Package 'HSAUR3'

September 29, 2026

Title A Handbook of Statistical Analyses Using R (3rd Edition)

Version 1.0-16

Description Functions, data sets, analyses and examples from the third edition of the book "A Handbook of Statistical Analyses Using R" (Torsten Hothorn and Brian S. Everitt, Chapman & Hall/CRC, 2014). The first chapter of the book, which is entitled "An Introduction to R", is completely included in this package, for all other chapters, a vignette containing all data analyses is available. In addition, Sweave source code for slides of selected chapters is included in this package (see HSAUR3/inst/slides). The publishers web page is '<https://www.routledge.com/A-Handbook-of-Statistical-Analyses-using-R/Hothorn-Everitt/p/book/9781482204582>'.

Depends R (>= 3.0.0)

Suggests boot (>= 1.3-9), lattice (>= 0.20-23), MASS (>= 7.3-29), mgcv (>= 1.7-27), rpart (>= 4.1-4), survival (>= 2.37-4), ape (>= 3.0-11), coin (>= 1.1-3), flexmix (>= 2.3-11), Formula (>= 1.1-1), gamair (>= 0.0.8), gamlss.data (>= 4.2.6), gee (>= 4.13-18), KernSmooth (>= 2.23-10), lme4 (>= 1.0-5), maps (>= 2.3-6), mboost (>= 2.2-3), mclust (>= 4.2), mlbench (>= 2.1-1), mice (>= 2.18), multcomp (>= 1.3-1), mvtnorm (>= 0.9-9996), partykit (>= 0.8-0), quantreg (>= 5.05), randomForest (>= 4.6-7), rmeta (>= 2.16), sandwich (>= 2.3-0), scatterplot3d (>= 0.3-34), sf (>= 1.0-9), sp (>= 1.0-14), TH.data (>= 1.0-2), vcd (>= 1.3-1), wordcloud (>= 2.4), HSAUR2

LazyData yes

License GPL-2

Encoding UTF-8

NeedsCompilation no

Author Torsten Hothorn [aut, cre] (ORCID: <https://orcid.org/0000-0001-8301-0471>), Brian S. Everitt [aut]

Maintainer Torsten Hothorn <Torsten.Hothorn@R-project.org>

Repository CRAN

Date/Publication 2026-09-29 14:50:02 UTC

Contents

agefat . . . 3
aspirin . . . 4
backpain . . . 5
BCG . . . 6
birds . . . 7
birthdeathrates . . . 8
bladdercancer . . . 8
bp . . . 9
BtheB . . . 10
CHFLS . . . 11
clouds . . . 13
CYGOB1 . . . 15
EFT . . . 16
epilepsy . . . 16
Forbes2000 . . . 17
foster . . . 18
gardenflowers . . . 19
GHQ . . . 19
heptathlon . . . 20
household . . . 21
HSAURtable . . . 22
Lanza . . . 23
mastectomy . . . 24
men1500m . . . 25
meteo . . . 26
orallesions . . . 27
phosphate . . . 27
pistonrings . . . 28
planets . . . 29
plasma . . . 30
polyps . . . 31
polyps3 . . . 32
pottery . . . 33
rearrests . . . 34
respiratory . . . 34
roomwidth . . . 35
schizophrenia . . . 36
schizophrenia2 . . . 37
schooldays . . . 38
skulls . . . 39
smoking . . . 40
Smoking_DollHill1950 . . . 41

Smoking_Mueller1940 . . . 41
Smoking_SchairerSchoeniger1944 . . . 42
Smoking_Wassink1945 . . . 43
students . . . 43
suicides . . . 44
suicides2 . . . 45
toenail . . . 45
toothpaste . . . 46
USairpollution . . . 47
USmelanoma . . . 48
USstates . . . 49
USTemp . . . 49
voting . . . 50
water . . . 51
watervoles . . . 52
waves . . . 52
weightgain . . . 53
womensrole . . . 54

Index **56**

`agefat` *Total Body Composision Data*

Description

Age and body fat percentage of 25 normal adults.

Usage

```
data("agefat")
```

Format

A data frame with 25 observations on the following 3 variables.

`age` the age of the subject.

`fat` the body fat percentage.

`gender` a factor with levels `female` and `male`.

Details

The data come from a study investigating a new methods of measuring body composition (see Mazess et al, 1984), and give the body fat percentage (percent fat), age and gender for 25 normal adults aged between 23 and 61 years. The questions of interest are how are age and percent fat related, and is there any evidence that the relationship is different for males and females.

Source

R. B. Mazess, W. W. Peppler and M. Gibbons (1984), Total body composition by dual-photon (153Gd) absorptiometry. *American Journal of Clinical Nutrition*, **40**, 834–839.

Examples

```
data("agefat", package = "HSAUR3")
plot(fat ~ age, data = agefat)
```

---

aspirin *Aspirin Data*

Description

Efficacy of Aspirin in preventing death after a myocardial infarct.

Usage

```
data("aspirin")
```

Format

A data frame with 7 observations on the following 4 variables.

`dp` number of deaths after placebo.

`tp` total number subjects treated with placebo.

`da` number of deaths after Aspirin.

`ta` total number of subjects treated with Aspirin.

Details

The data were collected for a meta-analysis of the effectiveness of Aspirin (versus placebo) in preventing death after a myocardial infarction.

Source

J. L. Fleiss (1993), The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2**, 121–145.

Examples

```
data("aspirin", package = "HSAUR3")
aspirin
```

backpain *Driving and Back Pain Data*

Description

A case-control study to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID).

Usage

```
data("backpain")
```

Format

A data frame with 434 observations on the following 4 variables.

`ID` a factor which identifies matched pairs.

`status` a factor with levels `case` and `control`.

`driver` a factor with levels `no` and `yes`.

`suburban` a factor with levels `no` and `yes` indicating a suburban resident.

Details

These data arise from a study reported in Kelsey and Hardy (1975) which was designed to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID). A case-control study was used with cases selected from people who had recently had X-rays taken of the lower back and had been diagnosed as having AHLID. The controls were taken from patients admitted to the same hospital as a case with a condition unrelated to the spine. Further matching was made on age and sex and a total of 217 matched pairs were recruited, consisting of 89 female pairs and 128 male pairs.

Source

Jennifer L. Kelsey and Robert J. Hardy (1975), Driving of Motor Vehicles as a Risk Factor for Acute Herniated Lumbar Intervertebral Disc. *American Journal of Epidemiology*, **102**(1), 63–73.

Examples

```
data("backpain", package = "HSAUR3")
summary(backpain)
```

BCG *BCG Vaccine Data*

Description

A meta-analysis on the efficacy of BCG vaccination against tuberculosis (TB).

Usage

```
data("BCG")
```

Format

A data frame with 13 observations on the following 7 variables.

`Study` an identifier of the study.

`BCGTB` the number of subjects suffering from TB after a BCG vaccination.

`BCGVacc` the number of subjects with BCG vaccination.

`NoVaccTB` the number of subjects suffering from TB without BCG vaccination.

`NoVacc` the total number of subjects without BCG vaccination.

`Latitude` geographic position of the place the study was undertaken.

`Year` the year the study was undertaken.

Details

Bacille Calmette Guerin (BCG) is the most widely used vaccination in the world. Developed in the 1930s and made of a live, weakened strain of Mycobacterium bovis, the BCG is the only vaccination available against tuberculosis today. Colditz et al. (1994) report data from 13 clinical trials of BCG vaccine each investigating its efficacy in the treatment of tuberculosis. The number of subjects suffering from TB with or without BCG vaccination are given here. In addition, the data contains the values of two other variables for each study, namely, the geographic latitude of the place where the study was undertaken and the year of publication. These two variables will be used to investigate and perhaps explain any heterogeneity among the studies.

Source

G. A. Colditz, T. F. Brewer, C. S. Berkey, M. E. Wilson, E. Burdick, H. V. Fineberg and F. Mosteller (1994), Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *Journal of the American Medical Association*, **271**(2), 698–702.

Examples

```
data("BCG", package = "HSAUR3")

### sort studies w.r.t. sample size
BCG <- BCG[order(rowSums(BCG[,2:5])),]

### to long format
BCGlong <- with(BCG, data.frame(Freq = c(BCGTB, BCGVacc - BCGTB,
                                         NoVaccTB, NoVacc - NoVaccTB),
                infected = rep(rep(factor(c("yes", "no")),
                               rep(nrow(BCG), 2)), 2),
                vaccined = rep(factor(c("yes", "no")),
                               rep(nrow(BCG) * 2, 2)),
                study = rep(factor(Study, levels = as.character(Study)),
                            4)))

### doubledecker plot
library("vcd")
doubledecker(xtabs(Freq ~ study + vaccined + infected,
                   data = BCGlong))
```

`birds` *Birds in Paramo Vegetation Data*

Description

The data were originally derived from a study which investigated numbers of bird species in isolated islands of paramo vegetation in the northern Andes.

Usage

```
data(birds)
```

Format

A data frame with 14 observations on the following 5 variables.

`N` number of species

`AR` area of island in thousands of square km

`EL` elevation in thousands of m

`Dec` distance from Ecuador in km

`DNI` distance to the nearest island in km

Source

F. Vuilleumier (1970), Insular biogeography in continental regions. I. The northern Andes of South America. *The American Naturalist* **104**, 373–388

birthdeathrates *Birth and Death Rates Data*

Description

Birth and death rates for 69 countries.

Usage

```
data("birthdeathrates")
```

Format

A data frame with 69 observations on the following 2 variables.

`birth` birth rate.

`death` death rate.

Source

J. A. Hartigan (1975), *Clustering Algorithms*. John Wiley & Sons, New York.

Examples

```
data("birthdeathrates", package = "HSAUR3")
plot(birthdeathrates)
```

bladdercancer *Bladder Cancer Data*

Description

Data arise from 31 male patients who have been treated for superficial bladder cancer, and give the number of recurrent tumours during a particular time after the removal of the primary tumour, along with the size of the original tumour.

Usage

```
data("bladdercancer")
```

Format

A data frame with 31 observations on the following 3 variables.

`time` the duration.

`tumorsize` a factor with levels `<=3cm` and `>3cm`.

`number` number of recurrent tumours.

Details

The aim is the estimate the effect of size of tumour on the number of recurrent tumours.

Source

G. U. H. Seeber (1998), Poisson Regression. In: *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds), John Wiley & Sons, Chichester.

Examples

```
data("bladdercancer", package = "HSAUR3")
mosaicplot(xtabs(~ number + tumorsize, data = bladdercancer))
```

---

bp *Lowering Blood Pressure Data*

Description

Lowering a patient's blood pressure during surgery, using a hypotensive drug.

Usage

```
data(bp)
```

Format

A data frame with 53 observations on the following 3 variables.

`logdose` the logarithm (base 10) of the dose of drug in milligrams

`bloodp` average systolic blood pressure achieved while the drug was being administered

`recovtime` time in minutes before the patient's systolic blood pressure returned to 100mm of mercury

Source

J. D. Robertson and P. Armitage (1959) Comparison of Two Hypotensive Agents, *Anaesthesia*, **14**(1), 53–64

BtheB *Beat the Blues Data*

Description

Data from a clinical trial of an interactive multimedia program called 'Beat the Blues'.

Usage

```
data("BtheB")
```

Format

A data frame with 100 observations of 100 patients on the following 8 variables.

drug did the patient take anti-depressant drugs (`No` or `Yes`).

length the length of the current episode of depression, a factor with levels `<6m` (less than six months) and `>6m` (more than six months).

treatment treatment group, a factor with levels `TAU` (treatment as usual) and `BtheB` (Beat the Blues)

bdi.pre Beck Depression Inventory II before treatment.

bdi.2m Beck Depression Inventory II after two months.

bdi.3m Beck Depression Inventory II after one month follow-up.

bdi.5m Beck Depression Inventory II after three months follow-up.

bdi.8m Beck Depression Inventory II after six months follow-up.

Details

Longitudinal data from a clinical trial of an interactive, multimedia program known as "Beat the Blues" designed to deliver cognitive behavioural therapy to depressed patients via a computer terminal. Patients with depression recruited in primary care were randomised to either the Beating the Blues program, or to "Treatment as Usual (TAU)".

Note that the data are stored in the wide form, i.e., repeated measurments are represented by additional columns in the data frame.

Source

J. Proudfoot, D. Goldberg, A. Mann, B. S. Everitt, I. Marks and J. A. Gray, (2003). Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological Medicine*, **33**(2), 217–227.

Examples

```
data("BtheB", package = "HSAUR3")
layout(matrix(1:2, nrow = 1))
ylim <- range(BtheB[,grep("bdi", names(BtheB))], na.rm = TRUE)
boxplot(subset(BtheB, treatment == "TAU")[,grep("bdi", names(BtheB))],
        main = "Treated as usual", ylab = "BDI",
        xlab = "Time (in months)", names = c(0, 2, 3, 5, 8), ylim = ylim)
boxplot(subset(BtheB, treatment == "BtheB")[,grep("bdi", names(BtheB))],
        main = "Beat the Blues", ylab = "BDI", xlab = "Time (in months)",
        names = c(0, 2, 3, 5, 8), ylim = ylim)
```

CHFLS *Chinese Health and Family Life Survey*

Description

The Chinese Health and Family Life Survey sampled 60 villages and urban neighborhoods chosen in such a way as to represent the full geographical and socioeconomic range of contemporary China.

Usage

```
data("CHFLS")
```

Format

A data frame with 1534 observations on the following 10 variables.

`R_region` a factor with levels `Coastal South`, `Coastal East`, `Inlands`, `North`, `Northeast`, `Central West`.

`R_age` age of the responding woman.

`R_edu` education level of the responding woman, an ordered factor with levels `Never attended school` < `Elementary school` < `Junior high school` < `Senior high school` < `Junior college` < `University`.

`R_income` monthly income of the responding woman.

`R_health` self-reported health status, an ordered factor with levels `Poor` < `Not good` < `Fair` < `Good` < `Excellent`.

`R_height` height of the responding woman.

`R_happy` self-reportet happiness of the responding woman, an ordered factor with levels `Very unhappy` < `Not too happy` < `Somewhat happy` < `Very happy`.

`A_height` height of the woman's partner.

`A_edu` level of education of the woman's partner, an ordered factor with levels `Never attended school` < `Elementary school` < `Junior high school` < `Senior high school` < `Junior college` < `University`.

`A_income` montjly income of the woman's partner.

Details

Contemporary China is on the leading edge of a sexual revolution, with tremendous regional and generational differences that provide unparalleled natural experiments for analysis of the antecedents and outcomes of sexual behavior. The Chinese Health and Family Life Study, conducted 1999–2000 as a collaborative research project of the Universities of Chicago, Beijing, and North Carolina, provides a baseline from which to anticipate and track future changes. Specifically, this study produces a baseline set of results on sexual behavior and disease patterns, using a nationally representative probability sample. The Chinese Health and Family Life Survey sampled 60 villages and urban neighborhoods chosen in such a way as to represent the full geographical and socioeconomic range of contemporary China excluding Hong Kong and Tibet. Eighty-three individuals were chosen at random for each location from official registers of adults aged between 20 and 64 years to target a sample of 5000 individuals in total. Here, we restrict our attention to women with current male partners for whom no information was missing, leading to a sample of 1534 women. The data have been extracted as given in the example section.

Source

https://sscs.uchicago.edu

References

William L. Parish, Edward O. Laumann, Myron S. Cohen, Suiming Pan, Heyi Zheng, Irving Hoffman, Tianfu Wang, and Kwai Hang Ng. (2003), Population-Based Study of Chlamydial Infection in China: A Hidden Epidemic. *Journal of the American Medican Association*, **289**(10), 1265–1273.

Examples

```
## Not run:

    ### for a description see http://popcenter.uchicago.edu/data/chfls.shtml
    library("TH.data")
    load(file.path(path.package(package="TH.data"), "rda", "CHFLS.rda"))

    tmp <- chfls1[, c("REGION6", "ZJ05", "ZJ06", "A35", "ZJ07", "ZJ16M", "INCRM",
                      "JK01", "JK02", "JK20", "HY04", "HY07", "A02", "AGEGAPM",
                  "A07M", "A14", "A21", "A22M", "A23", "AX16", "INCAM", "SEXNOW", "ZW04")]

     names(tmp) <- c("Region",
                 "Rgender",               ### gender of respondent
                 "Rage",                  ### age of respondent
  "RagestartA",   ### age of respondent at beginning of relationship
                                          ### with partner A
                 "Redu",                  ### education of respondent
                 "RincomeM",              ### rounded monthly income of respondent
  "RincomeComp",   ### inputed monthly income of respondent
                 "Rhealth",               ### health condition respondent
                 "Rheight",               ### respondent's height
                 "Rhappy",                ### respondent's happiness
                 "Rmartial",              ### respondent's marital status
                 "RhasA",                 ### R has current A partner
                 "Agender",               ### gender of partner A
```

```
                "RAagegap",             ### age gap
                "RAstartage",           ### age at marriage
                "Aheight",              ### height of partner A
                "Aedu",                 ### education of partner A
                "AincomeM",             ### rounded partner A income
                "AincomeEst",           ### estimated partner A income
                "orgasm",               ### orgasm frequency
                "AincomeComp",          ### imputed partner A income
                "Rsexnow",              ### has sex last year
                "Rhomosexual")          ### R is homosexual

 ### code missing values
 tmp$AincomeM[tmp$AincomeM < 0] <- NA
 tmp$RincomeM[tmp$RincomeM < 0] <- NA
 tmp$Aheight[tmp$Aheight < 0] <- NA

 olevels <- c("never", "rarely", "sometimes", "often", "always")
 tmpA <- subset(tmp, Rgender == "female" & Rhomosexual != "yes" & orgasm %in% olevels)

 ### 1534 subjects
 dim(tmpA)

CHFLS <- tmpA[, c("Region", "Rage", "Redu", "RincomeComp", "Rhealth", "Rheight", "Rhappy",
                  "Aheight", "Aedu", "AincomeComp")]
 names(CHFLS) <- c("R_region", "R_age", "R_edu", "R_income", "R_health", "R_height",
                  "R_happy", "A_height", "A_edu", "A_income")
 levels(CHFLS$R_region) <- c("Coastal South", "Coastal Easth", "Inlands", "North",
                             "Northeast", "Central West")

 CHFLS$R_edu <- ordered(as.character(CHFLS$R_edu), levels = c("no school", "primary",
                        "low mid", "up mid", "j col", "univ/grad"))
levels(CHFLS$R_edu) <- c("Never attended school", "Elementary school", "Junior high school",
                         "Senior high school", "Junior college", "University")
 CHFLS$A_edu <- ordered(as.character(CHFLS$A_edu), levels = c("no school", "primary",
                        "low mid", "up mid", "j col", "univ/grad"))
levels(CHFLS$A_edu) <- c("Never attended school", "Elementary school", "Junior high school",
                         "Senior high school", "Junior college", "University")

 CHFLS$R_health <- ordered(as.character(CHFLS$R_health), levels = c("poor", "not good",
                           "fair", "good", "excellent"))
 levels(CHFLS$R_health) <- c("Poor", "Not good", "Fair", "Good", "Excellent")

CHFLS$R_happy <- ordered(as.character(CHFLS$R_happy), levels = c("v unhappy", "not too",
                         "relatively", "very"))
levels(CHFLS$R_happy) <- c("Very unhappy", "Not too happy", "Relatively happy", "Very happy")

## End(Not run)
```

---

clouds *Cloud Seeding Data*

Description

Data from an experiment investigating the use of massive amounts of silver iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall.

Usage

```
data("clouds")
```

Format

A data frame with 24 observations on the following 7 variables.

seeding a factor indicating whether seeding action occured (`no` or `yes`).

time number of days after the first day of the experiment.

sne suitability criterion.

cloudcover the percentage cloud cover in the experimental area, measured using radar.

prewetness the total rainfall in the target area one hour before seeding (in cubic metres times `1e+8`).

echomotion a factor showing whether the radar echo was `moving` or `stationary`.

rainfall the amount of rain in cubic metres times `1e+8`.

Details

Weather modification, or cloud seeding, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall. Introduction of such material into a cloud that contains supercooled water, that is, liquid water colder than zero Celsius, has the aim of inducing freezing, with the consequent ice particles growing at the expense of liquid droplets and becoming heavy enough to fall as rain from clouds that otherwise would produce none.

The data available in `cloud` were collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide 100 to 1000 grams per cloud) in cloud seeding to increase rainfall. In the experiment, which was conducted in an area of Florida, 24 days were judged suitable for seeding on the basis that a measured suitability criterion (`SNE`).

Source

W. L. Woodley, J. Simpson, R. Biondini and J. Berkeley (1977), Rainfall results 1970-75: Florida area cumulus experiment. *Science* **195**, 735–742.

R. D. Cook and S. Weisberg (1980), Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495–508.

Examples

```
data("clouds", package = "HSAUR3")
layout(matrix(1:2, nrow = 2))
boxplot(rainfall ~ seeding, data = clouds, ylab = "Rainfall")
boxplot(rainfall ~ echomotion, data = clouds, ylab = "Rainfall")
```

CYGOB1 *CYG OB1 Star Cluster Data*

Description

Energy output and surface termperature for Star Cluster CYG OB1.

Usage

```
data("CYGOB1")
```

Format

A data frame with 47 observations on the following 2 variables.

`logst` log survface termperature of the star.

`logli` log light intensity of the star.

Details

The Hertzsprung-Russell (H-R) diagram forms the basis of the theory of stellar evolution. The diagram is essentially a plot of the energy output of stars plotted against their surface temperature. Data from the H-R diagram of Star Cluster CYG OB1, calibrated according to VanismaGreve1972 are given here.

Source

F. Vanisma and J. P. De Greve (1972), Close binary systems before and after mass transfer. *Astrophysics and Space Science*, **87**, 377–401.

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.

Examples

```
data("CYGOB1", package = "HSAUR3")
plot(logst ~ logli, data = CYGOB1)
```

EFT *Embedded Figures Test Data*

Description

Embedded figures test for 24 school children.

Usage

```
data(EFT)
```

Format

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

`group` a factor with levels `row` `corner`

`time` time to complete the pattern

`EFT` Embedded Figures Test

Source

M. Aitkin, D. Anderson, B. Francis, and J. Hinde (1989), *Statistical Modelling in GLIM*, Oxford University Press, New York, USA

epilepsy *Epilepsy Data*

Description

A randomised clinical trial investigating the effect of an anti-epileptic drug.

Usage

```
data("epilepsy")
```

Format

A data frame with 236 observations on the following 6 variables.

`treatment` the treatment group, a factor with levels `placebo` and `Progabide`.

`base` the number of seizures before the trial.

`age` the age of the patient.

`seizure.rate` the number of seizures (response variable).

`period` treatment period, an ordered factor with levels 1 to 4.

`subject` the patient ID, a factor with levels 1 to 59.

Details

In this clinical trial, 59 patients suffering from epilepsy were randomized to groups receiving either the anti-epileptic drug Progabide or a placebo in addition to standard chemotherapy. The numbers of seizures suffered in each of four, two-week periods were recorded for each patient along with a baseline seizure count for the 8 weeks prior to being randomized to treatment and age. The main question of interest is whether taking progabide reduced the number of epileptic seizures compared with placebo.

Source

P. F. Thall and S. C. Vail (1990), Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.

Examples

```
data("epilepsy", package = "HSAUR3")
library(lattice)
dotplot(I(seizure.rate / base) ~ period | subject, data = epilepsy,
        subset = treatment == "Progabide")
dotplot(I(seizure.rate / base) ~ period | subject, data = epilepsy,
        subset = treatment == "Progabide")
```

Forbes2000 *The Forbes 2000 Ranking of the World's Biggest Companies (Year 2004)*

Description

The Forbes 2000 list is a ranking of the world's biggest companies, measured by sales, profits, assets and market value.

Usage

```
data("Forbes2000")
```

Format

A data frame with 2000 observations on the following 8 variables.

rank the ranking of the company.

name the name of the company.

country a factor giving the country the company is situated in.

category a factor describing the products the company produces.

sales the amount of sales of the company in billion USD.

profits the profit of the company in billion USD.

assets the assets of the company in billion USD.

marketvalue the market value of the company in billion USD.

Source

https://www.forbes.com, assessed on November 26th, 2004.

Examples

```
data("Forbes2000", package = "HSAUR3")
summary(Forbes2000)
### number of countries
length(levels(Forbes2000$country))
### number of industries
length(levels(Forbes2000$category))
```

---

foster *Foster Feeding Experiment*

Description

The data are from a foster feeding experiment with rat mothers and litters of four different genotypes. The measurement is the litter weight after a trial feeding period.

Usage

```
data("foster")
```

Format

A data frame with 61 observations on the following 3 variables.

`litgen` genotype of the litter, a factor with levels `A`, `B`, `I`, and `J`.

`motgen` genotype of the mother, a factor with levels `A`, `B`, `I`, and `J`.

`weight` the weight of the litter after a feeding period.

Details

Here the interest lies in uncovering the effect of genotype of mother and litter on litter weight.

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.

Examples

```
data("foster", package = "HSAUR3")
plot.design(foster)
```

gardenflowers *Garden Flowers Data*

Description

The dissimilarity matrix of 18 species of garden flowers.

Usage

```
data("gardenflowers")
```

Format

An object of class `dist`.

Details

The dissimilarity was computed based on certain characteristics of the flowers.

Source

L. Kaufman and P. J. Rousseeuw (1990), *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, New York.

Examples

```
data("gardenflowers", package = "HSAUR3")
gardenflowers
```

GHQ *General Health Questionnaire*

Description

Data from an psychiatric screening questionnaire

Usage

```
data("GHQ")
```

Format

A data frame with 22 observations on the following 4 variables.

`GHQ` the General Health Questionnaire score.

`gender` a factor with levels `female` and `male`

`cases` the number of diseased subjects.

`non.cases` the number of healthy subjects.

Details

The data arise from a study of a psychiatric screening questionnaire called the GHQ (General Health Questionnaire, see Goldberg, 1972). Here the main question of interest is to see how caseness is related to gender and GHQ score.

Source

D. Goldberg (1972). *The Detection of Psychiatric Illness by Questionnaire*, Oxford University Press, Oxford, UK.

Examples

```
data("GHQ", package = "HSAUR3")
male <- subset(GHQ, gender == "male")
female <- subset(GHQ, gender == "female")
layout(matrix(1:2, ncol = 2))
barplot(t(as.matrix(male[,c("cases", "non.cases")])), main = "Male", xlab = "GHC score")
barplot(t(as.matrix(male[,c("cases", "non.cases")])), main = "Female", xlab = "GHC score")
```

`heptathlon` *Olympic Heptathlon Seoul 1988*

Description

Results of the olympic heptathlon competition, Seoul, 1988.

Usage

```
data("heptathlon")
```

Format

A data frame with 25 observations on the following 8 variables.

`hurdles` results 100m hurdles.

`highjump` results high jump.

`shot` results shot.

`run200m` results 200m race.

`longjump` results long jump.

`javelin` results javelin.

`run800m` results 800m race.

`score` total score.

Details

The first combined Olympic event for women was the pentathlon, first held in Germany in 1928. Initially this consisted of the shot putt, long jump, 100m, high jump and javelin events held over two days. The pentathlon was first introduced into the Olympic Games in 1964, when it consisted of the 80m hurdles, shot, high jump, long jump and 200m. In 1977 the 200m was replaced by the 800m and from 1981 the IAAF brought in the seven-event heptathlon in place of the pentathlon, with day one containing the events-100m hurdles, shot, high jump, 200m and day two, the long jump, javelin and 800m. A scoring system is used to assign points to the results from each event and the winner is the woman who accumulates the most points over the two days. The event made its first Olympic appearance in 1984.

In the 1988 Olympics held in Seoul, the heptathlon was won by one of the stars of women's athletics in the USA, Jackie Joyner-Kersee. The results for all 25 competitors are given here.

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.

Examples

```
data("heptathlon", package = "HSAUR3")
plot(heptathlon)
```

---

household *Household Expenditure Data*

Description

Survey data on household expenditure on four commodity groups.

Usage

```
data("household")
```

Format

A data frame with 40 observations on the following 5 variables.

`housing` expenditure on housing, including fuel and light.

`food` expenditure on foodstuffs, including alcohol and tobacco.

`goods` expenditure on other goods, including clothing, footwear and durable goods.

`service` expenditure on services, including transport and vehicles.

`gender` a factor with levels `female` and `male`

Details

The data are part of a data set collected from a survey of household expenditure and give the expenditure of 20 single men and 20 single women on four commodity groups. The units of expenditure are Hong Kong dollars,

Source

FIXME

Examples

```
data("household", package = "HSAUR3")
```

HSAURtable *Produce LaTeX Tables*

Description

Generate `longtable` LaTeX environments.

Usage

```
HSAURtable(object, ...)
## S3 method for class 'table'
HSAURtable(object, xname = deparse(substitute(object)), pkg = NULL,
    ...)
## S3 method for class 'data.frame'
HSAURtable(object, xname = deparse(substitute(object)), pkg = NULL,
    nrows = NULL,  ...)
## S3 method for class 'tabtab'
toLatex(object, caption = NULL, label = NULL,
    topcaption = TRUE, index = TRUE, ...)
## S3 method for class 'dftab'
toLatex(object, pcol = 1, caption = NULL,
    label = NULL, rownames = FALSE, topcaption = TRUE, index = TRUE,
    ...)
```

Arguments

object	an object of table or data.frame.
xname	the name of the object.
pkg	the package object comes from, optionally.
nrows	the number of rows actually printed for a data.frame.
caption	the (optional) caption of the table without label.
label	the (optional) label to be defined for this table.
pcol	the number of parallel columns.
rownames	logical, should the rownames be printed in the first row without column name?
topcaption	logical, should the captions be placed on top (default) of the table?
index	logical, should an index entry be generated?
...	additional arguments, currently ignored.

Details

Based on the data in object, an object from which a Latex table (in a longtable environment) may be constructed (via toLatex) is generated.

Value

An object of class tabtab or dftab for which toLatex methods are available.

toLatex produces objects of class Latex, a character vector, essentially.

Examples

```
data("rearrests", package = "HSAUR3")
toLatex(HSAURtable(rearrests),
        caption = "Rearrests of juvenile felons.",
        label = "rearrests_tab")
```

Lanza *Prevention of Gastointestinal Damages*

Description

Data from four randomised clinical trials on the prevention of gastointestinal damages by Misoprostol reported by Lanza et al. (1987, 1988a,b, 1989).

Usage

```
data("Lanza")
```

Format

A data frame with 198 observations on the following 3 variables.

`study` a factor with levels `I`, `II`, `III`, and `IV` describing the study number.

`treatment` a factor with levels `Misoprostol` `Placebo`

`classification` an ordered factor with levels `1` < `2` < `3` < `4` < `5` describing an ordered response variable.

Details

The response variable is defined by the number of haemorrhages or erosions.

Source

F. L. Lanza (1987), A double-blind study of prophylactic effect of misoprostol on lesions of gastric and duodenal mucosa induced by oral administration of tolmetin in healthy subjects. *British Journal of Clinical Practice*, May suppl, 91–101.

F. L. Lanza, R. L. Aspinall, E. A. Swabb, R. E. Davis, M. F. Rack, A. Rubin (1988a), Double-blind, placebo-controlled endoscopic comparison of the mucosal protective effects of misoprostol versus cimetidine on tolmetin-induced mucosal injury to the stomach and duodenum. *Gastroenterology*, **95**(2), 289–294.

F. L. Lanza, K. Peace, L. Gustitus, M. F. Rack, B. Dickson (1988b), A blinded endoscopic comparative study of misoprostol versus sucralfate and placebo in the prevention of aspirin-induced gastric and duodenal ulceration. *American Journal of Gastroenterology*, **83**(2), 143–146.

F. L. Lanza, D. Fakouhi, A. Rubin, R. E. Davis, M. F. Rack, C. Nissen, S. Geis (1989), A double-blind placebo-controlled comparison of the efficacy and safety of 50, 100, and 200 micrograms of misoprostol QID in the prevention of ibuprofen-induced gastric and duodenal mucosal lesions and symptoms. *American Journal of Gastroenterology*, **84**(6), 633–636.

Examples

```
data("Lanza", package = "HSAUR3")
layout(matrix(1:4, nrow = 2))
pl <- tapply(1:nrow(Lanza), Lanza$study, function(indx)
    mosaicplot(table(Lanza[indx,"treatment"],
                     Lanza[indx,"classification"]),
               main = "", shade = TRUE))
```

`mastectomy` *Survival Times after Mastectomy of Breast Cancer Patients*

Description

Survival times in months after mastectomy of women with breast cancer. The cancers are classified as having metastized or not based on a histochemical marker.

Usage

```
data("mastectomy")
```

Format

A data frame with 42 observations on the following 3 variables.

time survival times in months.

event a logical indicating if the event was observed (TRUE) or if the survival time was censored (FALSE).

metastasized a factor at levels yes and no.

Source

B. S. Everitt and S. Rabe-Hesketh (2001), *Analysing Medical Data using S-PLUS*, Springer, New York, USA.

Examples

```
data("mastectomy", package = "HSAUR3")
table(mastectomy$metastasized)
```

---

men1500m *Winners of the Olympic Men's 1500m*

Description

The data gives the winners of the men's 1500m race for the Olympic Games 1896 to 2004.

Usage

```
data("men1500m")
```

Format

A data frame with 25 observations on the following 5 variables.

`year` the olympic year.

`venue` city where the games took place.

`winner` winner of men's 1500m race.

`country` country the winner came from.

`time` time (in seconds) of the winner.

Examples

```
data("men1500m", package = "HSAUR3")
op <- par(las = 2)
plot(time ~ year, data = men1500m, axes = FALSE)
yrs <- seq(from = 1896, to = 2004, by = 4)
axis(1, at = yrs, labels = yrs)
axis(2)
box()
par(op)
```

---

meteo *Meteorological Measurements for 11 Years*

Description

Several meteorological measurements for a period between 1920 and 1931.

Usage

```
data("meteo")
```

Format

A data frame with 11 observations on the following 6 variables.

`year` the years.

`rainNovDec` rainfall in November and December (mm).

`temp` average July temperature.

`rainJuly` rainfall in July (mm).

`radiation` radiation in July (millilitres of alcohol).

`yield` average harvest yield (quintals per hectare).

Details

Carry out a principal components analysis of both the covariance matrix and the correlation matrix of the data and compare the results. Which set of components leads to the most meaningful interpretation?

Source

B. S. Everitt and G. Dunn (2001), *Applied Multivariate Data Analysis*, 2nd edition, Arnold, London.

Examples

```
data("meteo", package = "HSAUR3")
meteo
```

orallesions *Oral Lesions in Rural India*

Description

The distribution of the oral lesion site found in house-to-house surveys in three geographic regions of rural India.

Usage

```
data("orallesions")
```

Format

A two-way classification, see `table`.

Source

Cyrus R. Mehta and Nitin R. Patel (2003), *StatXact-6: Statistical Software for Exact Nonparametric Inference*, Cytel Software Cooperation, Cambridge, USA.

Examples

```
data("orallesions", package = "HSAUR3")
mosaicplot(orallesions)
```

phosphate *Phosphate Level Data*

Description

Plasma inorganic phosphate levels from 33 subjects.

Usage

```
data("phosphate")
```

Format

A data frame with 33 observations on the following 9 variables.

`group` a factor with levels `control` and `obese`.

`t0` baseline phosphate level,

`t0.5` phosphate level after 1/2 an hour.

`t1` phosphate level after one an hour.

`t1.5` phosphate level after 1 1/2 hours.

`t2` phosphate level after two hours.

`t3` phosphate level after three hours.

`t4` phosphate level after four hours.

`t5` phosphate level after five hours.

Source

C. S. Davis (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer, New York.

Examples

```
data("phosphate", package = "HSAUR3")
plot(t0 ~ group, data = phosphate)
```

`pistonrings` *Piston Rings Failures*

Description

Number of failures of piston rings in three legs of four steam-driven compressors.

Usage

```
data("pistonrings")
```

Format

A two-way classification, see `table`.

Details

The data are given in form of a `table`. The table gives the number of piston-ring failures in each of three legs of four steam-driven compressors located in the same building. The compressors have identical design and are oriented in the same way. The question of interest is whether the two classification variables (compressor and leg) are independent.

Source

S. J. Haberman (1973), The analysis of residuals in cross-classificed tables. *Biometrics* **29**, 205–220.

Examples

```
data("pistonrings", package = "HSAUR3")
mosaicplot(pistonrings)
```

planets *Exoplanets Data*

Description

Data on planets outside the Solar System.

Usage

```
data("planets")
```

Format

A data frame with 101 observations from 101 exoplanets on the following 3 variables.

mass Jupiter mass of the planet.

period period in earth days.

eccen the radial eccentricity of the planet.

Details

From the properties of the exoplanets found up to now it appears that the theory of planetary development constructed for the planets of the Solar System may need to be reformulated. The exoplanets are not at all like the nine local planets that we know so well. A first step in the process of understanding the exoplanets might be to try to classify them with respect to their known properties.

Source

M. Mayor and P. Frei (2003). *New Worlds in the Cosmos: The Discovery of Exoplanets*. Cambridge University Press, Cambridge, UK.

Examples

```
data("planets", package = "HSAUR3")
require("scatterplot3d")
scatterplot3d(log(planets$mass), log(planets$period), log(planets$eccen),
              type = "h", highlight.3d = TRUE,  angle = 55,
              scale.y = 0.7, pch = 16)
```

plasma *Blood Screening Data*

Description

The erythrocyte sedimentation rate and measurements of two plasma proteins (fibrinogen and globulin).

Usage

```
data("plasma")
```

Format

A data frame with 32 observations on the following 3 variables.

`fibrinogen` the fibrinogen level in the blood.

`globulin` the globulin level in the blood.

`ESR` the erythrocyte sedimentation rate, either less or greater 20 mm / hour.

Details

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If the ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as rheumatic diseases, chronic infections and malignant diseases, its determination might be useful in screening blood samples taken form people suspected to being suffering from one of the conditions mentioned. The absolute value of the ESR is not of great importance rather it is whether it is less than 20mm/hr since lower values indicate a healthy individual.

The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

Source

D. Collett and A. A. Jemain (1985), Residuals, outliers and influential observations in regression analysis. *Sains Malaysiana*, **4**, 493–511.

Examples

```
data("plasma", package = "HSAUR3")
layout(matrix(1:2, ncol = 2))
boxplot(fibrinogen ~ ESR, data = plasma, varwidth = TRUE)
boxplot(globulin ~ ESR, data = plasma, varwidth = TRUE)
```

polyps *Familial Andenomatous Polyposis*

Description

Data from a placebo-controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis (FAP).

Usage

```
data("polyps")
```

Format

A data frame with 20 observations on the following 3 variables.

`number` number of colonic polyps at 12 months.

`treat` treatment arms of the trail, a factor with levels `placebo` and `drug`.

`age` the age of the patient.

Details

Giardiello et al. (1993) and Piantadosi (1997) describe the results of a placebo-controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis (FAP). The trial was halted after a planned interim analysis had suggested compelling evidence in favour of the treatment. Here we are interested in assessing whether the number of colonic polyps at 12 months is related to treatment and age of patient.

Source

F. M. Giardiello, S. R. Hamilton, A. J. Krush, S. Piantadosi, L. M. Hylind, P. Celano, S. V. Booker, C. R. Robinson and G. J. A. Offerhaus (1993), Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *New England Journal of Medicine*, **328**(18), 1313–1316.

S. Piantadosi (1997), *Clinical Trials: A Methodologic Perspective*. John Wiley & Sons, New York.

Examples

```
data("polyps", package = "HSAUR3")
plot(number ~ age, data = polyps, pch = as.numeric(polyps$treat))
legend(40, 40, legend = levels(polyps$treat), pch = 1:2, bty = "n")
```

polyps3 *Familial Andenomatous Polyposis*

Description

Data from a placebo-controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis (FAP).

Usage

```
data("polyps3")
```

Format

A data frame with 22 observations on the following 5 variables.

`gender` a factor with levels `female` and `male`.

`treatment` a factor with levels `placebo` and `active`.

`baseline` the baseline number of polyps.

`age` the age of the patient.

`number3m` the number of polyps after three month.

Details

The data arise from the same study as the `polyps` data. Here, the number of polyps after three months are given.

Source

F. M. Giardiello, S. R. Hamilton, A. J. Krush, S. Piantadosi, L. M. Hylind, P. Celano, S. V. Booker, C. R. Robinson and G. J. A. Offerhaus (1993), Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *New England Journal of Medicine*, **328**(18), 1313–1316.

S. Piantadosi (1997), *Clinical Trials: A Methodologic Perspective*. John Wiley & Sons, New York.

Examples

```
data("polyps3", package = "HSAUR3")
plot(number3m ~ age, data = polyps3, pch = as.numeric(polyps3$treatment))
legend("topright", legend = levels(polyps3$treatment), pch = 1:2, bty = "n")
```

pottery *Romano-British Pottery Data*

Description

Chemical composition of Romano-British pottery.

Usage

```
data("pottery")
```

Format

A data frame with 45 observations on the following 9 chemicals.

Al2O3 aluminium trioxide.

Fe2O3 iron trioxide.

MgO magnesium oxide.

CaO calcium oxide.

Na2O natrium oxide.

K2O calium oxide.

TiO2 titanium oxide.

MnO mangan oxide.

BaO barium oxide.

kiln site at which the pottery was found.

Details

The data gives the chemical composition of specimens of Romano-British pottery, determined by atomic absorption spectrophotometry, for nine oxides.

Source

A. Tubb and N. J. Parker and G. Nickless (1980), The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, **22**, 153–171.

Examples

```
data("pottery", package = "HSAUR3")
plot(pottery)
```

rearrests *Rearrests of Juvenile Felons*

Description

Rearrests of juventile felons by type of court in which they were tried.

Usage

```
data("rearrests")
```

Format

A two-way classification, see `table`.

Details

The data (taken from Agresti, 1996) arise from a sample of juveniles convicted of felony in Florida in 1987. Matched pairs were formed using criteria such as age and the number of previous offences. For each pair, one subject was handled in the juvenile court and the other was transferred to the adult court. Whether or not the juvenile was rearrested by the end of 1988 was then noted. Here the question of interest is whether the true proportions rearrested were identical for the adult and juvenile court assignments?

Source

A. Agresti (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Examples

```
data("rearrests", package = "HSAUR3")
rearrests
```

respiratory *Respiratory Illness Data*

Description

The respiratory status of patients recruited for a randomised clinical multicenter trial.

Usage

```
data("respiratory")
```

Format

A data frame with 555 observations on the following 7 variables.

`centre` the study center, a factor with levels `1` and `2`.

`treatment` the treatment arm, a factor with levels `placebo` and `treatment`.

`gender` a factor with levels `female` and `male`.

`age` the age of the patient.

`status` the respiratory status (response variable), a factor with levels `poor` and `good`.

`month` the month, each patient was examined at months `0`, `1`, `2`, `3` and `4`.

`subject` the patient ID, a factor with levels `1` to `111`.

Details

In each of two centres, eligible patients were randomly assigned to active treatment or placebo. During the treatment, the respiratory status (categorised `poor` or `good`) was determined at each of four, monthly visits. The trial recruited 111 participants (54 in the active group, 57 in the placebo group) and there were no missing data for either the responses or the covariates. The question of interest is to assess whether the treatment is effective and to estimate its effect.

Note that the data are in long form, i.e, repeated measurments are stored as additional rows in the data frame.

Source

C. S. Davis (1991), Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, **10**, 1959–1980.

Examples

```
data("respiratory", package = "HSAUR3")
mosaicplot(xtabs( ~ treatment + month + status, data = respiratory))
```

---

roomwidth *Students Estimates of Lecture Room Width*

Description

Lecture room width estimated by students in two different units.

Usage

```
data("roomwidth")
```

Format

A data frame with 113 observations on the following 2 variables.

unit a factor with levels `feet` and `metres`.

width the estimated width of the lecture room.

Details

Shortly after metric units of length were officially introduced in Australia, each of a group of 44 students was asked to guess, to the nearest metre, the width of the lecture hall in which they were sitting. Another group of 69 students in the same room was asked to guess the width in feet, to the nearest foot. The data were collected by Professor T. Lewis and are taken from Hand et al (1994). The main question is whether estimation in feet and in metres gives different results.

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.

Examples

```
data("roomwidth", package = "HSAUR3")
convert <- ifelse(roomwidth$unit == "feet", 1, 3.28)
boxplot(I(width * convert) ~ unit, data = roomwidth)
```

---

schizophrenia *Age of Onset of Schizophrenia Data*

Description

Data on sex differences in the age of onset of schizophrenia.

Usage

```
data("schizophrenia")
```

Format

A data frame with 251 observations on the following 2 variables.

`age` age at the time of diagnosis.

`gender` a factor with levels `female` and `male`

Details

A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequently epidemiological studies of the disorder have consistently shown an earlier onset in men than in women. One model that has been suggested to explain this observed difference is know as the subtype model which postulates two type of schizophrenia, one characterised by early onset, typical symptoms and poor premorbid competence, and the other by late onset, atypical symptoms, and good premorbid competence. The early onset type is assumed to be largely a disorder of men and the late onset largely a disorder of women.

Source

E. Kraepelin (1919), *Dementia Praecox and Paraphrenia*. Livingstone, Edinburgh.

Examples

```
data("schizophrenia", package = "HSAUR3")
boxplot(age ~ gender, data = schizophrenia)
```

`schizophrenia2` *Schizophrenia Data*

Description

Though disorder and early onset of schizophrenia.

Usage

```
data("schizophrenia2")
```

Format

A data frame with 220 observations on the following 4 variables.

`subject` the patient ID, a factor with levels `1` to `44`.

`onset` the time of onset of the disease, a factor with levels `< 20 yrs` and `> 20 yrs`.

`disorder` whether thought disorder was `absent` or `present`, the response variable.

`month` month after hospitalisation.

Details

The data were collected in a follow-up study of women patients with schizophrenia. The binary response recorded at 0, 2, 6, 8 and 10 months after hospitalisation was thought disorder (absent or present). The single covariate is the factor indicating whether a patient had suffered early or late onset of her condition (age of onset less than 20 years or age of onset 20 years or above). The question of interest is whether the course of the illness differs between patients with early and late onset?

Source

Davis (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer, New York.

Examples

```
data("schizophrenia2", package = "HSAUR3")
mosaicplot(xtabs( ~ onset + month + disorder, data = schizophrenia2))
```

`schooldays` *Days not Spent at School*

Description

Data from a sociological study, the number of days absent from school is the response variable.

Usage

```
data("schooldays")
```

Format

A data frame with 154 observations on the following 5 variables.

`race` race of the child, a factor with levels `aboriginal` and `non-aboriginal`.

`gender` the gender of the child, a factor with levels `female` and `male`.

`school` the school type, a factor with levels `F0` (primary), `F1` (first), `F2` (second) and `F3` (third form).

`learner` how good is the child in learning things, a factor with levels `average` and `slow`.

`absent` number of days absent from school.

Details

The data arise from a sociological study of Australian Aboriginal and white children reported by Quine (1975).

In this study, children of both sexes from four age groups (final grade in primary schools and first, second and third form in secondary school) and from two cultural groups were used. The children in age group were classified as slow or average learners. The response variable was the number of days absent from school during the school year. (Children who had suffered a serious illness during the years were excluded.)

Source

S. Quine (1975), Achievement Orientation of Aboriginal and White Adolescents. Doctoral Dissertation, Australian National University, Canberra.

Examples

```
data("schooldays", package = "HSAUR3")
plot.design(schooldays)
```

---

skulls *Egyptian Skulls*

Description

Measurements made on Egyptian skulls from five epochs.

Usage

```
data("skulls")
```

Format

A data frame with 150 observations on the following 5 variables.

`epoch` the epoch the skull as assigned to, a factor with levels `c4000BC` `c3300BC`, `c1850BC`, `c200BC`, and `cAD150`, where the years are only given approximately, of course.

`mb` maximum breaths of the skull.

`bh` basibregmatic heights of the skull.

`bl` basialiveolar length of the skull.

`nh` nasal heights of the skull.

Details

The question is whether the measurements change over time. Non-constant measurements of the skulls over time would indicate interbreeding with immigrant populations.

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.

Examples

```
data("skulls", package = "HSAUR3")
means <- tapply(1:nrow(skulls), skulls$epoch, function(i)
             apply(skulls[i,colnames(skulls)[-1]], 2, mean))
means <- matrix(unlist(means), nrow = length(means), byrow = TRUE)
colnames(means) <- colnames(skulls)[-1]
rownames(means) <- levels(skulls$epoch)
pairs(means,
    panel = function(x, y) {
```

```
    text(x, y, levels(skulls$epoch))
})
```

---

smoking *Nicotine Gum and Smoking Cessation*

Description

Data from a meta-analysis on nicotine gum and smoking cessation

Usage

```
data("smoking")
```

Format

A data frame with 26 observations (studies) on the following 4 variables.

`qt` the number of treated subjetcs who stopped smoking.

`tt` the totla number of treated subjects.

`qc` the number of subjetcs who stopped smoking without being treated.

`tc` the total number of subject not being treated.

Details

Cigarette smoking is the leading cause of preventable death in the United States and kills more Americans than AIDS, alcohol, illegal drug use, car accidents, fires, murders and suicides combined. It has been estimated that 430,000 Americans die from smoking every year. Fighting tobacco use is, consequently, one of the major public health goals of our time and there are now many programs available designed to help smokers quit. One of the major aids used in these programs is nicotine chewing gum, which acts as a substitute oral activity and provides a source of nicotine that reduces the withdrawal symptoms experienced when smoking is stopped. But separate randomized clinical trials of nicotine gum have been largely inconclusive, leading Silagy (2003) to consider combining the results studies found from an extensive literature search. The results of these trials in terms of numbers of people in the treatment arm and the control arm who stopped smoking for at least 6 months after treatment are given here.

Source

C. Silagy (2003), Nicotine replacement therapy for smoking cessation (Cochrane Review). *The Cochrane Library*, **4**, John Wiley & Sons, Chichester.

Examples

```
data("smoking", package = "HSAUR3")
boxplot(smoking$qt/smoking$tt,
        smoking$qc/smoking$tc,
        names = c("Treated", "Control"), ylab = "Percent Quitters")
```

Smoking_DollHill1950 *Smoking and Lung Cancer (IV)*

Description

Number of smokers in a case-control study.

Usage

```
data(Smoking_DollHill1950)
```

Format

The format is: table [1:6, 1:2, 1:2] 2 33 250 196 136 32 27 55 293 190 ... - attr(*, "dimnames")=List of 3 ..$ Smoking : chr [1:6] "Nonsmoker" "1-" "5-" "15-" ... ..$ Diagnosis: chr [1:2] "Lung cancer" "Other" ..$ Sex : chr [1:2] "Male" "Female"

Details

This is Table V from Doll and Hill (1950).

Source

Richard Doll and A. Bradford Hill (1950), Smoking and Carcinoma of the Lung, *British Medical Journal*, **2**, 739-748

References

Richard Doll (1998), Uncovering the effects of smoking: historical perspective. *Statistical Methods in Medical Research*, **7**(87), 87-117

Smoking_Mueller1940 *Smoking and Lung Cancer (I)*

Description

Number of smokers in a case-control study.

Usage

```
data(Smoking_Mueller1940)
```

Format

The format is: table [1:5, 1:2] 25 18 13 27 3 4 5 22 41 14 - attr(*, "dimnames")=List of 2 ..$ Smoking_type: chr [1:5] "Extreme smoker" "Very heavy smoker" "Heavy smoker" "Moderate smoker" ... ..$ Group : chr [1:2] "Lung cancer" "Healthy control"

Details

Extreme smoker: 10-15 cigars, >35 cigarettes, or >50 g pipe tobacco/day. Very heavy smoker: 7-9 cigars, 26-35 cigarettes, or 36-50 g pipe tobacco/day. Heavy smoker: 4-6 cigars, 16-25 cigarettes, or 21-35 g pipe tobacco/day. Moderate smoker: 1-3 cigars, 1-15 cigarettes, or 1-20 g pipe tobacco/day.

Source

Franz-Hermann Mueller (1940), Tabakmissbrauch und Lungencarcinom. *Zeitschrift fuer Krebsforschung* **49**(1), 57-85.

References

Richard Doll (1998), Uncovering the effects of smoking: historical perspective. *Statistical Methods in Medical Research*, **7**(87), 87-117

Examples

```
data(Smoking_Mueller1940)
## maybe str(Smoking_Mueller1940) ; plot(Smoking_Mueller1940) ...
```

Smoking_SchairerSchoeniger1944

Smoking and Lung Cancer (II)

Description

Number of smokers in a case-control study.

Usage

```
data(Smoking_SchairerSchoeniger1944)
```

Format

The format is: table [1:5, 1:7] 3 11 31 19 29 2 0 4 6 3 ... - attr(*, "dimnames")=List of 2 ..$ Smoking : chr [1:5] "Nonsmoker" "Moderate smoker" "Medium smoker" "Heavy smoker" ... ..$ Diagnosis: chr [1:7] "Lung cancer" "Lip cancer" "Throat cancer" "Stomach cancer" ...

Source

E. Schairer and E. Sch\"oninger (1944), Lungenkrebs und Tabakverbrauch, *Zeitschrift fuer Krebsforschung*, **54**(4), 261-269

References

Richard Doll (1998), Uncovering the effects of smoking: historical perspective. *Statistical Methods in Medical Research*, **7**(87), 87-117

Smoking_Wassink1945 *Smoking and Lung Cancer (III)*

Description

Number of smokers in a case-control study.

Usage

```
data(Smoking_Wassink1945)
```

Format

The format is: table [1:4, 1:2] 6 18 36 74 19 36 25 20 - attr(*, "dimnames")=List of 2 ..$ Smoking : chr [1:4] "Nonsmoker" "Moderate smoker" "Heavy smoker" "Very heavy smoker" ..$ Diagnosis: chr [1:2] "Lung cancer" "Healthy control"

Source

W. F. Wassink (1945), Ontstaansvoorwaarden voor Longkanker, *Nederlands Tijdschrift voor Geneeskunde*, **92**, 3732–3747

References

Richard Doll (1998), Uncovering the effects of smoking: historical perspective. *Statistical Methods in Medical Research*, **7**(87), 87-117

students *Student Risk Taking*

Description

Students were administered two parallel forms of a test after a random assignment to three different treatments.

Usage

```
data("students")
```

Format

A data frame with 35 observations on the following 3 variables.

`treatment` a factor with levels `AA`, `C`, and `NC`.

`low` the result of the first test.

`high` the result of the second test.

Details

The data arise from a large study of risk taking (Timm, 2002). Students were randomly assigned to three different treatments labelled AA, C and NC. Students were administered two parallel forms of a test called `low` and `high`. The aim is to carry out a test of the equality of the bivariate means of each treatment population.

Source

N. H. Timm (2002), *Applied Multivariate Analysis*. Springer, New York.

Examples

```
data("students", package = "HSAUR3")
layout(matrix(1:2, ncol = 2))
boxplot(low ~ treatment, data = students, ylab = "low")
boxplot(high ~ treatment, data = students, ylab = "high")
```

---

suicides *Crowd Baiting Behaviour and Suicides*

Description

Data from a study carried out to investigate the causes of jeering or baiting behaviour by a crowd when a person is threatening to commit suicide by jumping from a high building.

Usage

```
data("suicides")
```

Format

A two-way classification, see `table`.

Source

L. Mann (1981), The baiting crowd in episodes of threatened suicide. *Journal of Personality and Social Psychology*, **41**, 703–709.

Examples

```
data("suicides", package = "HSAUR3")
mosaicplot(suicides)
```

suicides2 *Male Suicides Data*

Description

Number of suicides in different age groups and countries.

Usage

```
data("suicides2")
```

Format

A data frame with 15 observations on the following 5 variables.

`A25.34` number of suicides (per 100000 males) between 25 and 34 years old.

`A35.44` number of suicides (per 100000 males) between 35 and 44 years old.

`A45.54` number of suicides (per 100000 males) between 45 and 54 years old.

`A55.64` number of suicides (per 100000 males) between 55 and 64 years old.

`A65.74` number of suicides (per 100000 males) between 65 and 74 years old.

Details

Each of the numbers gives the number of suicides per 100000 male inhabitants of the countries given by the row names.

toenail *Toenail Infection Data*

Description

Results of a clinical trial to compare two competing oral antifungal treatments for toenail infection.

Usage

```
data("toenail")
```

Format

A data frame with 1908 observations on the following 5 variables.

`patientID` a unique identifier for each patient in the trial.

`outcome` degree of separation of the nail plate from the nail bed (onycholysis).

`treatment` a factor with levels `itraconazole` and `terbinafine`.

`time` the time in month when the visit actually took place.

`visit` number of visit attended.

Details

De Backer et al. (1998) describe a clinical trial to compare two competing oral antifungal treatments for toenail infection (dermatophyte onychomycosis). A total of 378 patients were randomly allocated into two treatment groups, one group receiving 250mg per day of terbinafine and the other group 200mg per day of itraconazole. Patients were evaluated at seven visits, intended to be at weeks 0, 4, 8, 12, 24, 36, and 48 for the degree of separation of the nail plate from the nail bed (onycholysis) dichotomized into `moderate` or `severe` and `none` or `mild`. But patients did not always arrive exactly at the scheduled time and the exact time in months that they did attend was recorded. The data is not balanced since not all patients attended for all seven planned visits.

Source

M. D. Backer and C. D. Vroey and E. Lesaffre and I. Scheys and P. D. Keyser (1998), Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. *Journal of the American Academy of Dermatology*, **38**, S57–S63.

Examples

```
data("toenail", package = "HSAUR3")
```

`toothpaste` *Toothpaste Data*

Description

Meta-analysis of studies comparing two different toothpastes.

Usage

```
data("toothpaste")
```

Format

A data frame with 9 observations on the following 7 variables.

`Study` the identifier of the study.

`nA` number of subjects using toothpaste A.

`meanA` mean DMFS index of subjects using toothpaste A.

`sdA` standard deviation of DMFS index of subjects using toothpaste A.

`nB` number of subjects using toothpaste B.

`meanB` mean DMFS index of subjects using toothpaste B.

`sdB` standard deviation of DMFS index of subjects using toothpaste B.

Details

The data are the results of nine randomised trials comparing two different toothpastes for the prevention of caries development. The outcomes in each trial was the change, from baseline, in the decayed, missing (due to caries) and filled surface dental index (DMFS).

Source

B. S. Everitt and A. Pickles (2000), *Statistical Aspects of the Design and Analysis of Clinical Trials*, Imperial College Press, London.

Examples

```
data("toothpaste", package = "HSAUR3")
toothpaste
```

USairpollution *Air Pollution in US Cities*

Description

Air pollution data of 41 US cities.

Usage

```
data("USairpollution")
```

Format

A data frame with 41 observations on the following 7 variables.

`SO2` SO2 content of air in micrograms per cubic metre.

`temp` average annual temperature in Fahrenheit.

`manu` number of manufacturing enterprises employing 20 or more workers.

`popul` population size (1970 census); in thousands.

`wind` average annual wind speed in miles per hour.

`precip` average annual precipitation in inches.

`predays` average number of days with precipitation per year.

Details

The annual mean concentration of sulphur dioxide, in micrograms per cubic metre, is a measure of the air pollution of the city. The question of interest here is what aspects of climate and human ecology as measured by the other six variables in the data determine pollution?

Source

R. R. Sokal and F. J. Rohlf (1981), *Biometry*, W. H. Freeman, San Francisco (2nd edition).

Examples

```
data("USairpollution", package = "HSAUR3")
```

USmelanoma *USA Malignant Melanoma Data*

Description

USA mortality rates for white males due to malignant melanoma 1950-1969.

Usage

```
data("USmelanoma")
```

Format

A data frame with 48 observations on the following 5 variables.

`mortality` number of white males died due to malignant melanoma 1950-1969 per one million inhabitants.

`latitude` latitude of the geographic centre of the state.

`longitude` longitude of the geographic centre of each state.

`ocean` a binary variable indicating contiguity to an ocean at levels `no` or `yes`.

Details

Fisher and van Belle (1993) report mortality rates due to malignant melanoma of the skin for white males during the period 1950-1969, for each state on the US mainland. Questions of interest about these data include how do the mortality rates compare for ocean and non-ocean states?

Source

Fisher and van Belle (1993)

Examples

```
data("USmelanoma", package = "HSAUR3")
```

USstates *US States*

Description

Socio-demographic variables for ten US states.

Usage

```
data(USstates)
```

Format

A data frame with 10 observations on the following 7 variables.

`Population` population size divided by 1000
`Income` average per capita income
`Illiteracy` illiteracy rate (per cent population)
`Life.Expectancy` life expectancy (years)
`Homicide` homicide rate (per 1000)
`Graduates` percentage of high school graduates
`Freezing` average number of days per below freezing

Details

The data set contains values of seven socio-demographic variables for ten states in the USA.

UStemp *Temperatures in 22 US cities*

Description

Lowest temperatures in Fahrenheit in 22 US cities in four months.

Usage

```
data(UStemp)
```

Format

A data frame with 22 observations on the following 4 variables.

`January` lowest temperature in Fahrenheit
`April` lowest temperature in Fahrenheit
`July` lowest temperature in Fahrenheit
`October` lowest temperature in Fahrenheit

voting *House of Representatives Voting Data*

Description

Voting results for 15 congressmen from New Jersey.

Usage

```
data("voting")
```

Format

A 15 times 15 matrix.

Details

Romesburg (1984) gives a set of data that shows the number of times 15 congressmen from New Jersey voted differently in the House of Representatives on 19 environmental bills. Abstentions are not recorded.

Source

H. C. Romesburg (1984), *Cluster Analysis for Researchers*. Lifetime Learning Publications, Belmont, Canada.

Examples

```
data("voting", package = "HSAUR3")
require("MASS")
voting_mds <- isoMDS(voting)
plot(voting_mds$points[,1], voting_mds$points[,2],
     type = "n", xlab = "Coordinate 1", ylab = "Coordinate 2",
     xlim = range(voting_mds$points[,1])*1.2)
text(voting_mds$points[,1], voting_mds$points[,2],
     labels = colnames(voting))
voting_sh <- Shepard(voting[lower.tri(voting)], voting_mds$points)
```

water *Mortality and Water Hardness*

Description

The mortality and drinking water hardness for 61 cities in England and Wales.

Usage

```
data("water")
```

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 north as Derby.

town the name of the town.

mortality averaged annual mortality per 100.000 male inhabitants.

hardness calcium concentration (in parts per million).

Details

The data were collected in an investigation of environmental causes of disease. They show the annual mortality per 100,000 for males, averaged over the years 1958-1964, and the calcium concentration (in parts per million) in the drinking water for 61 large towns in England and Wales. The higher the calcium concentration, the harder the water. Towns at least as far north as Derby are identified in the table. Here there are several questions that might be of interest including, are mortality and water hardness related, and do either or both variables differ between northern and southern towns?

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.

Examples

```
data("water", package = "HSAUR3")
plot(mortality ~ hardness, data = water,
     col = as.numeric(water$location))
```

watervoles *Water Voles Data*

Description

Percentage incidence of the 13 characteristics of water voles in 14 areas.

Usage

```
data("watervoles")
```

Format

A dissimilarity matrix for the following 14 variables, i.e, areas: `Surrey`, `Shropshire`, `Yorkshire`, `Perthshire`, `Aberdeen`, `Elean Gamhna`, `Alps`, `Yugoslavia`, `Germany`, `Norway`, `Pyrenees I`, `Pyrenees II`, `North Spain`, and `South Spain`.

Details

Corbet et al. (1970) report a study of water voles (genus Arvicola) in which the aim was to compare British populations of these animals with those in Europe, to investigate whether more than one species might be present in Britain. The original data consisted of observations of the presence or absence of 13 characteristics in about 300 water vole skulls arising from six British populations and eight populations from the rest of Europe. The data are the percentage incidence of the 13 characteristics in each of the 14 samples of water vole skulls.

Source

G. B. Corbet, J. Cummins, S. R. Hedges, W. J. Krzanowski (1970), The taxonomic structure of British water voles, genus *Arvicola*. *Journal of Zoology*, **61**, 301–316.

Examples

```
data("watervoles", package = "HSAUR3")
watervoles
```

waves *Electricity from Wave Power at Sea*

Description

Measurements of root mean square bending moment by two different mooring methods.

Usage

```
data("waves")
```

Format

A data frame with 18 observations on the following 2 variables.

method1 Root mean square bending moment in Newton metres, mooring method 1

method2 Root mean square bending moment in Newton metres, mooring method 2

Details

In a design study for a device to generate electricity from wave power at sea, experiments were carried out on scale models in a wave tank to establish how the choice of mooring method for the system affected the bending stress produced in part of the device. The wave tank could simulate a wide range of sea states and the model system was subjected to the same sample of sea states with each of two mooring methods, one of which was considerably cheaper than the other. The question of interest is whether bending stress differs for the two mooring methods.

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.

Examples

```
data("waves", package = "HSAUR3")
plot(method1 ~ method2, data = waves)
```

---

weightgain *Gain in Weight of Rats*

Description

The data arise 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

```
data("weightgain")
```

Format

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

`source` source of protein given, a factor with levels `Beef` and `Cereal`.

`type` amount of protein given, a factor with levels `High` and `Low`.

`weightgain` weigt gain in grams.

Details

Ten rats are randomized to each of the four treatments. The question of interest is how diet affects weight gain.

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.

Examples

```
data("weightgain", package = "HSAUR3")
interaction.plot(weightgain$type, weightgain$source,
                 weightgain$weightgain)
```

`womensrole` *Womens Role in Society*

Description

Data from a survey from 1974 / 1975 asking both female and male responders about their opinion on the statement: Women should take care of running their homes and leave running the country up to men.

Usage

```
data("womensrole")
```

Format

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

`education` years of education.

`gender` a factor with levels `Male` and `Female`.

`agree` number of subjects in agreement with the statement.

`disagree` number of subjects in disagreement with the statement.

Details

The data are from Haberman (1973) and also given in Collett (2003). The questions here are whether the response of men and women differ.

Source

S. J. Haberman (1973), The analysis of residuals in cross-classificed tables. *Biometrics*, **29**, 205–220.

D. Collett (2003), *Modelling Binary Data*. Chapman and Hall / CRC, London. 2nd edition.

Examples

```
data("womensrole", package = "HSAUR3")
summary(subset(womensrole, gender == "Female"))
summary(subset(womensrole, gender == "Male"))
```

Index

∗ **datasets**
agefat, 3
aspirin, 4
backpain, 5
BCG, 6
birds, 7
birthdeathrates, 8
bladdercancer, 8
bp, 9
BtheB, 10
CHFLS, 11
clouds, 13
CYGOB1, 15
EFT, 16
epilepsy, 16
Forbes2000, 17
foster, 18
gardenflowers, 19
GHQ, 19
heptathlon, 20
household, 21
Lanza, 23
mastectomy, 24
men1500m, 25
meteo, 26
orallesions, 27
phosphate, 27
pistonrings, 28
planets, 29
plasma, 30
polyps, 31
polyps3, 32
pottery, 33
rearrests, 34
respiratory, 34
roomwidth, 35
schizophrenia, 36
schizophrenia2, 37
schooldays, 38
skulls, 39
smoking, 40
Smoking_DollHill1950, 41
Smoking_Mueller1940, 41
Smoking_SchairerSchoeniger1944, 42
Smoking_Wassink1945, 43
students, 43
suicides, 44
suicides2, 45
toenail, 45
toothpaste, 46
USairpollution, 47
USmelanoma, 48
USstates, 49
USTemp, 49
voting, 50
water, 51
watervoles, 52
waves, 52
weightgain, 53
womensrole, 54
∗ **misc**
HSAURtable, 22

agefat, 3
aspirin, 4

backpain, 5
BCG, 6
birds, 7
birthdeathrates, 8
bladdercancer, 8
bp, 9
BtheB, 10

CHFLS, 11
clouds, 13
CYGOB1, 15

dist, *19*

EFT, 16
epilepsy, 16

Forbes2000, 17
foster, 18

gardenflowers, 19
GHQ, 19

heptathlon, 20
household, 21
HSAURtable, 22

Lanza, 23

mastectomy, 24
men1500m, 25
meteo, 26

orallesions, 27

phosphate, 27
pistonrings, 28
planets, 29
plasma, 30
polyps, 31, *32*
polyps3, 32
pottery, 33

rearrests, 34
respiratory, 34
roomwidth, 35

schizophrenia, 36
schizophrenia2, 37
schooldays, 38
skulls, 39
smoking, 40
Smoking_DollHill1950, 41
Smoking_Mueller1940, 41
Smoking_SchairerSchoeniger1944, 42
Smoking_Wassink1945, 43
students, 43
suicides, 44
suicides2, 45

table, *27*, *28*, *34*, *44*
toenail, 45
toLatex, *23*
toLatex.dftab *(*HSAURtable*)*, 22
toLatex.tabtab *(*HSAURtable*)*, 22
toothpaste, 46

USairpollution, 47
USmelanoma, 48
USstates, 49
UStemp, 49

voting, 50

water, 51
watervoles, 52
waves, 52
weightgain, 53
womensrole, 54