Package 'sur'

July 23, 2025

Type Package

Title Companion to ``Statistics Using R: An Integrative Approach''

Version 1.0.4

Depends R (>= 3.5.0)

Description Access to the datasets and many of the functions used in ``Statistics Using R: An Integrative Approach''. These datasets include a subset of the National Education Longitudinal Study, the Framingham Heart Study, as well as several simulated datasets used in the examples throughout the textbook. The functions included in the package reproduce some of the functionality of 'Stata' that is not directly available in 'R'. The package also contains a tutorial on basic data frame management, including how to handle missing data.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.0.0

Imports learnr

NeedsCompilation no

Author Daphna Harel [cre, aut],
Sharon Weinberg [ctb],
Sarah Abramowitz [ctb]

Maintainer Daphna Harel <daphna.harel@gmail.com>

Repository CRAN

Date/Publication 2020-08-25 22:30:02 UTC

Contents

Anscombe *Anscombe's Four Datasets*

Description

This dataset is used to illustrate the importance of statistical display as an adjunct to summary statistics. Anscombe (1973) fabricated four different bivariate datasets such that, for all datasets, the respective X and Y means, X and Y standard deviations, and correlations, slopes, intercepts, and standard errors of estimate are equal. Accordingly, without a visual representation of these four panels, one might assume that the data values for all four datasets are the same. Scatterplots illustrate, however, the extent to which these datasets are different from one another.

Usage

```
Anscombe
```

Format

A data frame with 11 rows and 8 variables:

x1 values of X for the first dataset

y1 values of Y for the first dataset

x2 values of X for the second dataset

y2 values of Y for the second dataset

x3 values of X for the third dataset

y3 values of Y for the third dataset

x4 values of X for the fourth dataset

y4 values of Y for the fourth dataset

Basketball *Heights and Weights of U.S. Basketball Players*

Description

The dataset consists of the heights and weights of the 24 scoring leaders, 12 each from the U.S. Women's and Men's National Basketball Association, for the 2014 – 2015 season. These data are taken from the ESPN website at espn.com.

Usage

```
Basketball
```

Format

A data frame with 20 rows and 6 variables:

player name of player

gender gender of player

heightin height of player in inches

weightlb weight of player in pounds

games number of games played

points average total points scored per game

Source

https://www.espn.com/

`Blood` *Blood Pressure Data of African-American Adult Males*

Description

The data were collected to determine whether an increase in calcium intake reduces blood pressure among African-American adult males. The data are based on a sample of 21 African-American adult males selected randomly from the population of African-American adult males. Ten of the 21 men were randomly assigned to a treatment condition that required them to take a calcium supplement for 12 weeks. The remaining 11 men received a placebo for the 12 weeks. At both the beginning and the end of this time period, systolic blood pressure readings of all men were recorded. These data are adapted from the Data and Story Library (DASL) website.

Usage

```
Blood
```

Format

A data frame with 21 rows and 4 variables:

id case number

treatmen treatment condition

systolc1 initial blood pressure

systolc2 final blood pressure

boot.mean *Bootstrapped Mean*

Description

Function to obtain a sampling distribution of means by bootstrapping.

Usage

```
boot.mean(x, B, n = length(x))
```

Arguments

x	original sample, given as a numeric or logical object, to be used to generate bootstrapped samples.
B	number of bootstrapped samples to be generated by randomly sampling with replacement.
n	size of each bootstrapped sample. Default setting is the size of the original sample.

Value

A list with components:

Replications	number of bootstrapped means computed.
mean	mean of bootstrapped means.
se	standard error, estimated as the standard deviation of bootstrapped means.
bootstrap.samples	means of bootstrapped samples.

Examples

```
# using simple vector
a = 1:10
set.seed(1234)
boot.mean(a, B = 500)

# using variable from data frame
set.seed(1234)
boot.mean(Framingham$AGE3, B = 1000)
```

Brainsz *Brain Size and IQ Data*

Description

The data are based on a study by Willerman et al. (1991) of the relationships between brain size, gender, and intelligence. The research participants consisted of 40 right-handed introductory psychology students with no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease who were selected from a larger pool of introductory psychology students with total Scholastic Aptitude Test Scores higher than 1350 or lower than 940. The students in the study took four subtests (Vocabulary, Similarities, Block Design, and Picture Completion) of the Wechsler (1981) Adult Intelligence Scale-Revised. Among the students with Wechsler full-scale IQ's less than 103, 10 males and 10 females were randomly selected. Similarly, among the students with Wechsler full-scale IQ's greater than 130, 10 males and 10 females were randomly selected, yielding a randomized blocks design. MRI scans were performed at the same facility for all 40 research participants to measure brain size. The scans consisted of 18 horizontal MRI images. The computer counted all pixels with non-zero gray scale in each of the 18 images, and the total count served as an index for brain size. The dataset and description are adapted from the Data and Story Library (DASL) website.

Usage

```
Brainsz
```

Format

A data frame with 40 rows and 7 variables:

ID case number

GENDER gender of student

FSIQ full-scale IQ score based on WAIS-R

VIQ verbal IQ score based on WAIS-R

PIQ performance IQ score based on WAIS-R

MRI pixel count from 18 MRI scans

IQDI group membership based on FSIQ score

Chapter14_Figures *Exercise 14.1 Figures*

Description

This dataset contains simulated data for the figures accompanying Exercise 14.1 of Chapter 14. The data represent the results of a fictional study to determine whether there is a relationship between gender, teaching method, and achievement in reading. Each set of scores reflects a scenario with a different relationship among the variables.

Usage

```
Chapter14_Figures
```

Format

A data frame with 12 rows and 7 variables:

sex individual's sex

score1 reading achievement score for first scenario

method teaching method

score2 reading achievement score for second scenario

score3 reading achievement score for third scenario

score4 reading achievement score for fourth scenario

score5 reading achievement score for fifth scenario

cumulative.table *Cumulative Percentage Table*

Description

Returns as a named vector the cumulative percentage frequency distribution of a variable `x` at each unique value.

Usage

```
cumulative.table(x)
```

Arguments

`x`	object containing data for a single variable.

Details

If x contains NA values (missing data), the cumulative percentage table will not reach 100. The table will end with the cumulative percentage of non-missing data within the object; the value remaining after subtracting this value from 100 represents the percentage of NA values within the object.

Value

A named numeric vector containing cumulative percentage frequencies, named by unique values of x and ordered numerically or alphabetically by name.

See Also

percent.table, cumsum, table

Examples

```
# using variable without NA values
cumulative.table(NELS$famsize)

# using variable with NA values
cumulative.table(NELS$parmarl8)
```

Currency *Value and Circulation of Currency*

Description

This dataset contains, for the smaller bill denominations, the value of the bill and the total value in circulation. The source for these data is *The World Almanac and Book of Facts 2014*.

Usage

```
Currency
```

Format

A data frame with 5 rows and 3 variables:

BillValue denomination

TotalCirculation total currency in circulation in U.S. dollars

NumberCirculation total number of bills in circulation

Exercise *Exercise, Food Intake, and Weight Loss*

Description

A fabricated dataset constructed by Darlington (1990) to demonstrate the importance of including all relevant variables in an analysis. This dataset contains information about exercise, food intake, and weight loss for a fictional set of dieters.

Usage

```
Exercise
```

Format

A data frame with 10 rows and 4 variables:

ID case number

Exercise average daily number of hours exercised in that week

FoodIntake average daily number of calories consumed in one particular week that is more than a baseline of 1,000 calories, as measured in increments of 100 calories

WeightLoss number of pounds lost in that week

References

"Regression and linear models." Darlington, R. B. (1990, ISBN:978-0070153721)

Exercise14_5 *Exercise 14.5 Data*

Description

This dataset contains simulated data for the figures accompanying Exercise 14.1 of Chapter 14. The data represent the results of a fictional study in which a college professor examines the effect of the grade level of the students and the time of the course on how well undergraduate students at her college do in her course.

Usage

```
Exercise14_5
```

Format

A data frame with 40 rows and 3 variables:

Time time of day student takes the course

Year year of college in which the student is enrolled

Score final exam score

Figure15_1 *Figure 15.1 Data*

Description

This dataset contains simulated data for Figure 15.1 of Chapter 15.

Usage

```
Figure15_1
```

Format

A list with 3 elements:

x an integer-scaled independent variable

y an integer-scaled outcome variable

f frequency of value pair

Figure15_12 *Figure 15.12 Data*

Description

This dataset contains simulated data for Figures 15.12 - 15.13 of Chapter 15.

Usage

```
Figure15_12
```

Format

A data frame with 9 rows and 4 variables:

x a numeric independent variable for Figure 15.12

y a numeric outcome variable for Figure 15.12

xpr a numeric independent variable for Figure 15.13

ypr a numeric outcome variable for Figure 15.13

Figure15_9 *Figure 15.9 Data*

Description

This dataset contains simulated data for Figures 15.9 - 15.11 of Chapter 15.

Usage

```
Figure15_9
```

Format

A data frame with 24 rows and 4 variables:

x a numeric independent variable for Figure 15.9

y a numeric outcome variable for Figure 15.9

res residual value for regression of `y` on `x`

log_y log of the outcome variable `y`

Figure2_4 *Figure 2.4. Annual Number of Deaths in New York City: Tobacco vs. Other*

Description

This dataset contains data on causes of death in New York City that were used for Figure 2.4 of Chapter 2.

Usage

```
Figure2_4
```

Format

A data frame with 591,200 rows and 1 variable:

causes cause of death

Figure3_2 *Figure 3.2 Data*

Description

This dataset contains simulated test scores of Spanish fluency used to generate Figure 3.2 of Chapter 3.

Usage

```
Figure3_2
```

Format

A data frame with 100 rows and 1 variable:

fluency score on test of Spanish fluency

Figure3_3 *Figure 3.3 Data*

Description

This dataset contains simulated scores used to generate Figure 3.3 of Chapter 3.

Usage

```
Figure3_3
```

Format

A data frame with 45 rows and 1 variable:

score numeric score from rectangular distribution

Figure3_5a *Figure 3.5(A) Data*

Description

This dataset contains simulated scores used to generate Figure 3.5(A) of Chapter 3.

Usage

```
Figure3_5a
```

Format

A data frame with 121 rows and 1 variable:

DistnA numeric score from a symmetric distribution

Figure3_5b *Figure 3.5(B) Data*

Description

This dataset contains simulated scores used to generate Figure 3.5(B) of Chapter 3.

Usage

```
Figure3_5b
```

Format

A data frame with 75 rows and 1 variable:

DistnB numeric score from a symmetric distribution

Figure3_6and7 *Figures 3.6 and 3.7 Data*

Description

This dataset contains simulated scores used to generate Figures 3.6 ad 3.7 of Chapter 3.

Usage

```
Figure3_6and7
```

Format

A data frame with 69 rows and 2 variables:

NegSkew numeric score from a distribution with severe negative skew

PosSkew numeric score from a distribution with severe positive skew

Figure5_5 *Figure 5.5 Data*

Description

This dataset contains simulated scores used to generate Figures 5.5(A) - 5.5(I) of Chapter 5.

Usage

```
Figure5_5
```

Format

A data frame with 10 rows and 18 variables:

ax days elapsed in a given year

ay days remaining in that same year

bx age of elementary school student

by number of seconds to run a 100-yard dash

cx introversion score of adolescent boy

cy aggression score of adolescent boy

dx moodiness score of college freshman

dy English ability score of college freshman

ex weight of male college student

ey achievement score in statistics of male college student

fx expected grade in course of college student

fy course evaluation score given by college student

gx IQ score of child in grades K – 3

gy reading achievement score of child in grades K – 3

hx arithmetic reasoning score of elementary school student

hy arithmetic fundamentals score of elementary school student

ix diameter of tree

iy circumference of tree

`Framingham` *Framingham Heart Study*

Description

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of non-institutionalized people in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1956 and 5,209 subjects were initially enrolled in the study. In our dataset, we included variables from the first examination in 1956 and the third examination, in 1968. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of any of the following types of coronary heart disease(CHD): angina pectoris, myocardial infarction, heart failure, and cerebrovascular disease.

Usage

```
Framingham
```

Format

A data frame with 400 rows and 33 variables:

ID case number

SEX sex

TOTCHOL1 serum cholesterol (mg/dL) at initial examination

AGE1 age (years) at initial examination

SYSBP1 systolic blood pressure (mmHg) at initial examination

DIABP1 diastolic blood pressure (mmHg) at initial examination

CURSMOKE1 indicator that participant currently is a cigarette smoker at initial examination

CIGPDAY1 cigarettes smoked per day at initial examination

BMI1 Body Mass Index (kg/(M*M)) at initial examination

DIABETES1 indicator that participant is diabetic at initial examination

BPMEDS1 use of anti-hypertensive medication at initial examination

HEARTRTE1 ventricular rate (beats/min) at initial examination

GLUCOSE1 casual glucose (mg/dL) at initial examination

PREVCHD1 prevalent CHD (angina pectoris, myocardial infarction, or coronary insufficiency) at initial examination

TIME1 days since initial examination

TIMECHD1 days from initial examination to any CHD event

TOTCHOL3 serum cholesterol (mg/dL) at third examination

AGE3 age (years) at third examination

SYSBP3 systolic blood pressure (mmHg) at third examination

DIABP3 diastolic blood pressure (mmHg) at third examination

CURSMOKE3 indicator that participant currently is a cigarette smoker at third examination

CIGPDAY3 cigarettes smoked per day at third examination

BMI3 Body Mass Index (kg/(M*M) at third examination

DIABETES3 indicator that participant is diabetic at third examination

BPMEDS3 use of anti-hypertensive medication at third examination

HEARTRTE3 ventricular rate (beats/min) at third examination

GLUCOSE3 casual glucose (mg/dL) at third examination

PREVCHD3 prevalent CHD (angina pectoris, myocardial infarction, or coronary insufficiency) at third examination

TIME3 days since initial examination at third examination

HDLC3 HDL cholesterol (mg/dL) at third examination

LDLC3 LDL cholesterol (mg/dL) at third examination

TIMECHD3 days from initial examination to any CHD event at third examination

ANYCHD4 indicator of event of hospitalized myocardial infarction, angina pectoris, coronary insufficiency, or fatal CHD by the end of the study

Details

The associated dataset is a subset of the data collected as part of the Framingham study and includes laboratory, clinic, questionnaire, and adjudicated event data on 400 participants. These participants for the dataset have been chosen so that among all male participants, 100 smokers and 100 non-smokers were selected at random. A similar procedure resulted in 100 female smokers and 100 female non-smokers. This procedure resulted in an over-sampling of smokers. The data for each participant is on one row. People who had any type of CHD in the initial examination period are not included in the dataset.

Hamburger *McDonald's Hamburger Nutrition Information*

Description

This dataset contains the fat grams and calories associated with the different types of hamburger sold by McDonald's. The data are from McDonald's Nutrition Information Center.

Usage

```
Hamburger
```

Format

A data frame with 5 rows and 4 variables:

name type of burger

fat grams of fat

calories total calories

cheese cheese added

IceCream *Ice Cream Sales Data*

Description

This dataset contains fabricated data for the temperature, relative humidity, and ice cream sales for 30 days randomly selected between May 15th and September 6th.

Usage

```
IceCream
```

Format

A data frame with 30 rows and 4 variables:

id case number

temp temperature in degrees Fahrenheit

barsold number of ice cream bars sold

relhumid relative humidity

Impeach *Clinton Impeachment Votes*

Description

On February 12, 1999, for only the second time in the nation's history, the U.S. Senate voted on whether to remove a president, based on impeachment articles passed by the U.S. House. Professor Alan Reifman of Texas Tech University created the dataset consisting of descriptions of each senator that can be used to understand some of the reasons that the senators voted the way they did. The data are taken from the Journal of Statistics Education [online].

Usage

```
Impeach
```

Format

A data frame with 100 rows and 11 variables:

name senator's name

state state the senator represents

region geographic region of the U.S.

vote1 vote on perjury

vote2 vote on obstruction of justice

guilty total number of guilty votes

party political party of senator

conserva conservatism score, defined as the senator's degree of ideological conservatism, based on 1997 voting records as judged by the American Conservative Union, where the scores ranged from 0 to 100 and 100 is most conservative

supportc state voter support for Clinton, defined as the percent of the vote Clinton received in the 1996 presidential election in the senator's state

reelect year the senator's seat is up for reelection

newbie indicator for whether the senator is in their first-term

Learndis *Learning Disabilities in Elementary Students*

Description

This dataset is a subset of data from a study by Susan Tomasi and Sharon L. Weinberg (1999), which profiled learning disabled students in an urban setting. According to Public Law 94.142, enacted in 1976, a team may determine that a child has a learning disability (LD) if a severe discrepancy exists between a child's actual achievement in, for example, math or reading, and his or her intellectual ability. The dataset consists of six variables, described below, on 105 elementary school children from an urban area who were classified as LD and who, as a result, had been receiving special education services for at least three years. Of the 105 children, 42 are female and 63 are male. There are two main types of placements for these students: part-time resource room placements, in which the students get additional instruction to supplement regular classroom instruction, and self-contained classroom placements, in which students are segregated full time. In this dataset, 66 students are in resource room placements while 39 are in self-contained classroom placements. For inferential purposes, we consider the children in the dataset to be a random sample of all children attending public elementary school in a certain city who have been diagnosed with learning disabilities. Many students in the dataset have missing values for either math or reading comprehension, or both. Such omissions can lead to problems when generalizing results. There are statistical remedies for missing data that are beyond the scope of this text. In this case, we will assume that there is no pattern to the missing values, so that our sample is representative of the population.

Usage

```
Learndis
```

Format

A data frame with 105 rows and 6 variables:

grade student's grade level

gender student's gender

placemen type of placement: "RR" for part time in resource room or "MIS" for full time in self-contained classroom

readcomp reading comprehension score, with possible range of 0 to 200

mathcomp math comprehension score, with possible range of 0 to 200

iq student'a intellectual ability, as measured by IQ score with possible range of 0 to 200

References

"Classifying children as learning disabled: An analysis of current practice in an urban setting." Tomasi, S., & Weinberg, S. L. (1999) <doi:10.2307/1511150>

`levenes.test` *Levene's Test for Homogeneity of Variance*

Description

Function to test the homogeneity of variance for two populations, an assumption of the independent samples t-test. The null hypothesis tested is that the two population variances are equal; the alternative is that the two population variances are not equal.

Usage

```
levenes.test(y, group)
```

Arguments

`y`	outcome variable of interest, given as a numeric object.
`group`	a factor or character object with two levels indicating group membership.

Value

An anova table containing test results: two values for degrees of freedom, the F-value, and the p-value.

See Also

`t.test`

Examples

```
# using simple data frame
value = c(7,2,4,4,8,3,61,2,80,4)
grp = rep(c("A","B"), each = 5)
ex_data = data.frame(value = value, grp = grp)
levenes.test(ex_data$value, group = ex_data$grp)

# using variable without NA values
levenes.test(NELS$famsize, group = NELS$gender)

# using variable with NA values
levenes.test(NELS$achrdg12, group = NELS$gender)
```

leverage *Leverage*

Description

Returns the leverage values for a linear regression model.

Usage

```
leverage(x)
```

Arguments

x	linear regression model given as an lm object.

Value

A numeric vector of leverage values.

See Also

lm, rstudent(), cooks.distance()

Examples

```
mod = lm(Framingham$SYSBP1 ~ Framingham$TOTCHOL1 + Framingham$AGE1)
leverage(mod)
```

Likert *Likert-Scale Assertiveness Measure*

Description

This dataset contains fabricated data for a single survey item measured on a Likert scale. It is given that a survey was administered to 30 individuals and included an item measuring assertiveness by having the individual indicate agreement with the statement: “I have the ability to stand up for my own rights without denying the rights of others.” The response options were: 1 = "strongly agree"; 2 = "agree"; 3 = "neutral"; 4 = "disagree"; 5 = "strongly disagree." Notice that on this scale, high scores are associated with low levels of assertiveness.

Usage

```
Likert
```

Format

A data frame with 30 rows and 1 variable:

Assertiveness five-point Likert-scale score of assertiveness, with high scores associated with low levels of assertiveness

---

`line.graph` *Line Graph*

Description

Function to plot the estimated density values of a variable as a line.

Usage

```
line.graph(x, ...)
```

Arguments

`x`	numeric object to be plotted.
`...`	additional arguments to be passed to the `plot()` function.

Value

A line graph of the estimated density distribution of a variable.

See Also

`plot()`

Examples

```
line.graph(Temp$Temperature[Temp$City == "SanFrancisco"])
line.graph(IceCream$barsold)
```

ManDext *Manual Dexterity*

Description

This fictional dataset contains the treatment group number and the manual dexterity scores for 30 individuals selected by the director of a drug rehabilitation center. There are three treatments, and the individuals are randomly assigned ten to a treatment. After five weeks of treatment, a manual dexterity test is administered for which a higher score indicates greater manual dexterity.

Usage

```
ManDext
```

Format

A data frame with 30 rows and 3 variables:

ManualDex manual dexterity score

Sex individual's sex

Treatment treatment group assignment

ManDext2 *Manual Dexterity (Dataset #2)*

Description

This is a second fictional dataset that expands on ManDext, adding predicted outcome variables from regression analyses under alternative scenarios.

Usage

```
ManDext2
```

Format

A data frame with 30 rows and 9 variables:

ManualDex manual dexterity score

Sex individual's sex

Treatment treatment group assignment

yhat predicted outcome for disordinal interaction scenario

yhat2 predicted outcome for ordinal interaction scenario

yhat3 predicted outcome for first no-interaction scenario

yhat4 predicted outcome for second no-interaction scenario

yhat5 predicted outcome for third no-interaction scenario

yhat6 predicted outcome for fourth no-interaction scenario

`Marijuana` *Marijuana Use of Twelfth Graders*

Description

The dataset contains the year and percentage of twelfth graders who have ever used marijuana for several recent years. The source for these data is The World Almanac and Book of Facts 2014.

Usage

```
Marijuana
```

Format

A data frame with 23 rows and 2 variables:

Year year for which data was collected

MarijuanaUse percentage of twelfth graders who reported that they have ever used marijuana

---

`NELS` *National Education Longitudinal Study (NELS) of 1988*

Description

In response to pressure from federal and state agencies to monitor school effectiveness in the United States, the National Center of Education Statistics (NCES) of the U.S. Department of Education conducted a survey in the spring of 1988, the National Education Longitudinal Study (NELS). The participants consisted of a nationally representative sample of approximately 25,000 eighth graders to measure achievement outcomes in four core subject areas (English, history, mathematics, and science), in addition to personal, familial, social, institutional, and cultural factors that might relate to these outcomes. Details on the design and initial analysis of this survey may be referenced in Horn, Hafner, and Owings (1992). A follow-up of these students was conducted during tenth grade in the spring of 1990; a second follow-up was conducted during the twelfth grade in the spring of 1992; and, finally, a third follow-up was conducted in the spring of 1994.

Usage

```
NELS
```

Format

A data frame with 500 rows and 48 variables:

id case number

advmath8 indicator for whether advanced math taken in eighth grade

urban urbanicity, a measure of the type of environment in which the student lives

region geographic region of school

gender student's gender

famsize student's family size

parmarl8 parents' marital status in eighth grade

homelang home language background

slfcnc08 self-concept in eighth grade

slfcnc10 self-concept in tenth grade

slfcnc12 self-concept in twelfth grade

schtyp8 school type in eighth grade

tcherint likert-scale variable classifying student agreement with the statement, "My teachers are interested in students"

late12 number of times late for school in twelfth grade

cuts12 number of times skipped/cut classes in twelfth grade

absent12 number of times student missed school in twelfth grade

approg indicator for whether advanced placement program taken

hwkin12 time spent on homework weekly in school per week in twelfth grade

hwkout12 time spent on homework out of school per week in twelfth grade

excurr12 time spent weekly on extracurricular activities in twelfth grade, in hours

computer indicator for whether computer owned by family in eighth grade

hsprog type of high school program

unitengl units in English (NAEP), or number of years of English taken in high school

unitmath units in mathematics (NAEP), or number of years of math taken in high school

unitcalc units in calculus (NAEP), or number of years of calculus taken in high school

schattrt school average daily attendance rate

apoffer number of advanced placement courses offered by school

nursery indicator for whether nursery school attended

algebra8 indicator for whether algebra taken in eighth grade

numinst number of post-secondary institutions attended

edexpect highest level of education expected

expinc30 expected income at age 30, in dollars

achrdg08 reading achievement in eighth grade

achmat08 math achievement in eighth grade

achsci08 science achievement in eighth grade

achsls08 social studies achievement in eighth grade

achrdg10 reading achievement in tenth grade

achmat10 math achievement in tenth grade

achsci10 science achievement in tenth grade

achsls10 social studies achievement in tenth grade

achrdg12 reading achievement in twelfth grade

achmat12 math achievement in twelfth grade

achsci12 science achievement in twelfth grade

achsls12 social studies achievement in twelfth grade

cigarett indicator for whether smoked cigarettes ever

alcbinge indicator for whether ever binged on alcohol

marijuan indicator for whether smoked marijuana ever

ses socioeconomic status score, ranging from 0 to 35, and given as a composite of father's education level, mother's education level, father's occupation, mother's education, and family income

Details

For this dataset, we have selected a sub-sample of 500 cases and 48 variables. The cases were sampled randomly from the approximately 5,000 students who responded to all four administrations of the survey, who were always at grade level (neither repeated nor skipped a grade), and who pursued some form of post-secondary education. The particular variables were selected to explore the relationships between student and home background variables, self-concept, educational and income aspirations, academic motivation, risk-taking behavior, and academic achievement.

References

"A profile of American eighth-grade mathematics and science instruction." Horn, L., Hafner, & Owings (1992) <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=92486>

`percent.table` *Percentage Table*

Description

For one variable, returns a frequency distribution table given in percentages. For two variables, returns a contingency table given in percentages.

Usage

```
percent.table(x, y = NULL)
```

Arguments

x	object containing data for a single variable.
y	optional second object to create a contingency table given in percentages. Default setting ignores second object by setting y = NULL.

Value

A table of frequency percentages (for one variable) or a contingency table of percentages (for two variables).

See Also

cumulative.table, table

Examples

```
# frequency table for one variable
percent.table(NELS$region)

# cross-tabulation for two variables
percent.table(Wages$south,Wages$occup)
```

Politics *Gender and Political Party Affiliation*

Description

This dataset contains data on a fabricated random sample of 200 individuals, 100 females and 100 males, drawn from a population of interest. There are only two variables, both of which are categorical: gender and political party affiliation.

Usage

```
Politics
```

Format

A data frame with 200 rows and 2 variables:

Gender individual's gender

Party individual's political party affiliation

se.skew *Standard Error of Skewness*

Description

Function to obtain the standard error of the skewness of a distribution of values.

Usage

```
se.skew(x)
```

Arguments

x	numeric object containing the values for a variable.

Details

Standard error of skewness is computed on non-missing values using the following equation.

$$\sqrt{(6 * N * (N-1)/((N-2) * (N+1) * (N+3)))}$$

Value

Standard error of skewness for x.

See Also

skew, skew.ratio

Examples

```
se.skew(Temp$Temperature[Temp$City == "Springfield"])
se.skew(Temp$Temperature[Temp$City == "SanFrancisco"])
```

skew *Skewness of a Distribution*

Description

Function to obtain the skewness value of a distribution of values.

Usage

```
skew(x)
```

Arguments

x	numeric object containing the values for a variable.

Details

Skewness value computed on non-missing values using the ratio of $\Sigma((x-m)^3)/N$ to $\sqrt{(\Sigma((x-m)^2)/N)^3}$.

Value

Skewness value of x.

See Also

se.skew, skew.ratio

Examples

```
skew(IceCream$relhumid)
skew(IceCream$temp)
```

---

skew.ratio *Skewness Ratio*

Description

Returns the ratio of a distribution's skewness value to its standard error of skewness.

Usage

```
skew.ratio(x)
```

Arguments

x	numeric object containing the values for a variable.

Details

skew.ratio relies on the functions skew and se.skew to compute the skewness value and standard error of skewness, respectively.

Value

Skewness ratio of x.

See Also

skew, se.skew

Examples

```
# skew ratio computed two ways
skew.ratio(NELS$achmat12)
skew(NELS$achmat12) / se.skew(NELS$achmat12)
```

States *Educational Measures of the 50 States and Washington, D.C.*

Description

This dataset includes different educational measures of the 50 states and Washington, D.C. These data are from The 2014 World Almanac and Book of Facts.

Usage

```
States
```

Format

A data frame with 51 rows and 10 variables:

state name of state

region region of the country in which the state is located

enrollmt total public school enrollment 2011 - 2012

stuteach average number of pupils per teacher 2011 - 2012

teachpay average annual salary for public school teachers 2011 - 2012

educexpe average expenditure per pupil 2011 - 2012

satcr average SAT Critical Reading score 2013

satm average SAT Math score 2013

satw average SAT Writing score 2013

pertak percentage of eligible students taking the SAT 2012

Statisticians *Significant Statisticians*

Description

This dataset includes data on 12 statisticians who each have contributed significantly to the field of modern statistics.

Usage

```
Statisticians
```

Format

A data frame with 12 rows and 5 variables:

Statistician name of statistician

Gender gender of statistician, where 1 = "Female" and 2 = "Male"

Birth year of birth

Death year of death

AmStat number of references in *The American Statistician*, 1995-2005

Stepping *Stepping and Heart Rate*

Description

Students at Ohio State University conducted an experiment in the fall of 1993 to explore the nature of the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable, heart rate, was measured in beats per minute. For each person, the resting heart rate was measured before a trial (HRInit) and after stepping (HRFinal). There were two different step heights (Height): 5.75 inches (coded as 1 = Low), and 11.5 inches (coded as 2 = High). There were three rates of stepping (Freq): 14 steps/min. (coded as 1 = Slow), 21 steps/min. (coded as 2 = Medium), and 28 steps/min. (coded as 3 = Fast). This resulted in six possible height/frequency combinations. Each subject performed the activity for three minutes. Subjects were kept on pace by the beat of an electric metronome. One experimenter counted the subject's heart rate, in beats per minute, for 20 seconds before and after each trial. The subject always rested between trials until her or his heart rate returned to close to the beginning rate. Another experimenter kept track of the time spent stepping. Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment. The dataset and description are adapted from the Data and Story Library (DASL) website.

Usage

```
Stepping
```

Format

A data frame with 30 rows and 6 variables:

Order overall performance order of the trial

Block subject and experimenters' block number

Height step height

Freq rate of stepping

HRInit resting heart rate of the subject before a trial, in beats per minute

HRFinal final heart rate of the subject after a trial, in beats per minute

---

Temp *Average Monthy Temperatures for Two Cities*

Description

This dataset gives the average monthly temperatures (in degrees Fahrenheit) for Springfield, MO and San Francisco, CA. These data are from Burrill and Hopensperger (1993).

Usage

```
Temp
```

Format

A data frame with 24 rows and 2 variables:

City city where temperature was measured

Temperature average monthly temperature, in degrees Fahrenheit

References

"Exploring Statistics with the T1-81" Burrill, G., & Hopensperger, P. (1993, ISBN:9780201524321)

the.mode *Mode*

Description

Function to obtain the mode(s) of a distribution.

Usage

```
the.mode(x)
```

Arguments

x	object containing data for a single variable.

Value

A numeric vector of the value(s) of the distribution that have the highest frequency of occurrence.

See Also

mean, median

Examples

```
# single mode for factor variable
the.mode(NELS$urban)
# bimodal numeric variable
a = c(14,24,62,12,12,12,36,17,11,99,99,99)
the.mode(a)
```

UpperBodyStrength *Upper Body Strength*

Description

This simulated dataset consists of the number of hours eight individuals spend at the gym on a weekly basis along with measures of their upper body strength.

Usage

```
UpperBodyStrength
```

Format

A data frame with 8 rows and 3 variables:

gym number of hours spent at the gym weekly

strength upper body strength score

gender individual's gender

`Wages` *Wage and Education Data from the 1985 Current Population Survey*

Description

This is a subsample of 100 males and 100 females randomly selected from the 534 cases that comprised the 1985 Current Population Survey in a way that controls for highest education level attained. The sample of 200 contains 20 males and 20 females with less than a high school diploma, 20 males and 20 females with a high school diploma, 20 males and 20 females with some college training, 20 males and 20 females with a college diploma, and 20 males and 20 females with some graduate school training. The data include information about gender, highest education level attained, and hourly wage.

Usage

```
Wages
```

Format

A data frame with 400 rows and 9 variables:

id case number

educ number of years of education

south indicator for whether individual lives in the South

sex individual's sex

exper number of years of work experience

wage wage (dollars per hour)

occup occupation category

marr marital status

ed highest education level

Index