mendebaldea", "SanJorge"),
  labels = c("Mendebaldea", "San Jorge")) + scale_fill_brewer()
p + geom_boxplot()
# remove the legend
p + geom_boxplot() + guides(fill = "none")
# violin plot
p + geom_violin(scale = 'area') + guides(fill = "none")
p + geom_violin(scale = 'count') + guides(fill = "none")
p + geom_violin() + geom_boxplot(width = 0.15, fill = 'black') + guides(fill = "none") +
  stat_summary(fun = median, geom = "point", fill = "white", shape = 23, size = 3)
# dotplot
p + geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 3) +
  guides(fill = "none")
p + geom_boxplot(width = 0.4) + geom_dotplot(binaxis = "y", stackdir = "center",
  binwidth = 3) + guides(fill = "none") + scale_fill_brewer(type = "qual", palette = 1)
# base graphics
boxplot(size ~ location, data = APTSIZ, col = c("red", "yellow"),
  ylab = "Apartment size (square meters)")
```

BABERUTH

George Herman Ruth

Description

Baseball statistics for George Herman Ruth (The Bambino or the Sultan of Swat)

Usage

BABERUTH

Format

A data frame with 22 observation of the following 14 variables:

- year (year in which the season occurred)
- team (team for which he played Bos-A, Bos-N, or NY-A)
- g (games played)
- ab (at bats)
- r (runs scored)
- h (hits)
- X2b (doubles)
- X3b (triples)
- hr (home runs)
- RBI (runs batted in)
- sb (stolen bases)
- bb (base on balls or walks)
- ba (batting average = h/ab)
- slg (slugging percentage = total bases/at bats)

Source

https://www.baseball-reference.com/about/bat_glossary.shtml

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = BABERUTH, aes(x = ba)) + geom_histogram(binwidth = 0.03) +  
facet_grid(team ~ .) + labs(x = "Batting average")  
ggplot(data = BABERUTH, aes(x = g, y = ab, color = rbi)) + geom_point() +  
labs(x = "Number of Games Played", y = "Times at Bat", color = "Runs\n Batted In",  
title = "George Herman Ruth")
```

 BAC

Blood Alcohol Content

Description

Two volunteers weighing 180 pounds each consumed a twelve ounce beer every fifteen minutes for one hour. One hour after the fourth beer was consumed, each volunteer's blood alcohol was measured with ten different breathalyzers from the same company. The numbers recorded in data frame BAC are the sorted blood alcohol content values reported with breathalyzers from company X and company Y.

Usage

BAC

Format

A data frame with 10 observations of the following 2 variables:

- X (blood alcohol content measured in g/L)
- Y (blood alcohol content measured in g/L)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = BAC,
var.test(X, Y, alternative = "less"))
# Convert data from wide to long format
# library(reshape2)
# BACL <- melt(BAC, variable.name = "company", value.name = "bac")
# ggplot(data = BACL, aes(x = company, y = bac, fill = company)) +
# geom_boxplot() + guides(fill = "none") + scale_fill_brewer() +
# labs(y = "blood alcohol content measured in g/L")
# Convert with reshape()
BACL <- reshape(BAC, varying = c("X", "Y"), v.names = "bac", timevar = "company",
direction = "long")
ggplot(data = BACL, aes(x = factor(company), y = bac, fill = factor(company))) +
geom_boxplot() + guides(fill = "none") + scale_fill_brewer() +
labs(y = "blood alcohol content measured in g/L", x = "") +
scale_x_discrete(breaks = c(1, 2), labels = c("Company X", "Company Y"))

# Base graphics
boxplot(BAC$Y, BAC$X)
```

BATTERY

Lithium Batteries

Description

A manufacturer of lithium batteries has two production facilities, A and B. Facility A batteries have an advertised life of 180 hours. Facility B batteries have an advertised life of 200 hours. Fifty randomly selected batteries from Facility A are selected and tested. Fifty randomly selected batteries from Facility B are selected and tested. The lifetimes for the tested batteries are stored in the variable `lifetime`.

Usage

BATTERY

Format

A data frame with 100 observations on the following two variables:

- `lifetime` (life time measured in hours)
- `facility` (factor with two levels A and B)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
p <- ggplot(data = BATTERY, aes(x = lifetime, color = facility))
p + geom_density()
q <- ggplot(data = BATTERY, aes(x = facility, y = lifetime))
q + geom_violin()
ggplot(data = BATTERY, aes(x = facility, y = lifetime, fill = facility)) +
  geom_violin() + scale_fill_brewer() + guides(fill = "none")
ggplot(data = BATTERY, aes(sample = lifetime)) + stat_qq() + facet_grid(. ~ facility)
ggplot(data = BATTERY, aes(sample = lifetime, color = facility)) + stat_qq()
# lattice approach
qqmath(~ lifetime|facility, data = BATTERY)
qqmath(~ lifetime, group = facility, type = c('p', 'r'), auto.key=TRUE, data = BATTERY)
```

bino.gen *Binomial Distribution Simulation*

Description

Function that generates and displays m repeated samples of n Bernoulli trials with a given probability of success

Usage

```
bino.gen(samples = 10000, n = 20, pi = 0.5)
```

Arguments

samples	number of repeated samples to generate
n	number of Bernoulli trials
pi	probability of success for each Bernoulli trial

Value

simulated.distribution	Simulated binomial distribution
theoretical.distribution	Theoretical binomial distribution

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```
bino.gen(samples=50000, n = 10, pi = 0.80)
```

BIOMASS *Beech Trees*

Description

Several measurements of 42 beech trees (*Fagus Sylvatica*) taken from a forest in Navarre (Spain)

Usage

```
BIOMASS
```

Format

A data frame with 42 observations on the following 4 variables:

- diameter (diameter of the stem in centimeters)
- height (height of the tree in meters)
- stemweight (weight of the stem in kilograms)
- aboveweight (aboveground weight in kilograms)

Source

Gobierno de Navarra and Gestion Ambiental Viveros y Repoblaciones de Navarra, 2006. The data were obtained within the European Project FORSEE.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
pairs(BIOMASS, col = "red", cex = 0.75)
plot(log(aboveweight) ~ log(diameter), data = BIOMASS)
# logarithmic axes
ggplot(data = BIOMASS, aes(x = diameter, y = aboveweight, color = log(stemweight))) +
  geom_point() + scale_x_log10() + scale_y_log10() +
  labs(x = "diameter of the stem in centimeters", y = "above ground weight in kilograms")
```

BODYFAT

Body Fat Composition

Description

Values from a study reported in the *American Journal of Clinical Nutrition* that investigated a new method for measuring body composition

Usage

BODYFAT

Format

A data frame with 18 observations on the following 3 variables:

- age (age in years)
- fat (percent body fat composition)
- sex (a factor with levels F for female and M for male)

Source

Mazess, R. B., Peppler, W. W., and Gibbons, M. (1984) "Total Body Composition by Dual-Photon (153 Gd) Absorptiometry." *American Journal of Clinical Nutrition*, **40**, **4**: 834-839.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# base graphics
boxplot(fat ~ sex, data = BODYFAT)
# ggplot2 approach
ggplot(data=BODYFAT, aes(x = sex, y = fat, fill = sex)) + geom_boxplot() +
labs(x = "", y = "Percent body fat") + scale_x_discrete(breaks=c("F", "M"),
labels =c("Female", "Male")) + guides(fill = "none") +
scale_fill_manual(values = c("red", "green"))
# Brewer Colors
ggplot(data=BODYFAT, aes(x = sex, y = fat, fill = sex)) + geom_boxplot() +
labs(x = "", y = "Percent body fat") + scale_x_discrete(breaks=c("F", "M"),
labels =c("Female", "Male")) + guides(fill = "none") + scale_fill_brewer()
ggplot(data=BODYFAT, aes(x = fat, fill = sex)) + geom_density(alpha = 0.4) +
scale_fill_brewer()
```

CALCULUS

Calculus Assessment Scores

Description

Mathematical assessment scores for 36 students enrolled in a biostatistics course according to whether or not the students had successfully completed a calculus course prior to enrolling in the biostatistics course

Usage

CALCULUS

Format

A data frame with 36 observations on the following 2 variables:

- score (assessment score for each student)
- calculus (a factor with levels NO and YES for students who did not and did successfully complete calculus prior to enrolling in the biostatistics course)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# ggplot2 approach
ggplot(data = CALCULUS, aes(sample = score)) + stat_qq() + facet_grid(. ~ calculus)
ggplot(data = CALCULUS, aes(x = calculus, y = score, fill = calculus)) + geom_boxplot() +
guides(fill = "none") + scale_fill_brewer()
ggplot(data = CALCULUS, aes(sample = score, color = calculus)) + stat_qq()
# lattice approach
qqmath(~score|calculus, data = CALCULUS)
qqmath(~score, group = calculus, type = c('p', 'r'), auto.key=TRUE, data = CALCULUS)
```

CARS2004

*Cars in the European Union (2004)***Description**

The numbers of cars per 1000 inhabitants (cars), the total number of known mortal accidents (deaths), and the country population/1000 (population) for the 25 member countries of the European Union for the year 2004

Usage

CARS2004

Format

A data frame with 25 observations on the following 4 variables:

- country (a factor with levels Austria, Belgium, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, and United Kingdom)
- cars (number of cars per 1000 inhabitants)
- deaths (total number of known mortal accidents)
- population (country population/1000)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
plot(deaths ~ cars, data = CARS2004)
ggplot(data = CARS2004, aes(x = population, y = deaths, color = cars)) + geom_point()
```

 checking.plots

Checking Plots

Description

Function that creates four graphs that can be used to help assess independence, normality, and constant variance

Usage

```
checking.plots(model, n.id = 3, COL = c("#0080FF", "#A9E2FF"))
```

Arguments

model	an aov or lm object
n.id	the number of points to identify
COL	vector of two colors

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[twoway.plots](#), [oneway.plots](#)

Examples

```
mod.aov <- aov(stopdist ~ tire, data = TIRE)
checking.plots(mod.aov)
rm(mod.aov)

# Similar graphs using ggplot2
#
mod.aov <- aov(stopdist ~ tire, data = TIRE)
fortify(mod.aov)
# library(gridExtra) used to place all graphs on the same device
p1 <- ggplot(data = mod.aov, aes(x = 1:dim(fortify(mod.aov))[1], y = .stdresid,
color = tire)) + geom_point() + labs(y = "Standardized Residuals",
x = "Ordered Residuals") + geom_hline(yintercept = c(-3,-2, 2, 3),
linetype = "dashed", col = "pink") + theme_bw()
p2 <- ggplot(data = mod.aov, aes(sample = .stdresid, color = tire)) +
stat_qq() + geom_abline(intercept = 0, slope = 1, linetype = "dashed", col = "pink") + theme_bw()
p3 <- ggplot(data = mod.aov, aes(x = .fitted, y = .stdresid, color = tire)) +
geom_point() + geom_hline(yintercept = 0, linetype = "dashed") +
labs(y = "Standardized Residuals", x = "Fitted Values") +
geom_hline(yintercept = c(-3, -2, 2, 3), linetype = "dashed", color = "pink") +
theme_bw()
p1
```

```
p2
p3
multiplot(p1, p2, p3, cols = 1)
# Or use the following (not run) to get all graphs on the same device
# library(gridExtra)
# grid.arrange(p1, p2, p3, nrow=3)
rm(mod.aov, p1, p2, p3)
```

CHIPS

Silicon Chips

Description

Two techniques of splitting chips are randomly assigned to 28 sheets so that each technique is applied to 14 sheets. The the number of usable chips from each silicon sheet is stored in the variable number.

Usage

CHIPS

Format

A data frame with 28 observations on the following 2 variables:

- number (number of usable chips from each silicon sheet)
- method (a factor with levels techniqueI and techniqueII)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# ggplot2 approach
ggplot(data = CHIPS, aes(sample = number)) + stat_qq() + facet_grid(. ~ method)
ggplot(data = CHIPS, aes(sample = number, color = method)) + stat_qq()
ggplot(data=BODYFAT, aes(x = fat, fill = sex)) + geom_density(alpha = 0.4) +
scale_fill_brewer()

# lattice approach
qqmath(~ number|method, data = CHIPS)
qqmath(~ number, group = method, type = c('p', 'r'), auto.key = TRUE, data = CHIPS)
```

CIRCUIT

Circuit Design Lifetime

Description

Results from an accelerated life test used to estimate the lifetime of four different circuit designs (lifetimes in thousands of hours)

Usage

CIRCUIT

Format

A data frame with 26 observations on the following 2 variables:

- lifetime (lifetimes in thousands of hours)
- design (a factor with levels DesignI and DesignII)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# ggplot2 approach
ggplot(data = CIRCUIT, aes(x = design, y = lifetime, fill = design)) + geom_boxplot() +
  labs(x = "", y = "Lifetime in thousands of hours") + guides(fill = "none") +
  scale_fill_brewer()
ggplot(data = CIRCUIT, aes(x = design, y = lifetime, fill = design)) + geom_violin() +
  labs(x = "", y = "Lifetime in thousands of hours") + guides(fill = "none") +
  scale_fill_brewer()
# Reorder the boxplots by medians
ggplot(data = CIRCUIT, aes(x = reorder(design, lifetime, FUN = median), lifetime,
  fill = design)) + geom_boxplot() + labs(x = "", y = "Lifetime in thousands of hours") +
  guides(fill = "none") + scale_fill_brewer()
```

cisim

Confidence Interval Simulation Program

Description

This program simulates random samples from which it constructs confidence intervals for either the population mean, the population variance, or the population proportion of successes.

Usage

```

cisim(
  samples = 100,
  n = 30,
  parameter = 0.5,
  sigma = 1,
  conf.level = 0.95,
  type = c("Mean", "Var", "Pi")
)

```

Arguments

<code>samples</code>	the number of samples desired.
<code>n</code>	the size of each sample
<code>parameter</code>	If constructing confidence intervals for the population mean or the population variance, <code>parameter</code> is the population mean (i.e., <code>type</code> is one of either "Mean" or "Var"). If constructing confidence intervals for the population proportion of successes, the value entered for <code>parameter</code> represents the population proportion of successes (P_i), and as such, must be a number between 0 and 1.
<code>sigma</code>	is the population standard deviation. <code>sigma</code> is not required if confidence intervals are of type "Pi".
<code>conf.level</code>	confidence level for the graphed confidence intervals, restricted to lie between zero and one
<code>type</code>	character string, one of "Mean", "Var", or "Pi", or just the initial letter of each, indicating the type of confidence interval simulation to perform

Details

Default is to construct confidence intervals for the population mean. Simulated confidence intervals for the population variance or population proportion of successes are possible by selecting the appropriate value in the `type` argument.

Value

Performs specified simulation and draws the resulting confidence intervals on a graphical device.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```

cisim(samples = 100, n = 30, parameter = 100, sigma = 10, conf.level = 0.90)
# Simulates 100 samples of size 30 from a normal distribution with mean 100
# and a standard deviation of 10. From the 100 simulated samples, 90% confidence
# intervals for the Mean are constructed and depicted in the graph.

cisim(100, 30, 100, 10, type = "Var")

```

```
# Simulates 100 sample of size 30 from a normal distribution with mean 100
# and a standard deviation of 10. From the 100 simulated samples, 95% confidence
# intervals for the variance are constructed and depicted in the graph.
```

```
cisim(100, 50, 0.5, type = "Pi", conf.level = 0.92)
# Simulates 100 samples of size 50 from a binomial distribution where the
# population proportion of successes is 0.5. From the 100 simulated samples,
# 92% confidence intervals for Pi are constructed and depicted in the graph.
```

COSAMA

Cosmed Versus Amatek

Description

The Cosmed is a portable metabolic system. A study at Appalachian State University compared the metabolic values obtained from the Cosmed to those of a reference unit (Amatek) over a range of workloads from easy to maximal to test the validity and reliability of the Cosmed. A small portion of the results for maximal oxygen consumption (VO₂ in ml/kg/min) measurements taken at a 150 watt workload are stored in COSAMA.

Usage

COSAMA

Format

A data frame with 14 observations on the following 3 variables:

- subject (subject number)
- cosmed (measured VO₂ with Cosmed)
- amatek (measured VO₂ with Amatek)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# ggplot2 approach
ggplot(data = COSAMA, aes(factor(1), y = cosmed - amatek)) + geom_boxplot() +
labs(x = "")
# Line Plots: First change data format from wide to long with melt() from reshape2
# library(reshape2)
# CA <- melt(COSAMA, id.vars = "subject", variable.name = "treatment",
# value.count = "VO2")
# ggplot(data = CA, aes(x = subject, y = value, color = treatment)) + geom_line()
# rm(CA)
```

```
# Convert to long format with reshape()
CA <- reshape(COSAMA, varying = c("cosmed", "amatek"), v.names = "V02",
timevar = "treatment", idvar = "subject", direction = "long")
ggplot(data = CA, aes(x = subject, y = V02, color = factor(treatment))) + geom_line() +
labs(color = "Treatment") + scale_color_discrete(labels = c("Cosmed", "Amatek"))
rm(CA)
# lattice approach
bwplot(~ (cosmed - amatek), data = COSAMA)
```

COWS

*Butterfat of Cows***Description**

Random samples of ten mature (five-years-old and older) and ten two-year-old cows were taken from each of five breeds. The average butterfat percentage of these 100 cows is stored in the variable `butterfat` with the type of cow stored in the variable `breed` and the age of the cow stored in the variable `age`.

Usage

COWS

Format

A data frame with 100 observations on the following 3 variables:

- `butterfat` (average butterfat percentage)
- `age` (a factor with levels 2 years old and Mature)
- `breed` (a factor with levels Ayrshire, Canadian, Guernsey, Holstein-Friesian, and Jersey)

Source

Canadian record book of purebred dairy cattle.

References

- Sokal, R. R. and Rohlf, F. J. 1994. *Biometry*. W. H. Freeman, New York, third edition.
- Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = COWS, aes(x = breed, y = butterfat, fill = age)) +
geom_boxplot(position = position_dodge(1.0)) +
labs(x = "", y = "Average butterfat percentage") + scale_fill_brewer()
summary(aov(butterfat ~ breed + age, data = COWS))
```

DEPEND	<i>Number of Dependent Children for 50 Families</i>
--------	---

Description

Number of dependent children for 50 randomly selected families

Usage

DEPEND

Format

A data frame with 50 observations on 1 variable:

- number (number of dependent children)

Source

Kitchens, L. J. 2003. *Basic Statistics and Data Analysis*. Duxbury.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
xtabs(~number, data = DEPEND)
ggplot(data = DEPEND, aes(x = factor(number))) +
  geom_bar(fill = "cornsilk", color = "orange") + labs(x = "Number of Dependent Children")
ggplot(data = DEPEND, aes(x = number)) + geom_density(fill = "pink", alpha = 0.3,
  color = "red") + labs(x = "Number of Dependent Children")
```

DROSOPHILA	<i>Drosophila Melanogaster</i>
------------	--------------------------------

Description

DROSOPHILA contains per diem fecundity (number of eggs laid per female per day for the first 14 days of life) for 25 females from each of three lines of *Drosophila melanogaster*. The three lines are Nonselected (control), Resistant, and Susceptible.

Usage

DROSOPHILA

Format

A data frame with 75 observations on the following 2 variables:

- fecundity (number of eggs laid per female per day for the first 14 days of life)
- line (a factor with levels Nonselected, Resistant, and Susceptible)

Source

The original measurements are from an experiment conducted by R. R. Sokal (*Biometry* by Sokal and Rohlf, 1994, p. 237).

References

- Sokal, R. R. and Rohlf, F. J. 1994. *Biometry*. W. H. Freeman, New York, third edition.
- Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = DROSOPHILA, aes(x = reorder(line, fecundity, FUN = median),
y = fecundity, fill = line)) + geom_boxplot() + guides(fill = "none") +
labs(y = "number of eggs laid per female \n per day for the first 14 days of life",
x = "") + scale_fill_brewer()
ggplot(data = DROSOPHILA, aes(x = reorder(line, fecundity, FUN = median),
y = fecundity, fill = line)) + geom_violin() + guides(fill = "none") +
labs(y = "number of eggs laid per female \n per day for the first 14 days of life",
x = "") + scale_fill_brewer()
summary(aov(fecundity ~ line, data = DROSOPHILA))
```

 eda

Exploratory Data Analysis

Description

Function that produces a histogram, density plot, boxplot, and Q-Q plot

Usage

```
eda(x, trim = 0.05, dec = 3)
```

Arguments

x	is a numeric vector where NAs and Infs are allowed but will be removed.
trim	is a fraction (between 0 and 0.5, inclusive) of values to be trimmed from each end of the ordered data such that if trim = 0.5, the result is the median.
dec	is a number specifying the number of decimals

Details

The function `eda()` will not return console window information on data sets containing more than 5000 observations. It will, however, still produce graphical output for data sets containing more than 5000 observations.

Value

Function returns various measures of center and location. The values returned for the quartiles are based on the default **R** definitions for quartiles. For more information on the definition of the quartiles, type `?quantile` and read about the algorithm used by `type = 7`.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```
eda(x = rnorm(100))
# Produces four graphs for the 100 randomly
# generated standard normal variates.
```

ENGINEER

Engineers' Salaries

Description

Salaries for engineering graduates 10 years after graduation

Usage

ENGINEER

Format

A data frame with 51 observations on the following 2 variables:

- salary (salary 10 years after graduation in thousands of dollars)
- university (one of three different engineering universities)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = ENGINEER, aes(x = university, y = salary, fill = university)) +
  geom_boxplot() + guides(fill = "none") + scale_fill_brewer() +
  labs(y = "salary 10 years after graduation \n in thousands of dollars")
# Violin Plots
ggplot(data = ENGINEER, aes(x = university, y = salary, fill = university)) +
  geom_violin() + guides(fill = "none") + scale_fill_brewer() +
  labs(y = "salary 10 years after graduation \n in thousands of dollars")
```

EPIDURAL

*Traditional Sitting Position Versus Hamstring Stretch Position***Description**

Initial results from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts (oc)

Usage

EPIDURAL

Format

A data frame with 85 observations on the following 7 variables:

- doctor (a factor with levels Dr. A, Dr. B, Dr. C, and Dr. D)
- kg (weight in kg of patient)
- cm (height in cm of patient)
- ease (a factor with levels Difficult, Easy, and Impossible indicating the physicians' assessments of how well bone landmarks could be felt in the patient)
- treatment (a factor with levels Hamstring Stretch and Traditional Sitting)
- oc (number of obstructive contacts)
- complications (a factor with levels Failure - person got dizzy, Failure - too many OCs, None, Paresthesia, and Wet Tap)

Source

Fisher, K. S., Arnholt, A. T., Douglas, M. E., Vandiver, S. L., Nguyen, D. H. 2009. "A Randomized Trial of the Traditional Sitting Position Versus the Hamstring Stretch Position for Labor Epidural Needle Placement." *Journal of Anesthesia & Analgesia*, Vol 109, No. 2: 532-534.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
xtabs(~ doctor + ease, data = EPIDURAL)
xtabs(~ doctor + factor(ease, levels = c("Easy", "Difficult", "Impossible")),
data = EPIDURAL)
```

EPIDURALF

*Traditional Sitting Position Versus Hamstring Stretch Position***Description**

Intermediate results from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts (oc)

Usage

EPIDURALF

Format

A data frame with 342 observations on the following 7 variables:

- doctor (a factor with levels Dr. A, Dr. B, Dr. C, and Dr. D)
- kg (weight in kg of patient)
- cm (height in cm of patient)
- ease (a factor with levels Difficult, Easy, and Impossible indicating the physicians' assessments of how well bone landmarks could be felt in the patient)
- treatment (a factor with levels Hamstring Stretch and Traditional Sitting)
- oc (number of obstructive contacts)
- complications (a factor with levels Failure - person got dizzy, Failure - too many OCs, None, Paresthesia, and Wet Tap)

Source

Fisher, K. S., Arnholt, A. T., Douglas, M. E., Vandiver, S. L., Nguyen, D. H. 2009. "A Randomized Trial of the Traditional Sitting Position Versus the Hamstring Stretch Position for Labor Epidural Needle Placement." *Journal of Anesthesia & Analgesia*, Vol 109, No. 2: 532-534.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = EPIDURALF, aes(x = treatment, y = oc, fill = treatment)) +
  geom_boxplot() + guides(fill = "none") + scale_fill_brewer() +
  labs(y = "number of obstructive contacts")
```

EURD

European Union Research and Development

Description

A random sample of 15 countries' research and development investments for the years 2002 and 2003 was taken, and the results in millions of Euros are stored in EURD.

Usage

EURD

Format

A data frame with 15 observations on the following 3 variables:

- country (a character vector with values Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, France, Hungary, Latvia, Lithuania, Malta, Portugal, Romania, Slovakia, and Slovenia)
- rd2002 (research and development investments in millions of Euros for 2002)
- rd2003 (research and development investments in millions of Euros for 2003)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = EURD, aes(x = rd2002, y = rd2003)) + geom_point() +  
  geom_smooth(method = "lm")  
ggplot(data = EURD, aes(sample = rd2003 - rd2002)) + stat_qq()  
# lattice approach  
qqmath(~ (rd2003 - rd2002), data = EURD, type = c("p", "r"))
```

FAGUS

Retained Carbon in Beech Trees

Description

The carbon retained by leaves measured in kg/ha is recorded for forty-one different plots of mountainous regions of Navarre (Spain), depending on the forest classification: areas with 90% or more beech trees (*Fagus Sylvatica*) are labeled monospecific, while areas with many species of trees are labeled multispecific.

Usage

FAGUS

Format

A data frame with 41 observations on the following 3 variables:

- plot (plot number)
- carbon (carbon retained by leaves measured in kg/ha)
- type (a factor with levels monospecific and multispecific)

Source

Gobierno de Navarra and Gestion Ambiental Viveros y Repoblaciones de Navarra, 2006. The data were obtained within the European Project FORSEE.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = FAGUS, aes(x = type, y = carbon)) + geom_boxplot()
```

FCD

Fat Cats

Description

In a weight loss study on obese cats, overweight cats were randomly assigned to one of three groups and boarded in a kennel. In each of the three groups, the cats' total caloric intake was strictly controlled (1 cup of generic cat food) and monitored for 10 days. The difference between the groups was that group A was given 1/4 of a cup of cat food every six hours, group B was given 1/3 a cup of cat food every eight hours, and group C was given 1/2 a cup of cat food every twelve hours. The weights of the cats at the beginning and end of the study were recorded, and the difference in weights (grams) was stored in the variable `Weight` of the data frame FCD.

Usage

FCD

Format

A data frame with 36 observations on the following 2 variables:

- weight (difference in weight (grams))
- diet (a factor with levels A, B, and C)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# checking.plots()?
p <- ggplot(data = FCD, aes(x = diet, y = weight))
p + geom_violin(fill = "blue")
aov(weight ~ diet, data = FCD)
```

FERTILIZE

Cross and Auto Fertilization

Description

Plants' heights in inches obtained from two seeds, one obtained by cross fertilization and the other by auto fertilization, in two opposite but separate locations of a pot are recorded.

Usage

FERTILIZE

Format

A data frame with 30 observations on the following 3 variables:

- height (height of plant in inches)
- fertilization (a factor with levels cross and self)
- pot (a factor with fifteen levels)

Source

Darwin, C. 1876. *The Effect of Cross and Self-Fertilization in the Vegetable Kingdom*. D. Appleton and Company.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
p <- ggplot(data = FERTILIZE, aes(x = height, color = fertilization))
p + geom_density()
t.test(height ~ fertilization, data = FERTILIZE)
```

FOOD

Carrot Shear

Description

Shear measured in kN on frozen carrots from four randomly selected freezers

Usage

FOOD

Format

A data frame with 16 observations on the following 2 variables:

- shear (carrot shear measured in kN)
- freezer (a factor with levels A, B, C, and D)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
summary(aov(shear ~ freezer, data = FOOD))
```

FORMULA1

Pit Stop Times

Description

Pit stop times for two teams at 10 randomly selected Formula 1 races

Usage

FORMULA1

Format

A data frame with 10 observations on the following 3 variables:

- race (number corresponding to a race site)
- team1 (pit stop times for team one)
- team2 (pit stop times for team two)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# Change data format from wide to long
# library(reshape2)
# F1L <- melt(data = FORMULA1, id.vars = "race", variable.name = "team",
# value.name = "time")
# ggplot(data = F1L, aes(x = team, y = time)) + geom_boxplot()
# Using reshape()
F1L <- reshape(FORMULA1, varying = c("team1", "team2"), v.names = "time",
timevar = "team", idvar = "race", direction = "long")
ggplot(data = F1L, aes(x = factor(team), y = time, fill = factor(team))) +
geom_boxplot() + guides(fill = "none") + scale_x_discrete(breaks = 1:2,
labels = c("Team 1", "Team 2")) + labs(x = "", y = "Pit stop times in seconds")
with(data = FORMULA1,
boxplot(team1, team2, col = c("red", "blue")))
```

GD

Times Until Failure

Description

Contains time until failure in hours for a particular electronic component subjected to an accelerated stress test

Usage

GD

Format

A data frame with 100 observations on the following variable:

- attf (times until failure in hours)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = GD, aes(x = attf, y = ..density..)) +
geom_histogram(binwidth = 2, fill = "cornsilk", color = "orange") +
geom_density(color = "gray", size = 1) + labs(x = "time until failure in hours")
```

GLUCOSE

Blood Glucose Levels

Description

Fifteen diabetic patients were randomly selected, and their blood glucose levels were measured in mg/100 ml with two different devices.

Usage

GLUCOSE

Format

A data frame with 15 observations on the following 3 variables:

- patient (patient number)
- old (blood glucose level in mg/100 ml using an old device)
- new (blood glucose level in mg/100 ml using a new device)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = GLUCOSE,  
      boxplot(old, new, col = c("red", "blue")))
```

GRADES

GPA and SAT Scores

Description

The admissions committee of a comprehensive state university selected, at random, the records of 200 second semester freshmen. The results, first semester college GPA and high school SAT scores, are stored in the data frame GRADES.

Usage

GRADES

Format

A data frame with 200 observations on the following 2 variables:

- sat (SAT score)
- gpa (grade point average)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# base scatterplot
plot(gpa ~ sat, data = GRADES)
# lattice scatterplot
xyplot(gpa ~ sat, data = GRADES, type = c("p", "smooth"))
# ggplot scatterplot
ggplot(data = GRADES, aes(x = sat, y = gpa)) + geom_point() + geom_smooth()
```

GROCERY

Grocery Spending

Description

The consumer expenditure survey, created by the U.S. Department of Labor, was administered to 30 households in Watauga County, North Carolina, to see how the cost of living in Watauga county with respect to total dollars spent on groceries compares with other counties. The amount of money each household spent per week on groceries is stored in the variable amount.

Usage

GROCERY

Format

A data frame with 30 observations on the following variable:

- amount (total dollars spent on groceries)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = GROCERY,
      z.test(amount, sigma.x = 25, mu = 100, alternative = "greater"))
hist(GROCERY$amount, xlab = "Weekly grocery bill", main = "")
ggplot(data = GROCERY, aes(x = amount, y = ..density..)) +
  geom_histogram(binwidth = 8, fill = "cornsilk", color = "gray80") +
  geom_density(color = "lightblue", size = 1, fill = "lightblue", alpha = .2) +
  labs(x = "Weekly grocery bill (in dollars)")
```

HARDWATER

Mortality and Water Hardness

Description

Mortality and drinking water hardness for 61 cities in England and Wales

Usage

HARDWATER

Format

A data frame with 61 observations on the following 4 variables:

- location (a factor with levels North and South indicating whether the town is as far north as Derby or further)
- town (the name of the town)
- mortality (average annual mortality per 100,000 males)
- hardness (calcium concentration (in parts per million))

Details

These data were collected in an investigation of environmental causes of disease. They show the annual mortality rate per 100,000 for males, averaged over the years 1958-1964, and the calcium concentration (in parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water.)

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski. 1994. *A Handbook of Small Datasets*. Chapman and Hall/CRC, London.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = HARDWATER, aes(x = hardness, y = mortality, color = location)) +  
  geom_point() + labs(y = "averaged annual mortality per 100,000 males",  
  x = "calcium concentration (in parts per million)")
```

HOUSE

House Prices

Description

Random sample of house prices (in thousands of dollars) for three bedroom/two bath houses in Watauga County, NC

Usage

HOUSE

Format

A data frame with 14 observations on the following 2 variables:

- neighborhood (a factor with levels Blowing Rock, Cove Creek, Green Valley, Park Valley, Parkway, and Valley Crucis)
- price (price of house in thousands of dollars)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = HOUSE,  
  t.test(price, mu = 225))
```

 HSWRESTLER

High School Wrestlers

Description

The body fat percentage of 78 high school wrestlers was measured using three separate techniques, and the results are stored in the data frame HSWRESTLER. The techniques used were hydrostatic weighing (hwfat), skin fold measurements (skfat), and the Tanita body fat scale (tanfat).

Usage

```
HSWRESTLER
```

Format

A data frame with 78 observations on the following 9 variables:

- age (age of wrestler in years)
- ht (height of wrestler in inches)
- wt (weight of wrestler in pounds)
- abs (abdominal fat)
- triceps (tricep fat)
- subscap (subscapular fat)
- hwfat (hydrostatic measure of percent fat)
- tanfat (Tanita measure of percent fat)
- skfat (skin fold measure of percent fat)

Source

Data provided by Dr. Alan Utter, Department of Health Leisure and Exercise Science, Appalachian State University

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
FAT <- c(HSWRESTLER$hwfat, HSWRESTLER$tanfat, HSWRESTLER$skfat)
GROUP <- factor(rep(c("hwfat", "tanfat", "skfat"), rep(78, 3)))
BLOCK <- factor(rep(1:78, 3))
friedman.test(FAT ~ GROUP | BLOCK)
rm(FAT, BLOCK, GROUP)
ggplot(data = HSWRESTLER, aes(x = tanfat, y = hwfat, color = age)) + geom_point() +
  geom_smooth() + labs(x = "Tanita measure of percent fat",
  y = "hydrostatic measure of percent fat")
```

HUBBLE

Hubble Telescope

Description

The Hubble Space Telescope was put into orbit on April 25, 1990. Unfortunately, on June 25, 1990, a spherical aberration was discovered in Hubble's primary mirror. To correct this, astronauts had to work in space. To prepare for the mission, two teams of astronauts practiced making repairs under simulated space conditions. Each team of astronauts went through 15 identical scenarios. The times to complete each scenario were recorded in days.

Usage

HUBBLE

Format

A data frame with 15 observations on the following 2 variables:

- team1 (days to complete scenario)
- team2 (days to complete scenario)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = HUBBLE,
      qqnorm(team1 - team2))
with(data = HUBBLE,
      qqline(team1 - team2))
# Trellis Approach
qqmath(~(team1 - team2), data = HUBBLE, type=c("p", "r"))
# ggplot approach
ggplot(data = HUBBLE, aes(sample = team1 - team2)) + stat_qq(color = "blue")
```

INSURQUOTES

Insurance Quotes

Description

Insurance quotes for two insurers of hazardous waste jobs

Usage

```
INSURQUOTES
```

Format

A data frame with 15 observations on the following 2 variables:

- companyA (quotes from company A in Euros)
- companyB (quotes from company B in Euros)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = INSURQUOTES, aes(sample = companyA - companyB)) +  
  stat_qq(col = "orange", size = 4)  
with(data = INSURQUOTES,  
  t.test(companyA, companyB))
```

interval.plot*Interval Plot*

Description

Function to graph intervals

Usage

```
interval.plot(l1, u1, parameter = 0)
```

Arguments

l1	vector of lower values
u1	vector of upper values
parameter	value of the desired parameter (used when graphing confidence intervals)

Value

Draws user-given intervals on a graphical device.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```
set.seed(385)
samples <- 100
n <- 625
ll <- numeric(samples)
ul <- numeric(samples)
xbar <- numeric(samples)
for (i in 1:samples){
  xbar[i] <- mean(rnorm(n, 80, 25))
  ll[i] <- xbar[i] - qnorm(.975)*25/sqrt(n)
  ul[i] <- xbar[i] + qnorm(.975)*25/sqrt(n)
}
interval.plot(ll, ul, parameter = 80)
```

JANKA

Australian Eucalypt Hardwoods

Description

The dataset consists of density and hardness measurements from 36 Australian Eucalypt hardwoods.

Usage

JANKA

Format

A data frame with 36 observations on the following 2 variables:

- `density` (a measure of density of the timber)
- `hardness` (the Janka hardness of the timber)

Details

Janka hardness is a structural property of Australian hardwood timbers. The Janka hardness test measures the force required to imbed a steel ball into a piece of wood.

Source

Williams, E.J. 1959. *Regression Analysis*. John Wiley & Sons, New York.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = JANKA, aes(x = density, y = hardness)) + geom_point() + geom_smooth()
```

KINDER

Kindergarten Class

Description

The data frame KINDER contains the height in inches and weight in pounds of 20 children from a kindergarten class.

Usage

```
KINDER
```

Format

A data frame with 20 observations on the following 2 variables:

- ht (height in inches of each child)
- wt (weight in pounds of each child)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = KINDER, aes(x = ht, y = wt)) + geom_point(color = "blue") +  
geom_smooth(method = "lm", color = "red") + labs(x = "height in inches",  
y = "weight in pounds")
```

ksdist *Simulated Distribution of D_n (Kolmogorov-Smirnov)*

Description

Function to visualize the sampling distribution of D_n (the Kolmogorov-Smirnov one sample statistic) and to find simulated critical values.

Usage

```
ksdist(n = 10, sims = 10000, alpha = 0.05)
```

Arguments

n	sample size
sims	number of simulations to perform
alpha	desired α level

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[ksldist](#)

Examples

```
ksdist(n = 10, sims = 15000, alpha = 0.05)
```

ksldist *Simulated Lilliefors' Test of Normality Values*

Description

Function to visualize the sampling distribution of D_n (the Kolmogorov-Smirnov one sample statistic) for simple and composite hypotheses

Usage

```
ksldist(n = 10, sims = 10000, alpha = 0.05)
```

Arguments

n	sample size
sims	number of simulations to perform
alpha	desired α level

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[ksdist](#)

Examples

```
ksldist(n = 10, sims = 1500, alpha = 0.05)
```

LEDDIODE

LED Diodes

Description

The diameter in millimeters for a random sample of 15 diodes from each of the two suppliers is stored in the data frame LEDDIODE.

Usage

```
LEDDIODE
```

Format

A data frame with 30 observations on the following 2 variables:

- diameter (diameter of diode measured in millimeters)
- supplier (factor with levels supplierA and supplierB)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = LEDDIODE, aes(supplier, diameter)) + geom_boxplot()
```

LOSTR

Lost Revenue Due to Worker Illness

Description

Data set containing the lost revenue in dollars/day and number of workers absent due to illness for a metallurgic company

Usage

LOSTR

Format

A data frame with 25 observations on the following 2 variables:

- numbersick (number of absent workers due to illness)
- lostrevenue (lost revenue in dollars)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = LOSTR, aes(x = numbersick, y = lostrevenue)) + geom_point(color = "red",  
pch = 21, fill = "pink", size = 4) + geom_smooth(method = "lm") +  
labs(x = "number of absent workers due to illness", y = "lost revenue in dollars")
```

MILKCARTON

Milk Carton Drying Times

Description

A plastics manufacturer makes two sizes of milk containers: half gallon and gallon sizes. The time required for each size to dry is recorded in seconds in the data frame MILKCARTON.

Usage

MILKCARTON

Format

A data frame with 80 observations on the following 2 variables:

- seconds (drying time in seconds)
- size (factor with levels halfgallon and wholegallon)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = MILKCARTON, aes(x = size, y = seconds)) + geom_boxplot()
ggplot(data = MILKCARTON, aes(x = size, y = seconds, fill = size)) + geom_boxplot() +
guides(fill = "none") + scale_fill_brewer() +
labs(x = "size of container", y = "drying time in seconds")
```

multiplot

Complex Plot Arrangements for ggplot Objects

Description

Arrange multiple ggplot objects on graphics device

Usage

```
multiplot(..., plotlist = NULL, cols = 1, layout = NULL)
```

Arguments

...	ggplot objects to be passed to the function
plotlist	a list of ggplot plots to plot
cols	number of columns in layout
layout	a matrix specifying the layout

Author(s)

Winston Chang <winston@stdout.org>

See Also

[layout](#)

Examples

```
p1 <- ggplot(data = HSWRESTLER, aes(x = skfat, y = hwfat)) + geom_point()
p2 <- ggplot(data = HSWRESTLER, aes(x = tanfat, y = hwfat)) + geom_point()
multiplot(p1, p2, cols = 2)
multiplot(p1, p2, cols = 2, layout=matrix(c(1, 0, 0, 2), byrow = TRUE, nrow = 2))
```

NC2010DMG

North Carolina Demographics

Description

North Carolina county demographics for 2010 and county voter information for the North Carolina Amendment 1 ballot initiative which took place May 8, 2012, are stored in the data frame NC2010DMG.

Usage

NC2010DMG

Format

A data frame with 100 observations (counties) on the following 32 variables:

- countyName (Name of North Carolina county)
- pop2010 (Total population of the county in 2010)
- medage (Median age of the county in 2010)
- divorced (Number of divorced adults in 2010)
- pctrural (The percent of the population that lived in a rural area of the county in 2010)
- edu_baorup (The total number of people with a Bachelor's degree in 2010)
- medinc (The median household income adjusted for inflation in 2010)
- col_enroll (The number of people enrolled in college in 2010)
- age18-24 (The number of people between the ages of 18 and 24 in the county in 2010)
- age25-29 (The number of people between the ages of 25 and 29 in the county in 2010)
- age60up (The number of people over the age of 60 in the county in 2010)
- white (The number of white people in the county in 2010)
- black (The number of black people in the county in 2010)
- MaleBachelor (The number of males with a Bachelor's degree in 2010)
- MaleMaster (The number of males with a Master's degree in 2010)
- MaleProfessional (The number of males with a professional degree in 2010)
- MaleDoctorate (The number of males with a Doctorate degree in 2010)
- FemaleBachelor (The number of females with a Bachelor's degree in 2010)
- FemaleMaster (The number of females with a Master's degree in 2010)
- FemaleProfessional (The number of females with a professional degree in 2010)
- FemaleDoctorate (The number of females with a Doctorate degree in 2010)
- Owneroccupied (The number of homes that are owner occupied in 2010)
- Renteroccupied (The number of homes that are renter occupied in 2010)

- popden (The number of people per square mile in 2010)
- pctfor (The percent of voters that voted for Amendment 1 on May 8, 2012)
- turnout (The percent of registered voters who voted May 8, 2012)
- obama08 (The percent of voters who voted for Barack Obama in the 2008 presidential election)
- mccain08 (The percent of voters who voted for John McCain in the 2008 presidential election)
- evanrate (Evangelical rates of adherence per 1,000 population in 2010)
- churches (The number of churches in the county in 2010)
- colleges (The number of colleges in the county in 2010)

Source

The original data was provided by E.L. Davison, Department of Sociology, Appalachian State University. Variables countyName through popden were obtained from <https://data.census.gov/cedsci/> and further cleaned by Maureen O'Donnell and Eitan Lees. The variables pctfor through mccain08 were obtained from <https://www.ncsbe.gov/>. The variables evanrate and churches were obtained from <https://thearda.com>, while the information for colleges was obtained from <https://collegestats.org/colleges/north-carolina/>.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = MILKCARTON, aes(x = size, y = seconds)) + geom_boxplot()
ggplot(data = MILKCARTON, aes(x = size, y = seconds, fill = size)) + geom_boxplot() +
  guides(fill = "none") + scale_fill_brewer() +
  labs(x = "size of container", y = "drying time in seconds")
```

normarea

Normal Area

Description

Function that computes and draws the area between two user specified values in a user specified normal distribution with a given mean and standard deviation

Usage

```
normarea(lower = -Inf, upper = Inf, m = 0, sig = 1)
```

Arguments

lower	the desired lower value
upper	the desired upper value
m	the mean for the population (default is the standard normal with $m = 0$)
sig	the standard deviation of the population (default is the standard normal with $\text{sig} = 1$)

Value

Draws the specified area in a graphics device

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```
# Finds and graphically illustrates  $P(70 < X < 130)$  given  $X$  is  $N(100, 15)$ 
normarea(lower = 70, upper = 130, m = 100, sig = 15)
```

nsize	<i>Required Sample Size</i>
-------	-----------------------------

Description

Function to determine required sample size to be within a given margin of error

Usage

```
nsize(b, sigma = NULL, p = 0.5, conf.level = 0.95, type = c("mu", "pi"))
```

Arguments

b	the desired bound
sigma	population standard deviation; not required if using type "pi"
p	estimate for the population proportion of successes; not required if using type "mu"
conf.level	confidence level for the problem, restricted to lie between zero and one
type	character string, one of "mu" or "pi", or just the initial letter of each, indicating the appropriate parameter; default value is "mu"

Details

Answer is based on a normal approximation when using type "pi".

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

Examples

```
nsize(b = 0.015, p = 0.5, conf.level = 0.95, type = "pi")
# Returns the required sample size (n) to estimate the population
# proportion of successes with a 0.95 confidence interval
# so that the margin of error is no more than 0.015 when the
# estimate of the population proportion of successes is 0.5.
nsize(b = 0.02, sigma = 0.1, conf.level = 0.95, type = "mu")
# Returns the required sample size (n) to estimate the population
# mean with a 0.95 confidence interval so that the margin
# of error is no more than 0.02.
```

ntester

Normality Tester

Description

Q-Q plots of randomly generated normal data of the same sample size as the tested data are generated and plotted on the perimeter of the graph while a Q-Q plot of the actual data is depicted in the center of the graph.

Usage

```
ntester(actual.data)
```

Arguments

`actual.data` is a numeric vector. Missing and infinite values are allowed, but are ignored in the calculation. The length of `actual.data` must be less than 5000 after dropping nonfinite values.

Details

Q-Q plots of randomly generated normal data of the same size as the tested data are generated and plotted on the perimeter of the graph sheet while a Q-Q plot of the actual data is depicted in the center of the graph. The p-values are calculated based on the Shapiro-Wilk W-statistic. Function will only work on numeric vectors containing less than or equal to 5000 observations. Best used for moderate sized samples ($n < 50$).

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

Shapiro, S.S. and Wilk, M.B. 1965. *An analysis of variance test for normality (complete samples)*. *Biometrika* **52**: 591-611.

Examples

```
ntester(actual.data = rexp(40, 1))
# Q-Q plot of random exponential data in center plot
# surrounded by 8 Q-Q plots of randomly generated
# standard normal data of size 40.
```

oneway.plots

Exploratory Graphs for Single Factor Designs

Description

Function to create dotplots, boxplots, and design plot (means) for single factor designs

Usage

```
oneway.plots(Y, fac1, COL = c("#A9E2FF", "#0080FF"))
```

Arguments

Y	response variable for a single factor design
fac1	predictor variable (factor)
COL	a vector with two colors

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[twoway.plots](#), [checking.plots](#)

Examples

```
with(data = TIRE, oneway.plots(stopdist, tire))
## Similar graphs with ggplot2
ggplot(data = TIRE, aes(tire, stopdist, fill = tire)) +
  geom_dotplot(binaxis = "y", stackdir = "center") + coord_flip() + theme_bw()
ggplot(data = TIRE, aes(tire, stopdist, fill = tire)) + geom_boxplot() +
  guides(fill = "none") + theme_bw()
```

PAMTEMP

Pamplona Temperatures

Description

The data frame PAMTEMP has records of the temperature and precipitation for Pamplona, Spain from January 1, 1990 to December 31, 2010.

Usage

PAMTEMP

Format

A data frame with 7547 observations on the following 7 variables:

- tmax (maximum daily temperature in Celsius)
- tmin (minimum daily temperature in Celsius)
- precip (daily precipitation in mm)
- day (day of the month)
- month (month of the year)
- year (year)
- tmean (the average of tmax and tmin)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
str(PAMTEMP)
levels(PAMTEMP$month)
PAMTEMP$month <- factor(PAMTEMP$month, levels = month.abb[1:12])
levels(PAMTEMP$month)
ggplot(data = PAMTEMP, aes(x = 1:dim(PAMTEMP)[1], y = tmean)) +
  geom_line() +
  theme_bw() +
  labs(x = "", y = "Average Temperature (Celcius)")
```

PHENYL

Phenylketonuria

Description

The data frame PHENYL records the level of Q10 at four different times for 46 patients diagnosed with phenylketonuria. The variable Q10.1 contains the level of Q10 measured in micromoles for the 46 patients. Q10.2, Q10.3, and Q10.4 are the values recorded at later times, respectively, for the 46 patients.

Usage

PHENYL

Format

A data frame with 46 observations on the following 4 variables:

- Q10.1 (level of Q10 at time 1 in micromoles)
- Q10.2 (level of Q10 at time 2 in micromoles)
- Q10.3 (level of Q10 at time 3 in micromoles)
- Q10.4 (level of Q10 at time 4 in micromoles)

Details

Phenylketonuria (PKU) is a genetic disorder that is characterized by an inability of the body to utilize the essential amino acid, phenylalanine. Research suggests patients with phenylketonuria have deficiencies in coenzyme Q10.

Source

Artuch, R., *et. al.* 2004. "Study of Antioxidant Status in Phenylketonuric Patients." *Clinical Biochemistry*, **37**: 198-203.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
PL <- stack(PHENYL)
PL$sub <- factor(rep(1:46, 4))
ggplot(data = PL, aes(x= ind, y = values, group = sub, color = sub)) + geom_line() +
  guides(color = FALSE)
with(data = PHENYL,
  t.test(Q10.1, conf.level = 0.99))
```

PHONE *Telephone Call Times*

Description

PHONE contains times in minutes of long distance telephone calls during a one month period for a small business.

Usage

PHONE

Format

A data frame with 23 observations on the following variable:

- `call.time` (time spent on long distance calls in minutes)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = PHONE,  
SIGN.test(call.time, md = 2.1))
```

RAT *Rat Survival Time*

Description

The survival time in weeks of 20 male rats exposed to high levels of radiation

Usage

RAT

Format

A data frame with 20 observations on the following variable:

- `survival.time` (number of weeks survived)

Source

Lawless, J. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = RAT, aes(sample = survival.time)) + stat_qq()  
ggplot(data = RAT, aes(x = survival.time)) + geom_density(alpha = 0.2, fill = "blue") +  
labs(x = "Survival time in weeks")
```

RATBP

Rat Blood Pressure

Description

Twelve rats were chosen, and a drug was administered to six rats, the treatment group, chosen at random. The other six rats, the control group, received a placebo. The drops in blood pressure (mmHg) for the treatment group (with probability distribution F) and the control group (with probability distribution G), respectively, were recorded.

Usage

RATBP

Format

A data frame with 12 observations on the following 2 variables:

- mmHg (drops in blood pressure in mm of Hg where positive values are decreases, negative values are increases)
- group (factor with levels control and treatment)

Source

The data is originally from Ott and Mendenhall (*Understanding Statistics*, 1985, problem 8.17).

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# Boxplot  
ggplot(data = RATBP, aes(x = group, y = mmHg)) + geom_boxplot()  
ggplot(data = RATBP, aes(x = group, y = mmHg, fill = group)) + geom_boxplot() +  
guides(fill = "none") + labs(x = "", y = "drops in blood pressure in mm of Hg") +  
scale_fill_brewer()
```

 REFRIGERATOR

Refrigerator Energy Consumption

Description

Sixty 18 cubic feet refrigerators were randomly selected from a company's warehouse. The first thirty had their motors modified while the last thirty were left intact. The energy consumption (kilowatts) for a 24 hour period for each refrigerator was recorded and stored in the variable `kilowatts`.

Usage

```
REFRIGERATOR
```

Format

A data frame with 60 observations on the following 2 variables:

- `kilowatts` (energy consumption in kilowatts for a 24 hour period)
- `group` (factor with levels `original` and `modified`)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# Boxplot
ggplot(data = REFRIGERATOR, aes(x = group, y = kilowatts)) + geom_boxplot()
ggplot(data = REFRIGERATOR, aes(x = group, y = kilowatts, fill = group)) +
  geom_boxplot() + labs(y = "energy consumption in kilowatts for a 24 hour period") +
  guides(fill = "none") + scale_fill_brewer()
```

 ROACHEGGS

Oriental Cockroaches

Description

A laboratory is interested in testing a new child friendly pesticide on *Blatta orientalis* (oriental cockroaches). Scientists apply the new pesticide to 81 randomly selected *Blatta orientalis* oothecae (eggs). The results from the experiment are stored in the data frame `ROACHEGGS` in the variable `eggs`. A zero in the variable `eggs` indicates that nothing hatched from the egg while a 1 indicates the birth of a cockroach.

Usage

```
ROACHEGGS
```

Format

A data frame with 81 observations on the following variable:

- eggs (numeric vector where a 0 indicates nothing hatched while a 1 indicates the birth of a cockroach.)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
p <- seq(0.1, 0.9, 0.001)
negloglike <- function(p){
  -(sum(ROACHEGGS$eggs)*log(p) + sum(1-ROACHEGGS$eggs)*log(1-p))
}
nlm(negloglike, .2)
rm(p, negloglike)
```

SALINITY

*Surface-Water Salinity***Description**

Surface-water salinity measurements were taken in a bottom-sampling project in Whitewater Bay, Florida.

Usage

```
SALINITY
```

Format

A data frame with 48 observations on the following variable:

- salinity (surface-water salinity measurements)

Source

Davis, J. 1986. *Statistics and Data Analysis in Geology*. John Wiley, New York.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# Boxplot
ggplot(data = SALINITY, aes(x = salinity)) + geom_density(fill = "yellow", alpha = 0.3)
```

SATFRUIT

Fruit Trees

Description

To estimate the total surface occupied by fruit trees in 3 small areas (R63, R67, and R68) of Navarre (Spain) in 2001, a sample of 47 square segments has been taken. The experimental units are square segments or quadrats of 4 hectares, obtained by random sampling after overlaying a square grid on the study domain.

Usage

SATFRUIT

Format

A data frame with 47 observations on the following 17 variables:

- `quadrat` (number of the sampled segment or quadrat)
- `smallarea` (the small area, a factor with levels R63, R67, and R68)
- `wheat` (area classified as wheat in the sampled segment)
- `barley` (area classified as barley in the sampled segment)
- `nonarable` (area classified as non-arable in the sampled segment)
- `corn` (area classified as corn in the sampled segment)
- `sunflower` (area classified as sunflower in the sampled segment)
- `vineyard` (area classified as vineyard in the sampled segment)
- `grass` (area classified as grass in the sampled segment)
- `asparagus` (area classified as asparagus in the sampled segment)
- `alfalfa` (area classified as alfalfa in the sampled segment)
- `rape` (area classified as rape in the sampled segment)
- `rice` (area classified as rice in the sampled segment)
- `almonds` (area classified as almonds in the sampled segment)
- `olives` (area classified as olives in the sampled segment)
- `fruit` (area classified as fruit trees in the sampled segment)
- `observed` (the observed area of fruit trees in the sampled segment)

Source

Militino, A. F., *et. al.* 2006. "Using Small Area Models to Estimate the Total Area Occupied by Olive Trees." *Journal of Agricultural, Biological and Environmental Statistics*, **11**: 450-461.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
pairs(SATFRUIT[,15:17])
```

SBIQ

County IQ

Description

A school psychologist administered the Stanford-Binet intelligence quotient (IQ) test in two counties. Forty randomly selected, gifted and talented students were selected from each county. The Stanford-Binet IQ test is said to follow a normal distribution with a mean of 100 and standard deviation of 16.

Usage

```
SBIQ
```

Format

A data frame with 80 observations on the following 2 variables:

- score (IQ score)
- county (factor with levels County1 and County2)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SBIQ, aes(sample = score, color = county)) + stat_qq()
```

 SCHIZO

Dopamine Activity

Description

Twenty-five patients with schizophrenia were classified as psychotic or nonpsychotic after being treated with an antipsychotic drug. Samples of cerebral fluid were taken from each patient and assayed for dopamine β -hydroxylase (DBH) activity. The dopamine measurements for the two groups are in nmol/ml-hour per milligram of protein.

Usage

SCHIZO

Format

A data frame with 25 observations on the following 2 variables:

- dopamine (dopamine activity level)
- classification (factor with levels psychotic and nonpsychotic)

Source

Sternberg, D. E., Van Kammen, D. P., and Bunney, W. E. 1982. "Schizophrenia: Dopamine β -Hydroxylase Activity and Treatment Response." *Science*, **216**: 1423-1425.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SCHIZO, aes(x = classification, y = dopamine)) + geom_boxplot()
```

SCORE

Standardized Test Scores

Description

Standardized test scores from a random sample of twenty college freshmen

Usage

SCORE

Format

A data frame with 20 observations on the following variable:

- scores (standardized test score)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SCORE, aes(sample = scores)) + stat_qq()
```

SDS4

*M1 Motorspeedway Times***Description**

The times recorded are those for 41 successive vehicles travelling northwards along the M1 motorway in England when passing a fixed point near Junction 13 in Bedfordshire on Saturday, March 23, 1985. After subtracting the times, the following 40 interarrival times reported to the nearest second are stored in SDS4 under the variable times.

Usage

```
SDS4
```

Format

A data frame with 40 observations on the following variable:

- times (interarrival times to the nearest second)

Source

Hand, D. J., *et. al.* 1994. *A Handbook of Small Data Sets*. Chapman & Hall, London.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SDS4, aes(x = times)) + geom_histogram(binwidth = 2)
ggplot(data = SDS4, aes(x = times, y = ..density..)) +
geom_histogram(binwidth = 2, color = "red", fill = "pink", alpha = 0.5) +
geom_density(fill = "cornsilk", alpha = 0.5) +
labs(x = "interarrival times to the nearest second", y = "")
```

SIGN.test

*Sign Test***Description**

This function will test a hypothesis based on the sign test and reports linearly interpolated confidence intervals for one sample problems.

Usage

```
SIGN.test(
  x,
  y = NULL,
  md = 0,
  alternative = "two.sided",
  conf.level = 0.95,
  ...
)
```

Arguments

x	numeric vector; NAs and Infs are allowed but will be removed.
y	optional numeric vector; NAs and Infs are allowed but will be removed.
md	a single number representing the value of the population median specified by the null hypothesis
alternative	is a character string, one of "greater", "less", or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, alternative refers to the true median of the parent population in relation to the hypothesized value of the median.
conf.level	confidence level for the returned confidence interval, restricted to lie between zero and one
...	further arguments to be passed to or from methods

Details

Computes a "Dependent-samples Sign-Test" if both x and y are provided. If only x is provided, computes the "Sign-Test."

Value

A list of class `htest_S`, containing the following components:

statistic	the S-statistic (the number of positive differences between the data and the hypothesized median), with names attribute "S".
p.value	the p-value for the test

<code>conf.int</code>	is a confidence interval (vector of length 2) for the true median based on linear interpolation. The confidence level is recorded in the attribute <code>conf.level</code> . When the alternative is not "two.sided", the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values k for which one would not reject the null hypothesis that the true mean or difference in means is k . Here infinity will be represented by <code>Inf</code> .
<code>estimate</code>	is a vector of length 1, giving the sample median; this estimates the corresponding population parameter. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	is the value of the median specified by the null hypothesis. This equals the input argument <code>md</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : "greater", "less", or "two.sided"
<code>data.name</code>	a character string (vector of length 1) containing the actual name of the input vector <code>x</code>
<code>Confidence.Intervals</code>	a 3 by 3 matrix containing the lower achieved confidence interval, the interpolated confidence interval, and the upper achieved confidence interval

Null Hypothesis

For the one-sample sign-test, the null hypothesis is that the median of the population from which x is drawn is md . For the two-sample dependent case, the null hypothesis is that the median for the differences of the populations from which x and y are drawn is md . The alternative hypothesis indicates the direction of divergence of the population median for x from md (i.e., "greater", "less", "two.sided".)

Assumptions

The median test assumes the parent population is continuous.

Note

The reported confidence interval is based on linear interpolation. The lower and upper confidence levels are exact.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

- Gibbons, J.D. and Chakraborti, S. 1992. *Nonparametric Statistical Inference*. Marcel Dekker Inc., New York.
- Kitchens, L.J. 2003. *Basic Statistics and Data Analysis*. Duxbury.
- Conover, W. J. 1980. *Practical Nonparametric Statistics, 2nd ed.* Wiley, New York.
- Lehmann, E. L. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Holden and Day, San Francisco.

See Also

[z.test](#), [zsum.test](#), [tsum.test](#)

Examples

```
with(data = PHONE, SIGN.test(call.time, md = 2.1))
# Computes two-sided sign-test for the null hypothesis
# that the population median is 2.1. The alternative
# hypothesis is that the median is not 2.1. An interpolated
# upper 95% upper bound for the population median will be computed.
```

SIMDATAST

Simulated Data (Predictors)

Description

Simulated data for five variables

Usage

SIMDATAST

Format

A data frame with 200 observations on the following 5 variables:

- y1 (a numeric vector)
- y2 (a numeric vector)
- x1 (a numeric vector)
- x2 (a numeric vector)
- x3 (a numeric vector)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SIMDATAST, aes(x = x1, y = y1)) + geom_point() + geom_smooth()
```

SIMDATAXT	<i>Simulated Data (Logarithms)</i>
-----------	------------------------------------

Description

Simulated data for four variables

Usage

SIMDATAXT

Format

A data frame with 200 observations on the following 4 variables:

- y1 (a numeric vector)
- y2 (a numeric vector)
- x1 (a numeric vector)
- x2 (a numeric vector)
- x3 (a numeric vector)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = SIMDATAXT, aes(x = x1, y = y)) + geom_point() + geom_smooth()
```

SOCCER	<i>World Cup Soccer</i>
--------	-------------------------

Description

SOCCER contains how many goals were scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002.

Usage

SOCCER

Format

A data frame with 575 observations on the following 3 variables:

- `cgt` (cumulative goal time in minutes - total time accumulated when a particular goal is scored)
- `game` (game in which goals were scored)
- `goals` (number of goals scored in regulation period)

Details

The World Cup is played once every four years. National teams from all over the world compete. In 2002 and in 1998, thirty-six teams were invited; whereas, in 1994 and in 1990, only 24 teams participated. The data frame `SOCGER` contains three columns: `cgt`, `game`, and `goals`. All of the information contained in Soccer is indirectly available from the FIFA World Cup website, located at <https://www.fifa.com/>.

Source

Chu, S. 2003. "Using Soccer Goals to Motivate the Poisson Process." *INFORMS Transaction on Education*, 3, 2: 62-68.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
xtabs(~goals, data = SOCCER)
```

srs

Simple Random Sample

Description

Computes all possible samples from a given population using simple random sampling

Usage

```
srs(popvalues, n)
```

Arguments

<code>popvalues</code>	are values of the population. NAs and Infs are allowed but will be removed from the population.
<code>n</code>	the sample size

Details

If non-finite values are entered as part of the population, they are removed; and the returned simple random sample computed is based on the remaining finite values.

Value

The function `srs()` returns a matrix containing the possible simple random samples of size `n` taken from a population of finite values `popvalues`.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[combn](#)

Examples

```
srs(popvalues = c(5, 8, 3, NA, Inf), n = 2)
```

STATTEMPS

Student Temperatures

Description

In a study conducted at Appalachian State University, students used digital oral thermometers to record their temperatures each day they came to class. A randomly selected day of student temperatures is provided in STATTEMPS. Information is also provided with regard to subject gender and the hour of the day when the students' temperatures were measured.

Usage

```
STATTEMPS
```

Format

A data frame with 34 observations on the following 3 variables:

- temperature (temperature in Fahrenheit)
- gender (a factor with levels Female and Male)
- class (a factor with levels 8 a.m. and 9 a.m.)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
p <- ggplot(data = STATTEMPS, aes(x = gender, y =temperature, fill = class))
p + geom_violin()
```

STSCHOOL

School Satisfaction

Description

A questionnaire is randomly administered to 11 students from State School x and to 15 students from State School y. The results have been ordered and stored in the data frame STSCHOOL.

Usage

STSCHOOL

Format

A data frame with 26 observations on the following 4 variables:

- x (satisfaction score)
- y (satisfaction score)
- satisfaction (combined satisfaction scores)
- school (a factor with levels x and y)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = STSCHOOL, t.test(x, y, var.equal=TRUE))
```

SUNDIG

Workstation Comparison

Description

To compare the speed differences between two different brands of workstations (Sun and Digital), the times each brand took to complete complex simulations were recorded. Five complex simulations were selected, and the five selected simulations were run on both workstations. The resulting times in minutes for the five simulations are stored in data frame SUNDIG.

Usage

SUNDIG

Format

A data frame with 5 observations on the following 3 variables:

- sun (time in seconds for a Sun workstation to complete a simulation)
- digital (time in seconds for a Digital workstation to complete a simulation)
- difference (difference between sun and digital)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = SUNDIG, t.test(sun, digital, paired=TRUE)$conf)
```

SUNFLOWER

Sunflower Defoliation

Description

Seventy-two field trials were conducted by applying four defoliation treatments (non-defoliated control, 33%, 66%, and 100%) at different growth stages (*stage*) ranging from pre-flowering (1) to physiological maturity (5) in four different locations of Navarre, Spain: Carcastillo (1), Melida (2), Murillo (3), and Unciti (4). There are two response variables: *yield* in kg/ha of the sunflower and *numseed*, the number of seeds per sunflower head. Data are stored in the data frame SUNFLOWER.

Usage

SUNFLOWER

Format

A data frame with 72 observations on the following 5 variables:

- location (a factor with levels A, B, C, and D for locations Carcastillo, Melida, Murillo, and Unciti, respectively)
- stage (a factor with levels stage1, stage2, stage3, stage4, and stage5)
- defoli (a factor with levels control, treat1, treat2, and treat3)
- yield (sunflower yield in kg/ha)
- numseed (number of seeds per sunflower head)

Source

Muro, J., *et. al.* 2001. “Defoliation Effects on Sunflower Yield Reduction.” *Agronomy Journal*, **93**: 634-637.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
summary(aov(yield ~ stage + defoli + stage:defoli, data = SUNFLOWER))
ggplot(data = SUNFLOWER, aes(numseed, yield, color = defoli)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + facet_grid(location ~ .)
```

SURFACESPAIN

Surface Area for Spanish Communities

Description

Surface area (km²) for seventeen autonomous Spanish communities.

Usage

SURFACESPAIN

Format

A data frame with 17 observations on the following 2 variables:

- community (a factor with levels Andalucia, Aragon,Asturias, Baleares, C.Valenciana, Canarias, Cantabria, Castilla-La Mancha, Castilla-Leon, Cataluna, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarre, and P.Vasco)
- surface (surface area in km²)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
# Base Graphs
with(data = SURFACESPAIN, barplot(surface, names.arg = community, las = 2))
# ggplot2
ggplot(data = SURFACESPAIN, aes(x = reorder(community, surface), y = surface)) +
  geom_bar(stat = "identity", fill = "yellow", color = "gold") + coord_flip() +
  labs(x = "", y = "squared kilometers")
# Trellis Approach
barchart(community ~ surface, data = SURFACESPAIN)
```

SWIMTIMES

Swim Times

Description

Swimmers' improvements in seconds for two diets are stored in the data frame SWIMTIMES. The values in seconds represent the time improvement in seconds for swimmers.

Usage

```
SWIMTIMES
```

Format

A data frame with 28 observations on the following 2 variables:

- seconds (time improvement in seconds)
- diet (a factor with levels lowfat and highfat)

Details

Times for the thirty-two swimmers for the 200 yard individual medley were taken right after the swimmers' conference meet. The swimmers were randomly assigned to follow one of the diets. One group followed a low fat diet the entire year but lost two swimmers along the way. The other group followed a high fat diet the entire year and also lost two swimmers.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
wilcox.test(seconds ~ diet, data = SWIMTIMES)
ggplot(data = SWIMTIMES, aes(x = diet, y = seconds, fill = diet)) + geom_violin() +
  guides(fill = "none") + scale_fill_brewer()
```

TENNIS

Speed Detector

Description

The Yonalasee tennis club has two systems to measure the speed of a tennis ball. The local tennis pro suspects one system (speed1) consistently records faster speeds. To test her suspicions, she sets up both systems and records the speeds of 12 serves (three serves from each side of the court). The values are stored in the data frame TENNIS in the variables speed1 and speed2. The recorded speeds are in kilometers per hour.

Usage

TENNIS

Format

A data frame with 12 observations on the following 2 variables:

- speed1 (speed in kilometers per hour)
- speed2 (speed in kilometers per hour)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data = TENNIS, boxplot(speed1, speed2))
```

TESTSCORES

Statistics Grades

Description

Test grades of 29 students taking a basic statistics course

Usage

TESTSCORES

Format

A data frame with 29 observations on the following variable:

- grade (test score)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = TESTSCORES, aes(x = grade)) + geom_histogram(binwidth = 5,  
fill = "cornsilk", color = "gray60", alpha = 0.7)
```

TIRE

Stopping Distance

Description

The data frame TIRE has the stopping distances measured to the nearest foot for a standard sized car to come to a complete stop from a speed of sixty miles per hour. There are six measurements of the stopping distance for four different tread patterns labeled A, B, C, and D. The same driver and car were used for all twenty-four measurements.

Usage

TIRE

Format

A data frame with 24 observations on the following 3 variables:

- stopdist (stopping distance measured to the nearest foot)
- tire (a factor with levels A, B, C, and D)
- order (order the experiment was conducted)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = TIRE, aes(x = reorder(tire, stopdist, FUN = median), y = stopdist,
  fill = tire)) + geom_boxplot() + guides(fill = "none") +
labs(y = "Stopping distance in feet", x = "Tire Brand") + scale_fill_brewer()
summary(aov(stopdist ~ tire, data = TIRE))
p <- ggplot(data = TIRE, aes(x = reorder(tire, stopdist, FUN = mean),
  y = stopdist, fill = tire))
p + geom_boxplot(width = 0.6) + geom_dotplot(binaxis = "y", stackdir = "center",
  binwidth = 2) + guides(fill = "none") + scale_fill_brewer() +
stat_summary(fun = mean, geom = "point", fill = "black", shape = 23, size = 3) +
labs(x = "Tire Brand", y = "Stopping distance in feet")
```

TIREWEAR

Tire Wear

Description

The data frame TIREWEAR contains measurements for the amount of tread loss in thousandths of an inch after 10,000 miles of driving.

Usage

TIREWEAR

Format

A data frame with 16 observations on the following 3 variables:

- wear (tread loss measured in thousandths of an inch)
- treat (a factor with levels A, B, C, and D)
- block (a factor with levels Car1, Car2, Car3, and Car4)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
par(mfrow=c(1, 2), cex = 0.8)
with(data = TIREWEAR,
  interaction.plot(treat, block, wear, type = "b", legend = FALSE))
with(data = TIREWEAR,
  interaction.plot(block, treat, wear, type = "b", legend = FALSE))
par(mfrow=c(1, 1), cex = 1)
```

TITANIC3

Titanic Survival Status

Description

The TITANIC3 data frame describes the survival status of individual passengers on the Titanic. The TITANIC3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

Usage

TITANIC3

Format

A data frame with 1309 observations on the following 14 variables:

- pclass (a factor with levels 1st, 2nd, and 3rd)
- survived (Survival where 0 = No; 1 = Yes)
- name (Name)
- sex (a factor with levels female and male)
- age (age in years)
- sibsp (Number of Siblings/Spouses Aboard)
- parch (Number of Parents/Children Aboard)
- ticket (Ticket Number)
- fare (Passenger Fare)
- cabin (Cabin)
- embarked (a factor with levels Cherbourg, Queenstown, and Southampton)
- boat (Lifeboat Number)
- body (Body Identification Number)
- home.dest (Home/Destination)

Details

Thomas Cason from the University of Virginia has greatly updated and improved the `titanic` data frame using the *Encyclopedia Titanica* and created a new dataset called TITANIC3. This dataset reflects the state of data available as of August 2, 1999. Some duplicate passengers have been dropped; many errors have been corrected; many missing ages have been filled in; and new variables have been created.

Source

<https://hbiostat.org/data/repo/titanic.html>

References

- Harrell, F. E. 2001. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.
- Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(TITANIC3, table(pclass, sex))
```

TOE

Nuclear Energy

Description

Nuclear energy (in TOE, tons of oil equivalent) produced in 12 randomly selected European countries during 2003

Usage

TOE

Format

A data frame with 12 observations on the following variable:

- energy (nuclear energy measured in tons of oil equivalent)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = TOE, aes(x = energy)) + geom_density(color = "red", alpha = 0.3,  
fill = "pink")
```

TOP20

Tennis Income

Description

TOP20 contains data (in millions of dollars) corresponding to the earnings of 15 randomly selected tennis players whose earnings fall somewhere in positions 20 through 100 of ranked earnings.

Usage

TOP20

Format

A data frame with 15 observations on the following variable:

- income (yearly income in millions of dollars)

Source

<https://www.atptour.com/>

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = TOP20, aes(x = income)) +  
  geom_histogram(binwidth = 1, fill = "lightblue", color = "blue") +  
  labs(x = "yearly income in millions of dollars")
```

tsum.test

Summarized t-test

Description

Performs a one-sample, two-sample, or a Welch modified two-sample t-test based on user supplied summary information. Output is identical to that produced with `t.test`.

Usage

```
tsum.test(
  mean.x,
  s.x = NULL,
  n.x = NULL,
  mean.y = NULL,
  s.y = NULL,
  n.y = NULL,
  alternative = c("two.sided", "less", "greater"),
  mu = 0,
  var.equal = FALSE,
  conf.level = 0.95,
  ...
)
```

Arguments

mean.x	a single number representing the sample mean of x
s.x	a single number representing the sample standard deviation of x
n.x	a single number representing the sample size of x
mean.y	a single number representing the sample mean of y
s.y	a single number representing the sample standard deviation of y
n.y	a single number representing the sample size of y
alternative	is a character string, one of "greater", "less", or "two.sided", or just the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, alternative refers to the true mean of the parent population in relation to the hypothesized value mu. For the standard two-sample tests, alternative refers to the difference between the true population mean for x and that for y, in relation to mu. For the one-sample and paired t-tests, alternative refers to the true mean of the parent population in relation to the hypothesized value mu. For the standard and Welch modified two-sample t-tests, alternative refers to the difference between the true population mean for x and that for y, in relation to mu. For the one-sample t-tests, alternative refers to the true mean of the parent population in relation to the hypothesized value mu. For the standard and Welch modified two-sample t-tests, alternative refers to the difference between the true population mean for x and that for y, in relation to mu.
mu	is a single number representing the value of the mean or difference in means specified by the null hypothesis.
var.equal	logical flag: if TRUE, the variances of the parent populations of x and y are assumed equal. Argument var.equal should be supplied only for the two-sample tests.
conf.level	is the confidence level for the returned confidence interval; it must lie between zero and one.
...	Other arguments passed onto tsum.test()

Details

If `y` is `NULL`, a one-sample t-test is carried out with `x`. If `y` is not `NULL`, either a standard or Welch modified two-sample t-test is performed, depending on whether `var.equal` is `TRUE` or `FALSE`.

Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the t-statistic, with names attribute <code>"t"</code>
<code>parameters</code>	is the degrees of freedom of the t-distribution associated with <code>statistic</code> . Component <code>parameters</code> has names attribute <code>"df"</code> .
<code>p.value</code>	the p-value for the test
<code>conf.int</code>	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute <code>conf.level</code> . When <code>alternative</code> is not <code>"two.sided"</code> , the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values <code>k</code> for which one would not reject the null hypothesis that the true mean or difference in means is <code>k</code> . Here infinity will be represented by <code>Inf</code> .
<code>estimate</code>	is a vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	is the value of the mean or difference in means specified by the null hypothesis. This equals the input argument <code>mu</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : <code>"greater"</code> , <code>"less"</code> or <code>"two.sided"</code> .
<code>data.name</code>	is a character string (vector of length 1) containing the names <code>x</code> and <code>y</code> for the two summarized samples.

Null Hypothesis

For the one-sample t-test, the null hypothesis is that the mean of the population from which `x` is drawn is `mu`. For the standard and Welch modified two-sample t-tests, the null hypothesis is that the population mean for `x` less that for `y` is `mu`.

The alternative hypothesis in each case indicates the direction of divergence of the population mean for `x` (or difference of means for `x` and `y`) from `mu` (i.e., `"greater"`, `"less"`, or `"two.sided"`).

Test Assumptions

The assumption of equal population variances is central to the standard two-sample t-test. This test can be misleading when population variances are not equal, as the null distribution of the test statistic is no longer a t-distribution. If the assumption of equal variances is doubtful with respect to a particular dataset, the Welch modification of the t-test should be used.

The t-test and the associated confidence interval are quite robust with respect to level toward heavy-tailed non-Gaussian distributions (e.g., data with outliers). However, the t-test is non-robust with respect to power, and the confidence interval is non-robust with respect to average length, toward these same types of distributions.

Confidence Intervals

For each of the above tests, an expression for the related confidence interval (returned component `conf.int`) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when `alternative` is not `"two.sided"`; infinity will be represented by `Inf`.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

- Kitchens, L.J. 2003. *Basic Statistics and Data Analysis*. Duxbury.
- Hogg, R. V. and Craig, A. T. 1970. *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.
- Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.
- Snedecor, G. W. and Cochran, W. G. 1980. *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

See Also

[z.test](#), [zsum.test](#)

Examples

```
# 95% Confidence Interval for mu1 - mu2, assuming equal variances
round(tsum.test(mean.x = 53/15, mean.y = 77/11, s.x=sqrt((222 - 15*(53/15)^2)/14),
s.y = sqrt((560 - 11*(77/11)^2)/10), n.x = 15, n.y = 11, var.equal = TRUE)$conf, 2)
# One Sample t-test
tsum.test(mean.x = 4, s.x = 2.89, n.x = 25, mu = 2.5)
```

tway.plots

Exploratory Graphs for Two Factor Designs

Description

Function creates side-by-side boxplots for each factor, a design plot (means), and an interaction plot.

Usage

```
tway.plots(Y, fac1, fac2, COL = c("#A9E2FF", "#0080FF"))
```

Arguments

Y	response variable
fac1	factor one
fac2	factor two
COL	a vector with two colors

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

See Also

[oneway.plots](#), [checking.plots](#)

Examples

```
with(data = TIREWEAR, twoway.plots(wear, treat, block))
#####
## Similar graphs with ggplot2 ##
#####
p1 <- ggplot(data = TIREWEAR, aes(x = treat, y = wear, fill = treat)) +
  geom_boxplot() + guides(fill = FALSE) + theme_bw()
p2 <- ggplot(data = TIREWEAR, aes(x = block, y = wear, fill = block)) +
  geom_boxplot() + guides(fill = FALSE) + theme_bw()
p3 <- ggplot(data = TIREWEAR, aes(x = treat, y = wear, color = block,
  group = block)) + stat_summary(fun.y = mean, geom = "point", size = 4) +
  stat_summary(fun.y = mean, geom = "line") + theme_bw()
p4 <- ggplot(data = TIREWEAR, aes(x = treat, y = wear, color = treat)) +
  geom_boxplot() + facet_grid(. ~ block) + theme_bw()
p1
p2
p3
p4
## To get all plots on the same device use gridExtra (not run)
## library(gridExtra)
## grid.arrange(p1, p2, p3, p4, nrow=2)
```

URLADDRESS

Megabytes Downloaded

Description

The manager of a URL commercial address is interested in predicting the number of megabytes downloaded, `megasd`, by clients according to the number minutes they are connected, `mconnected`. The manager randomly selects (megabyte, minute) pairs, and records the data. The pairs (`megasd`, `mconnected`) are stored in the data frame `URLADDRESS`.

Usage

URLADDRESS

Format

A data frame with 30 observations on the following 2 variables:

- megasd (megabytes downloaded)
- mconnected (number of minutes connected)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = URLADDRESS, aes(x = mconnected, y = megasd)) +  
  geom_point(color = "blue") +  
  labs(x = "number of minutes connected", y = "megabytes downloaded")
```

VIT2005

Apartments in Vitoria

Description

Descriptive information and the appraised total price (in Euros) for apartments in Vitoria, Spain

Usage

VIT2005

Format

A data frame with 218 observations on the following 5 variables:

- totalprice (the market total price (in Euros) of the apartment including garage(s) and storage room(s))
- area (the total living area of the apartment in square meters)
- zone (a factor indicating the neighborhood where the apartment is located with levels Z11, Z21, Z31, Z32, Z34, Z35, Z36, Z37, Z38, Z41, Z42, Z43, Z44, Z45, Z46, Z47, Z48, Z49, Z52, Z53, Z56, Z61, and Z62)
- category (a factor indicating the condition of the apartment with levels 2A, 2B, 3A, 3B, 4A, 4B, and 5A ordered so that 2A is the best and 5A is the worst)
- age (age of the apartment in years)
- floor (floor on which the apartment is located)

- rooms (total number of rooms including bedrooms, dining room, and kitchen)
- out (a factor indicating the percent of the apartment exposed to the elements: The levels E100, E75, E50, and E25, correspond to complete exposure, 75% exposure, 50% exposure, and 25% exposure, respectively.)
- conservation (is an ordered factor indicating the state of conservation of the apartment. The levels 1A, 2A, 2B, and 3A are ordered from best to worst conservation.)
- toilets (the number of bathrooms)
- garage (the number of garages)
- elevator (indicates the absence (0) or presence (1) of elevators.)
- streetcategory (an ordered factor from best to worst indicating the category of the street with levels S2, S3, S4, and S5)
- heating (a factor indicating the type of heating with levels 1A, 3A, 3B, and 4A which correspond to: no heating, low-standard private heating, high-standard private heating, and central heating, respectively.)
- storage (the number of storage rooms outside of the apartment)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = VIT2005, aes(x = area, y = totalprice, color = factor(elevator))) +
  geom_point()
modTotal <- lm(totalprice ~ area + as.factor(elevator) + area:as.factor(elevator),
  data = VIT2005)
modSimpl <- lm(totalprice ~ area, data = VIT2005)
anova(modSimpl, modTotal)
rm(modSimpl, modTotal)
```

WAIT

Waiting Time

Description

A statistician records how long he must wait for his bus each morning.

Usage

WAIT

Format

A data frame with 15 observations on the following variable:

- minutes (waiting time in minutes)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
with(data= WAIT, wilcox.test(minutes, mu = 6, alternative = "less"))
```

WASHER

Washer Diameter

Description

Diameter of circular metal disk

Usage

WASHER

Format

A data frame with 20 observations on the following variable:

- diameter (diameter of washer in cm)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WASHER, aes(x = diameter)) + geom_density(fill = "blue", alpha = 0.2)
```

WATER

Sodium Content of Water

Description

An independent agency measures the sodium content in 20 samples from source x and in 10 samples from source y and stores them in the data frame WATER.

Usage

WATER

Format

A data frame with 30 observations on the following 4 variables:

- x (sodium content measured in mg/L)
- y (sodium content measured in mg/L)
- sodium (combined sodium content measured in mg/L)
- source (a factor with levels x and y)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WATER, aes(x = sodium, y = ..density.., fill = source)) +  
  geom_density(alpha = 0.2)  
t.test(sodium ~ source, data = WATER, alternative = "less")
```

WCST

Wisconsin Card Sorting Test

Description

The following data are the test scores from a group of 50 patients from the *Virgen del Camino* Hospital (Pamplona, Spain) on the Wisconsin Card Sorting Test.

Usage

WCST

Format

A data frame with 50 observations on the following variable:

- score (score on the Wisconsin Card Sorting Test)

Details

The “Wisconsin Card Sorting Test” is widely used by psychiatrists, neurologists, and neuropsychologists with patients who have a brain injury, neurodegenerative disease, or a mental illness such as schizophrenia. Patients with any sort of frontal lobe lesion generally do poorly on the test.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WCST, aes(x = score)) + geom_density(fill = "lightblue", alpha = 0.8,
color = "blue")
```

WEIGHTGAIN

Weight Gain in Rats

Description

The data come from an experiment to study the gain in weight of rats fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal).

Usage

```
WEIGHTGAIN
```

Format

A data frame with 40 observations on the following 3 variables:

- `proteinsource` (a factor with levels Beef and Cereal)
- `proteinamount` (a factor with levels High and Low)
- `weightgain` (weight gained in grams)

Details

The design of the experiment is a completely randomized design with ten rats in each of the four treatments.

Source

Hand, D. J., F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. 1994. *A Handbook of Small Datasets*. Chapman and Hall/CRC, London.

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. London: Chapman & Hall.

Examples

```
ggplot(data = WEIGHTGAIN, aes(x = proteinamount, y = weightgain,
fill = proteinsource)) + geom_boxplot()
aov(weightgain ~ proteinsource*proteinamount, data = WEIGHTGAIN)
```

WHEATSPAIN

Wheat Surface Area in Spain

Description

Seventeen Spanish communities and their corresponding surface area (in hecatares) dedicated to growing wheat

Usage

WHEATSPAIN

Format

A data frame with 17 observations on the following 3 variables:

- community (a factor with levels Andalucia, Aragon, Asturias, Baleares, C.Valenciana, Canarias, Cantabria, Castilla-La Mancha, Castilla-Leon, Cataluna, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarre, and P.Vasco)
- hectares (surface area measured in hectares)
- acres (surface area measured in acres)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WHEATSPAIN, aes(x = reorder(community, acres), y = acres)) +  
  geom_bar(stat="identity", color = "orange", fill = "gold") + coord_flip() +  
  labs(x = "")
```

WHEATUSA2004

USA Wheat Surface 2004

Description

USA's 2004 harvested wheat surface by state

Usage

WHEATUSA2004

Format

A data frame with 30 observations on the following 2 variables:

- `states` (a factor with levels AR, CA, CO, DE, GA, ID, IL, IN, KS, KY, MD, MI, MO, MS, MT, NC, NE, NY, OH, OK, OR, Other, PA, SC, SD, TN, TX, VA, WA, and WI)
- `acres` (wheat surface area measured in thousands of acres)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WHEATUSA2004, aes(x = reorder(states, acres), y = acres)) +
  geom_bar(stat = "identity", color = "gold", fill = "yellow") + coord_flip() +
  labs(x = "")
```

wilcox.test

Wilcoxon Exact Test

Description

Performs exact one sample and two sample Wilcoxon tests on vectors of data

Usage

```
wilcox.test(
  x,
  y = NULL,
  mu = 0,
  paired = FALSE,
  alternative = c("two.sided", "less", "greater"),
  conf.level = 0.95
)
```

Arguments

<code>x</code>	is a numeric vector of data values. Non-finite (i.e. infinite or missing) values will be omitted.
<code>y</code>	an optional numeric vector of data values
<code>mu</code>	a number specifying an optional parameter used to form the null hypothesis
<code>paired</code>	a logical indicating whether you want a paired test
<code>alternative</code>	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "less", or "greater". You can specify just the initial letter.
<code>conf.level</code>	confidence level of the interval

Details

If only x is given, or if both x and y are given and `paired = TRUE`, a Wilcoxon signed rank test of the null hypothesis that the distribution of x (in the one sample case) or of $x - y$ (in the paired two sample case) is symmetric about μ is performed.

Otherwise, if both x and y are given and `paired = FALSE`, a Wilcoxon rank sum test is done. In this case, the null hypothesis is that the distribution of x and y differ by a location shift μ , and the alternative is that they differ by some other location shift (and the one-sided alternative "greater" is that x is shifted to the right of y).

Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the value of the test statistic with a name describing it
<code>p.value</code>	the p-value for the test
<code>null.value</code>	the location parameter μ
<code>alternative</code>	a character string describing the alternative hypothesis
<code>method</code>	the type of test applied
<code>data.name</code>	a character string giving the names of the data
<code>conf.int</code>	a confidence interval for the location parameter
<code>estimate</code>	an estimate of the location parameter

Note

The function is rather primitive and should only be used for problems with fewer than 19 observations as the memory requirements are rather large.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

- Gibbons, J.D. and Chakraborti, S. 1992. *Nonparametric Statistical Inference*. Marcel Dekker Inc., New York.
- Hollander, M. and Wolfe, D.A. 1999. *Nonparametric Statistical Methods*. New York: John Wiley & Sons.

See Also

[wilcox.test](#)

Examples

```
# Wilcoxon Signed Rank Test
PH <- c(7.2, 7.3, 7.3, 7.4)
wilcox.test(PH, mu = 7.25, alternative = "greater")
# Wilcoxon Signed Rank Test (Dependent Samples)
with(data = AGGRESSION,
wilcox.test(violence, noviolence, paired = TRUE, alternative = "greater"))
# Wilcoxon Rank Sum Test
x <- c(7.2, 7.2, 7.3, 7.3)
y <- c(7.3, 7.3, 7.4, 7.4)
wilcox.test(x, y)
rm(PH, x, y)
```

WOOL

Wool Production

Description

Random sample of wool production in thousands of kilograms on 5 different days at two different locations

Usage

WOOL

Format

A data frame with 30 observations on the following 2 variables:

- production (wool production in thousands of kilograms)
- location (a factor with levels textileA and textileB.)

References

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. 2015. *Probability and Statistics with R*, Second Edition. Chapman & Hall / CRC.

Examples

```
ggplot(data = WOOL, aes(location, production, fill = location)) + geom_boxplot() +
guides(fill = "none") + scale_fill_brewer()
t.test(production ~ location, data = WOOL)
```

z.test

z-Test

Description

This function is based on the standard normal distribution and creates confidence intervals and tests hypotheses for both one and two sample problems.

Usage

```
z.test(
  x,
  sigma.x = NULL,
  y = NULL,
  sigma.y = NULL,
  sigma.d = NULL,
  alternative = c("two.sided", "less", "greater"),
  mu = 0,
  paired = FALSE,
  conf.level = 0.95,
  ...
)
```

Arguments

x	a (non-empty) numeric vector of data values
sigma.x	a single number representing the population standard deviation for x
y	an optional (non-empty) numeric vector of data values
sigma.y	a single number representing the population standard deviation for y
sigma.d	a single number representing the population standard deviation for the paired differences
alternative	character string, one of "greater", "less", or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, alternative refers to the true mean of the parent population in relation to the hypothesized value mu. For the standard two-sample tests, alternative refers to the difference between the true population mean for x and that for y, in relation to mu.
mu	a single number representing the value of the mean or difference in means specified by the null hypothesis
paired	a logical indicating whether you want a paired z-test
conf.level	confidence level for the returned confidence interval, restricted to lie between zero and one
...	Other arguments passed onto z.test()

Details

If `y` is `NULL`, a one-sample z-test is carried out with `x` provided `sigma.x` is not `NULL`. If `y` is not `NULL`, a standard two-sample z-test is performed provided both `sigma.x` and `sigma.y` are finite. If `paired = TRUE`, a paired z-test where the differences are defined as $x - y$ is performed when the user enters a finite value for `sigma.d` (the population standard deviation for the differences).

Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the z-statistic, with names attribute <code>z</code>
<code>p.value</code>	the p-value for the test
<code>conf.int</code>	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute <code>conf.level</code> . When <code>alternative</code> is not <code>"two.sided"</code> , the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values k for which one would not reject the null hypothesis that the true mean or difference in means is k . Here, infinity will be represented by <code>Inf</code> .
<code>estimate</code>	vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	the value of the mean or difference of means specified by the null hypothesis. This equals the input argument <code>mu</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : <code>"greater"</code> , <code>"less"</code> , or <code>"two.sided"</code> .
<code>data.name</code>	a character string (vector of length 1) containing the actual names of the input vectors <code>x</code> and <code>y</code>

Null Hypothesis

For the one-sample z-test, the null hypothesis is that the mean of the population from which `x` is drawn is μ . For the standard two-sample z-test, the null hypothesis is that the population mean for `x` less that for `y` is μ . For the paired z-test, the null hypothesis is that the mean difference between `x` and `y` is μ .

The alternative hypothesis in each case indicates the direction of divergence of the population mean for `x` (or difference of means for `x` and `y`) from μ (i.e., `"greater"`, `"less"`, or `"two.sided"`).

Test Assumptions

The assumption of normality for the underlying distribution or a sufficiently large sample size is required along with the population standard deviation to use Z procedures.

Confidence Intervals

For each of the above tests, an expression for the related confidence interval (returned component `conf.int`) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when `alternative` is not "two.sided"; infinity will be represented by `Inf`.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

- Kitchens, L.J. 2003. *Basic Statistics and Data Analysis*. Duxbury.
- Hogg, R. V. and Craig, A. T. 1970. *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.
- Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.
- Snedecor, G. W. and Cochran, W. G. 1980. *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

See Also

[zsum.test](#), [tsum.test](#)

Examples

```
with(data = GROCERY, z.test(x = amount, sigma.x = 30, conf.level = 0.97)$conf)
# Example 8.3 from PASWR.
x <- rnorm(12)
z.test(x, sigma.x = 1)
# Two-sided one-sample z-test where the assumed value for
# sigma.x is one. The null hypothesis is that the population
# mean for 'x' is zero. The alternative hypothesis states
# that it is either greater or less than zero. A confidence
# interval for the population mean will be computed.
x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7., 6.4, 7.1, 6.7, 7.6, 6.8)
y <- c(4.5, 5.4, 6.1, 6.1, 5.4, 5., 4.1, 5.5)
z.test(x, sigma.x=0.5, y, sigma.y=0.5, mu=2)
# Two-sided standard two-sample z-test where both sigma.x
# and sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is 2.
# The alternative hypothesis is that this difference is not 2.
# A confidence interval for the true difference will be computed.
z.test(x, sigma.x = 0.5, y, sigma.y = 0.5, conf.level = 0.90)
# Two-sided standard two-sample z-test where both sigma.x and
# sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is zero.
# The alternative hypothesis is that this difference is not
# zero. A 90% confidence interval for the true difference will
# be computed.
```

```
rm(x, y)
```

```
zsum.test
```

```
Summarized z-test
```

Description

This function is based on the standard normal distribution and creates confidence intervals and tests hypotheses for both one and two sample problems based on summarized information the user passes to the function. Output is identical to that produced with `z.test`.

Usage

```
zsum.test(
  mean.x,
  sigma.x = NULL,
  n.x = NULL,
  mean.y = NULL,
  sigma.y = NULL,
  n.y = NULL,
  alternative = c("two.sided", "less", "greater"),
  mu = 0,
  conf.level = 0.95,
  ...
)
```

Arguments

<code>mean.x</code>	a single number representing the sample mean of x
<code>sigma.x</code>	a single number representing the population standard deviation for x
<code>n.x</code>	a single number representing the sample size for y
<code>mean.y</code>	a single number representing the sample mean of y
<code>sigma.y</code>	a single number representing the population standard deviation for y
<code>n.y</code>	a single number representing the sample size for y
<code>alternative</code>	is a character string, one of "greater", "less", or "two.sided", or the initial letter of each, indicating the specification of the alternative hypothesis. For one-sample tests, <code>alternative</code> refers to the true mean of the parent population in relation to the hypothesized value <code>mu</code> . For the standard two-sample tests, <code>alternative</code> refers to the difference between the true population mean for x and that for y, in relation to <code>mu</code> .
<code>mu</code>	a single number representing the value of the mean or difference in means specified by the null hypothesis
<code>conf.level</code>	confidence level for the returned confidence interval, restricted to lie between zero and one
<code>...</code>	Other arguments passed onto <code>z.test()</code>

Details

If `y` is `NULL`, a one-sample z-test is carried out with `x` provided `sigma.x` is finite. If `y` is not `NULL`, a standard two-sample z-test is performed provided both `sigma.x` and `sigma.y` are finite.

Value

A list of class `htest`, containing the following components:

<code>statistic</code>	the z-statistic, with names attribute <code>z</code>
<code>p.value</code>	the p-value for the test
<code>conf.int</code>	is a confidence interval (vector of length 2) for the true mean or difference in means. The confidence level is recorded in the attribute <code>conf.level</code> . When <code>alternative</code> is not <code>"two.sided"</code> , the confidence interval will be half-infinite, to reflect the interpretation of a confidence interval as the set of all values <code>k</code> for which one would not reject the null hypothesis that the true mean or difference in means is <code>k</code> . Here, infinity will be represented by <code>Inf</code> .
<code>estimate</code>	vector of length 1 or 2, giving the sample mean(s) or mean of differences; these estimate the corresponding population parameters. Component <code>estimate</code> has a names attribute describing its elements.
<code>null.value</code>	the value of the mean or difference in means specified by the null hypothesis. This equals the input argument <code>mu</code> . Component <code>null.value</code> has a names attribute describing its elements.
<code>alternative</code>	records the value of the input argument <code>alternative</code> : <code>"greater"</code> , <code>"less"</code> , or <code>"two.sided"</code> .
<code>data.name</code>	a character string (vector of length 1) containing the names <code>x</code> and <code>y</code> for the two summarized samples.

Null Hypothesis

For the one-sample z-test, the null hypothesis is that the mean of the population from which `x` is drawn is `mu`. For the standard two-sample z-test, the null hypothesis is that the population mean for `x` less that for `y` is `mu`.

The alternative hypothesis in each case indicates the direction of divergence of the population mean for `x` (or difference of means for `x` and `y`) from `mu` (i.e., `"greater"`, `"less"`, or `"two.sided"`).

Test Assumptions

The assumption of normality for the underlying distribution or a sufficiently large sample size is required along with the population standard deviation to use Z procedures.

Confidence Intervals

For each of the above tests, an expression for the related confidence interval (returned component `conf.int`) can be obtained in the usual way by inverting the expression for the test statistic. Note that, as explained under the description of `conf.int`, the confidence interval will be half-infinite when `alternative` is not `"two.sided"`; infinity will be represented by `Inf`.

Author(s)

Alan T. Arnholt <arnholtat@appstate.edu>

References

- Kitchens, L.J. 2003. *Basic Statistics and Data Analysis*. Duxbury.
- Hogg, R. V. and Craig, A. T. 1970. *Introduction to Mathematical Statistics, 3rd ed.* Toronto, Canada: Macmillan.
- Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill.
- Snedecor, G. W. and Cochran, W. G. 1980. *Statistical Methods, 7th ed.* Ames, Iowa: Iowa State University Press.

See Also

[z.test](#), [tsum.test](#)

Examples

```
zsum.test(mean.x = 56/30, sigma.x = 2, n.x = 30, alternative="greater", mu = 1.8)
# Example 9.7 part a. from PASWR.
x <- rnorm(12)
zsum.test(mean(x), sigma.x = 1, n.x = 12)
# Two-sided one-sample z-test where the assumed value for
# sigma.x is one. The null hypothesis is that the population
# mean for 'x' is zero. The alternative hypothesis states
# that it is either greater or less than zero. A confidence
# interval for the population mean will be computed.
# Note: returns same answer as:
z.test(x, sigma.x = 1)

x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)
y <- c(4.5, 5.4, 6.1, 6.1, 5.4, 5.0, 4.1, 5.5)
zsum.test(mean(x), sigma.x = 0.5, n.x = 11, mean(y), sigma.y = 0.5, n.y = 8, mu = 2)
# Two-sided standard two-sample z-test where both sigma.x
# and sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is 2.
# The alternative hypothesis is that this difference is not 2.
# A confidence interval for the true difference will be computed.
# Note: returns same answer as:
z.test(x, sigma.x = 0.5, y, sigma.y = 0.5)
#
zsum.test(mean(x), sigma.x = 0.5, n.x = 11, mean(y), sigma.y = 0.5, n.y = 8,
conf.level=0.90)
# Two-sided standard two-sample z-test where both sigma.x and
# sigma.y are both assumed to equal 0.5. The null hypothesis
# is that the population mean for 'x' less that for 'y' is zero.
# The alternative hypothesis is that this difference is not
# zero. A 90% confidence interval for the true difference will
# be computed. Note: returns same answer as:
```

```
z.test(x, sigma.x=0.5, y, sigma.y=0.5, conf.level=0.90)  
rm(x, y)
```

Index

* datasets

AGGRESSION, 4
APPLE, 5
APTSIZE, 6
BABERUTH, 7
BAC, 8
BATTERY, 9
BIOMASS, 10
BODYFAT, 11
CALCULUS, 12
CARS2004, 13
CHIPS, 15
CIRCUIT, 16
COSAMA, 18
COWS, 19
DEPEND, 20
DROSOPHILA, 20
ENGINEER, 22
EPIDURAL, 23
EPIDURALF, 24
EURD, 25
FAGUS, 25
FCD, 26
FERTILIZE, 27
FOOD, 28
FORMULA1, 28
GD, 29
GLUCOSE, 30
GRADES, 30
GROCERY, 31
HARDWATER, 32
HOUSE, 33
HSWRESTLER, 34
HUBBLE, 35
INSURQUOTES, 36
JANKA, 37
KINDER, 38
LEDDIODE, 40
LOSTR, 41
MILKCARTON, 41
NC2010DMG, 43
PAMTEMP, 48
PHENYL, 49
PHONE, 50
RAT, 50
RATBP, 51
REFRIGERATOR, 52
ROACHEGGS, 52
SALINITY, 53
SATFRUIT, 54
SBIQ, 55
SCHIZO, 56
SCORE, 56
SDS4, 57
SIMDATAST, 60
SIMDATAXT, 61
SOCCER, 61
STATTEMPS, 63
STSCHOOL, 64
SUNDIG, 65
SUNFLOWER, 65
SURFACESPAIN, 66
SWIMTIMES, 67
TENNIS, 68
TESTSCORES, 69
TIRE, 69
TIREWEAR, 70
TITANIC3, 71
TOE, 72
TOP20, 73
URLADDRESS, 77
VIT2005, 78
WAIT, 79
WASHER, 80
WATER, 80
WCST, 81
WEIGHTGAIN, 82
WHEATSPAIN, 83

- WHEATUSA2004, 83
- WOOL, 86
- * **hplot**
 - checking.plots, 14
 - eda, 21
 - ksdist, 39
 - ksldist, 39
 - normarea, 44
 - ntester, 46
 - oneway.plots, 47
 - twoway.plots, 76
- * **htest**
 - tsum.test, 73
 - wilcoxe.test, 84
 - z.test, 87
 - zsum.test, 90
- * **package**
 - PASWR2-package, 4
- * **programming**
 - bino.gen, 10
 - cisim, 16
 - interval.plot, 36
 - multiplot, 42
 - nsizе, 45
 - srs, 62
- AGGRESSION, 4
- APPLE, 5
- APTSIZE, 6
- BABERUTH, 7
- BAC, 8
- BATTERY, 9
- bino.gen, 10
- BIOMASS, 10
- BODYFAT, 11
- CALCULUS, 12
- CARS2004, 13
- checking.plots, 14, 47, 77
- CHIPS, 15
- CIRCUIT, 16
- cisim, 16
- combn, 63
- COSAMA, 18
- COWS, 19
- DEPEND, 20
- DROSOPHILA, 20
- eda, 21
- ENGINEER, 22
- EPIDURAL, 23
- EPIDURALF, 24
- EURD, 25
- FAGUS, 25
- FCD, 26
- FERTILIZE, 27
- FOOD, 28
- FORMULA1, 28
- GD, 29
- GLUCOSE, 30
- GRADES, 30
- GROCERY, 31
- HARDWATER, 32
- HOUSE, 33
- HSWRESTLER, 34
- HUBBLE, 35
- INSURQUOTES, 36
- interval.plot, 36
- JANKA, 37
- KINDER, 38
- ksdist, 39, 40
- ksldist, 39, 39
- layout, 42
- LEDDIODE, 40
- LOSTR, 41
- MILKCARTON, 41
- multiplot, 42
- NC2010DMG, 43
- normarea, 44
- nsizе, 45
- ntester, 46
- oneway.plots, 14, 47, 77
- PAMTEMP, 48
- PASWR2-package, 4
- PHENYL, 49
- PHONE, 50
- RAT, 50

RATBP, 51
REFRIGERATOR, 52
ROACHEGGS, 52

SALINITY, 53
SATFRUIT, 54
SBIQ, 55
SCHIZO, 56
SCORE, 56
SDS4, 57
SIGN. test, 58
SIMDATAST, 60
SIMDATAXT, 61
SOCCER, 61
srs, 62
STATTEMPS, 63
STSCHOOL, 64
SUNDIG, 65
SUNFLOWER, 65
SURFACESPAIN, 66
SWIMTIMES, 67

TENNIS, 68
TESTSCORES, 69
TIRE, 69
TIREWEAR, 70
TITANIC3, 71
TOE, 72
TOP20, 73
tsum. test, 60, 73, 89, 92
twoway.plots, 14, 47, 76

URLADDRESS, 77

VIT2005, 78

WAIT, 79
WASHER, 80
WATER, 80
WCST, 81
WEIGHTGAIN, 82
WHEATSPAIN, 83
WHEATUSA2004, 83
wilcox. test, 85
wilcoxe. test, 84
WOOL, 86

z. test, 60, 76, 87, 92
zsum. test, 60, 76, 89, 90