Package ‘UsingR’

February 19, 2015

Version 2.0-4
Title Data Sets, etc. for the Text "Using R for Introductory Statistics", Second Edition
Author John Verzani <verzani@math.csi.cuny.edu>
Maintainer John Verzani <verzani@math.csi.cuny.edu>
Depends R (>= 2.15.0), MASS, HistData, Hmisc
Suggests zoo, coin, ggplot2, vcd, lubridate, aplpack
License GPL (>= 2)
LazyData TRUE
NeedsCompilation no
Repository CRAN
Date/Publication 2015-01-22 08:35:41

R topics documented:

- age.universe ........................................ 5
- aid ............................................. 6
- alaska.pipeline .................................. 6
- alltime.movies .................................. 7
- answers ........................................... 8
- aosat ............................................. 8
- arctic.oscillations .............................. 9
- babies .......................................... 10
- babyboom ....................................... 11
- batting .......................................... 12
- baycheck ........................................ 13
- best.times ....................................... 14
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood</td>
<td>15</td>
</tr>
<tr>
<td>breakdown</td>
<td>16</td>
</tr>
<tr>
<td>bright.stars</td>
<td>16</td>
</tr>
<tr>
<td>brightness</td>
<td>17</td>
</tr>
<tr>
<td>bumpers</td>
<td>18</td>
</tr>
<tr>
<td>BushApproval</td>
<td>18</td>
</tr>
<tr>
<td>bycatch</td>
<td>19</td>
</tr>
<tr>
<td>cabinet</td>
<td>20</td>
</tr>
<tr>
<td>camp</td>
<td>21</td>
</tr>
<tr>
<td>cancer</td>
<td>22</td>
</tr>
<tr>
<td>carbon</td>
<td>23</td>
</tr>
<tr>
<td>carsafety</td>
<td>23</td>
</tr>
<tr>
<td>central.park</td>
<td>24</td>
</tr>
<tr>
<td>central.park.cloud</td>
<td>25</td>
</tr>
<tr>
<td>ceo2013</td>
<td>26</td>
</tr>
<tr>
<td>cfb</td>
<td>26</td>
</tr>
<tr>
<td>chicken</td>
<td>27</td>
</tr>
<tr>
<td>chips</td>
<td>28</td>
</tr>
<tr>
<td>co2emiss</td>
<td>29</td>
</tr>
<tr>
<td>coins</td>
<td>29</td>
</tr>
<tr>
<td>coldvermont</td>
<td>30</td>
</tr>
<tr>
<td>confint.htest</td>
<td>31</td>
</tr>
<tr>
<td>corn</td>
<td>31</td>
</tr>
<tr>
<td>crime</td>
<td>32</td>
</tr>
<tr>
<td>deflection</td>
<td>33</td>
</tr>
<tr>
<td>demos</td>
<td>33</td>
</tr>
<tr>
<td>DensityPlot</td>
<td>34</td>
</tr>
<tr>
<td>diamond</td>
<td>35</td>
</tr>
<tr>
<td>divorce</td>
<td>36</td>
</tr>
<tr>
<td>DOTplot</td>
<td>36</td>
</tr>
<tr>
<td>dottodot</td>
<td>37</td>
</tr>
<tr>
<td>dowdata</td>
<td>38</td>
</tr>
<tr>
<td>dvdsales</td>
<td>39</td>
</tr>
<tr>
<td>emissions</td>
<td>39</td>
</tr>
<tr>
<td>ewr</td>
<td>40</td>
</tr>
<tr>
<td>exec.pay</td>
<td>41</td>
</tr>
<tr>
<td>fat</td>
<td>41</td>
</tr>
<tr>
<td>father.son</td>
<td>43</td>
</tr>
<tr>
<td>female.inc</td>
<td>44</td>
</tr>
<tr>
<td>firstchi</td>
<td>45</td>
</tr>
<tr>
<td>five.yr.temperature</td>
<td>45</td>
</tr>
<tr>
<td>florida</td>
<td>46</td>
</tr>
<tr>
<td>galileo</td>
<td>47</td>
</tr>
<tr>
<td>galton</td>
<td>48</td>
</tr>
<tr>
<td>gap</td>
<td>49</td>
</tr>
<tr>
<td>gasprices</td>
<td>49</td>
</tr>
<tr>
<td>getAnswer</td>
<td>50</td>
</tr>
<tr>
<td>goalspergame</td>
<td>51</td>
</tr>
</tbody>
</table>
topics documented:

- google ........................................... 51
- grades ........................................... 52
- grip .............................................. 52
- hall.fame ....................................... 53
- headtail ........................................ 54
- healthy ......................................... 55
- heartate ........................................ 56
- home ............................................. 56
- homedata ....................................... 57
- homeprice ..................................... 58
- homework ...................................... 59
- HUMMER ....................................... 59
- income_percentiles .............................. 60
- iq ................................................ 60
- kid.weights .................................... 61
- KSI ............................................... 62
- last.tie ......................................... 62
- lawsuits ........................................ 63
- lorem .......................................... 64
- malpract ........................................ 64
- mandms ......................................... 65
- math ............................................. 66
- maydow ......................................... 66
- Medicare ...................................... 67
- midsize ......................................... 68
- MLBattend ...................................... 68
- movies ......................................... 68
- movie_data_2011 ............................... 70
- mw.ages ........................................ 70
- nba.draft ...................................... 71
- nisdc ............................................ 72
- normtemp ....................................... 72
- npdb ............................................. 73
- nym.2002 ....................................... 74
- ObamaApproval .................................. 75
- OBP ............................................. 76
- oral.lesion ..................................... 76
- ozonemonthly ................................... 77
- paradise ....................................... 78
- pi2000 .......................................... 78
- primes ........................................... 79
- puerto .......................................... 79
- QQplot .......................................... 80
- rat ............................................... 81
- reaction.time ................................... 82
- reddrum ......................................... 83
- salmon.rate ..................................... 83
- salmonharvest .................................. 84
<table>
<thead>
<tr>
<th>R topics documented:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>samhda</td>
<td>85</td>
</tr>
<tr>
<td>SAT</td>
<td>86</td>
</tr>
<tr>
<td>scatter.with.hist</td>
<td>87</td>
</tr>
<tr>
<td>scrabble</td>
<td>88</td>
</tr>
<tr>
<td>simple.chutes</td>
<td>89</td>
</tr>
<tr>
<td>simple.densityplot</td>
<td>90</td>
</tr>
<tr>
<td>simple.eda</td>
<td>91</td>
</tr>
<tr>
<td>simple.eda.ts</td>
<td>92</td>
</tr>
<tr>
<td>simple.fancy.stripchart</td>
<td>93</td>
</tr>
<tr>
<td>simple.freqpoly</td>
<td>93</td>
</tr>
<tr>
<td>simple.hist.and.boxplot</td>
<td>94</td>
</tr>
<tr>
<td>simple.lag</td>
<td>95</td>
</tr>
<tr>
<td>simple.lm</td>
<td>96</td>
</tr>
<tr>
<td>simple.median.test</td>
<td>97</td>
</tr>
<tr>
<td>simple.scatterplot</td>
<td>98</td>
</tr>
<tr>
<td>simple.sim</td>
<td>99</td>
</tr>
<tr>
<td>simple.violinplot</td>
<td>100</td>
</tr>
<tr>
<td>simple.z.test</td>
<td>101</td>
</tr>
<tr>
<td>skateranks</td>
<td>102</td>
</tr>
<tr>
<td>slc</td>
<td>103</td>
</tr>
<tr>
<td>smokyph</td>
<td>103</td>
</tr>
<tr>
<td>snacks</td>
<td>104</td>
</tr>
<tr>
<td>south</td>
<td>104</td>
</tr>
<tr>
<td>southernosc</td>
<td>105</td>
</tr>
<tr>
<td>sp500.excess</td>
<td>106</td>
</tr>
<tr>
<td>Split.zoo</td>
<td>106</td>
</tr>
<tr>
<td>squareplot</td>
<td>107</td>
</tr>
<tr>
<td>stud.recs</td>
<td>108</td>
</tr>
<tr>
<td>student.expenses</td>
<td>109</td>
</tr>
<tr>
<td>superbarpplot</td>
<td>109</td>
</tr>
<tr>
<td>tastesgreat</td>
<td>111</td>
</tr>
<tr>
<td>tcm1y</td>
<td>111</td>
</tr>
<tr>
<td>tempsalinity</td>
<td>112</td>
</tr>
<tr>
<td>too.young</td>
<td>113</td>
</tr>
<tr>
<td>twins</td>
<td>113</td>
</tr>
<tr>
<td>u2</td>
<td>114</td>
</tr>
<tr>
<td>urchin.growth</td>
<td>115</td>
</tr>
<tr>
<td>vacation</td>
<td>115</td>
</tr>
<tr>
<td>violinplot</td>
<td>116</td>
</tr>
<tr>
<td>watertemp</td>
<td>117</td>
</tr>
<tr>
<td>wchomes</td>
<td>118</td>
</tr>
<tr>
<td>wellbeing</td>
<td>118</td>
</tr>
<tr>
<td>yahoo.get.hist.quote</td>
<td>119</td>
</tr>
<tr>
<td>yellowfin</td>
<td>120</td>
</tr>
</tbody>
</table>

**Index**

122
Description

For years people have tried to estimate the age of the universe. This data set collects a few estimates starting with lower bounds using estimates for the earth’s age.

Usage

data(age.universe)

Format

A data frame with 16 observations on the following 4 variables.

- `lower` a numeric vector
- `upper` a numeric vector
- `year` a numeric vector
- `source` Short description of source

Details

In the last two decades estimates for the age of the universe have been greatly improved. As of 2013, the best guess is 13.7 billion years with a margin of error of 1 percent. This last estimate is found by WMAP using microwave background radiation. Previous estimates were also based on estimates of Hubble’s constant, and dating of old stars.

Source


Examples

data(age.universe)
$n <- nrow(age.universe)$
$x <- 1:n$

names(x) = age.universe$year

plot(x,age.universe$upper,ylim=c(0,20))
points(x,age.universe$lower)

with(age.universe,sapply(x,function(i) lines(c(i,i),c(lower[i],upper[i]))))
Description
monthly payment for federal program

Usage
data(aid)

Format
The format is: Named num [1:51] 57.2 253.5 114.2 68.2 199.6 ... - attr(*, "names")= chr [1:51]
"Alabama" "Alaska" "Arizona" "Arkansas" ...

Source
From Kitchen's Exploring Statistics

Examples
data(aid)
hist(aid)

alaska.pipeline

Description
Comparison of in-field and laboratory measurement of defects

The Alaska pipeline data consists of in-field ultrasonic measurements of the depths of defects in the Alaska pipeline. The depth of the defects were then re-measured in the laboratory. These measurements were performed in six different batches.

Usage
data(alaska.pipeline)

Format
A data frame with 107 observations on the following 3 variables.

field.defect  Depth of defect as measured in field
lab.defect    Depth of defect as measured in lab
batch        One of 6 batches
Source


Examples

data(alaska.pipeline)
res = lm(lab.defect ~ field.defect, alaska.pipeline)
plot(lab.defect ~ field.defect, alaska.pipeline)
abline(res)
plot(fitted(res), resid(res))

```
alltime.movies  Top movies of all time
```

Description

The top 79 all-time movies as of 2003 by domestic (US) gross receipts.

Usage

data(alltime.movies)

Format

A data frame with 79 observations on the following 2 variables.

Gross a numeric vector
Release.Year a numeric vector

The row names are the titles of the movies.

Source

This data was found on [http://movieweb.com/movie/alltime.html](http://movieweb.com/movie/alltime.html) on June 17, 2003. The source of the data is attributed to (partially) Exhibitor Relations Co.

Examples

data(alltime.movies)
hist(alltime.movies$Gross)
---

## answers

### Description

Opens pdf file containing answers to selected problems

### Usage

```r
appliedR
answers()
```

### Value

Called for its side-effect of opening a pdf

### Examples

```r
# answers()
```

---

## aosat

### Description

Artic Oscillation data based on SAT data


### Usage

```r
data(aosat)
```

### Format

The format is: first column is date in years with fraction to indicate month. The second column is the measurement.

### Details

See [http://jisao.washington.edu/ao/](http://jisao.washington.edu/ao/) for more details on the importance of this time series.

### Source

This data came from the file AO\_SATindex\_JFM\_Jan1851March1997.ascii at [http://www.atmos.colostate.edu/ao/Data/ao_index.html](http://www.atmos.colostate.edu/ao/Data/ao_index.html)
arctic.oscillations

Examples

data(aosat)
## Not run:
library(zoo)
z = zoo(aosat[,2], order.by=aosat[,1])
plot(z)
## yearly
plot(aggregate(z, floor(index(z)), mean))
## decade-long means
plot(aggregate(z, 10*floor(index(z)/10), mean))

## End(Not run)

---

arctic.oscillations  Measurement of sea-level pressure at the arctic

Description

A monthly time series from January 1899 to June 2002 of sea-level pressure measurements relative to some baseline.

Usage

data(arctic.oscillations)

Format

The format is: chr "arctic.oscillations"

Details

See http://jisao.washington.edu/ao/ for more details on the importance of this time series.

Source

The data came from the file AO\_TREN\_NCEP\_Jan1899Current.ascii found at http://www.atmos.colostate.edu/ao/Data/ao_index.html.

Examples

data(arctic.oscillations)
x = ts(arctic.oscillations, start=c(1899,1), frequency=12)
plot(x)
Mothers and their babies data

Description
A collection of variables taken for each new mother in a Child and Health Development Study.

Usage
data(babies)

Format
A data frame with 1,236 observations on the following 23 variables.

Variables in data file

<table>
<thead>
<tr>
<th>id</th>
<th>identification number</th>
</tr>
</thead>
<tbody>
<tr>
<td>plurality</td>
<td>5= single fetus</td>
</tr>
<tr>
<td>outcome</td>
<td>1= live birth that survived at least 28 days</td>
</tr>
<tr>
<td>date</td>
<td>birth date where 1096=January1,1961</td>
</tr>
<tr>
<td>gestation</td>
<td>length of gestation in days</td>
</tr>
<tr>
<td>sex</td>
<td>infant’s sex 1=male 2=female 9=unknown</td>
</tr>
<tr>
<td>wt</td>
<td>birth weight in ounces (999 unknown)</td>
</tr>
<tr>
<td>parity</td>
<td>total number of previous pregnancies including fetal deaths and still births, 99=unknown</td>
</tr>
<tr>
<td>race</td>
<td>mother’s race 0-5=white 6=mex 7=black 8=asian 9=mixed 99=unknown</td>
</tr>
<tr>
<td>age</td>
<td>mother’s age in years at termination of pregnancy, 99=unknown</td>
</tr>
<tr>
<td>ed</td>
<td>mother’s education 0= less than 8th grade, 1 = 8th -12th grade - did not graduate, 2= HS graduate–no other schooling , 3= HS+trade, 4=HS+some college 5= College graduate, 6&amp;7 Trade school HS unclear, 9=unknown</td>
</tr>
<tr>
<td>ht</td>
<td>mother’s height in inches to the last completed inch 99=unknown</td>
</tr>
<tr>
<td>wt1</td>
<td>mother prepregnancy wt in pounds, 999=unknown</td>
</tr>
<tr>
<td>drace</td>
<td>father’s race, coding same as mother’s race.</td>
</tr>
<tr>
<td>dage</td>
<td>father’s age, coding same as mother’s age.</td>
</tr>
<tr>
<td>ded</td>
<td>father’s education, coding same as mother’s education.</td>
</tr>
<tr>
<td>dht</td>
<td>father’s height, coding same as for mother’s height</td>
</tr>
<tr>
<td>dwt</td>
<td>father’s weight coding same as for mother’s weight</td>
</tr>
<tr>
<td>marital</td>
<td>1=married, 2= legally separated, 3= divorced, 4=widowed, 5=never married</td>
</tr>
<tr>
<td>inc</td>
<td>family yearly income in $2500 increments 0 = under 2500, 1=2500-4999, ..., 8= 12,500-14,999, 9=15000+, 98=unknown, 99=not asked</td>
</tr>
<tr>
<td>smoke</td>
<td>does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9=unknown</td>
</tr>
</tbody>
</table>
**time** If mother quit, how long ago? 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+ yrs ago, 9=quit and don’t know, 98=unknown, 99=not asked

**number** number of cigs smoked per day for past and current smokers 0=never, 1=1-4, 2=5-9, 3=10-14, 4=15-19, 5=20-29, 6=30-39, 7=40-60, 8=60+, 9=smoke but don’t know, 98=unknown, 99=not asked

**Source**

This dataset is found from [http://www.stat.berkeley.edu/users/statlabs/labs.html](http://www.stat.berkeley.edu/users/statlabs/labs.html). It accompanies the excellent text *Stat Labs: Mathematical Statistics through Applications* Springer-Verlag (2001) by Deborah Nolan and Terry Speed.

**Examples**

```r
data(babies)
plot(wt ~ factor(smoke), data=babies)
plot(wt1 ~ dwt, data=babies, subset=wt1 < 800 & dwt < 800)
```

---

**babyboom**

*Babyboom: data for 44 babies born in one 24-hour period.*

**Description**

The babyboom dataset contains the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia.

**Usage**

```r
data(babyboom)
```

**Format**

A data frame with 44 observations on the following 4 variables.

- **clock.time** Time on clock
- **gender** a factor with levels girl boy
- **wt** weight in grams of child
- **running.time** minutes after midnight of birth

**Source**

This data set was submitted to the *Journal of Statistical Education*, [http://www.amstat.org/publications/jse/secure/v7n3/datasets.dunn.cfm](http://www.amstat.org/publications/jse/secure/v7n3/datasets.dunn.cfm), by Peter K. Dunn.
Description

This dataset contains batting statistics for the 2002 baseball season. The data allows you to compute batting averages, on base percentages, and other statistics of interest to baseball fans. The data only contains players with more than 100 at bats for a team in the year. The data is excerpted with permission from the Lahman baseball database at http://www.baseball1.com/.

Usage

data(batting)

Format

A data frame with 438 observations on the following 22 variables.

- **playerID**: This is coded, but those familiar with the players should be able to find their favorites.
- **yearID**: a numeric vector. Always 2002 in this dataset.
- **stintID**: a numeric vector. Player’s stint (order of appearances within a season)
- **teamID**: a factor with Team
- **lgID**: a factor with levels AL NL
- **G**: number of games played
- **AB**: number of at bats
- **R**: number of runs
- **H**: number of hits
- **DOUBLE**: number of doubles. "2B" in original data base
- **TRIPLE**: number of triples. "3B" in original data base
- **HR**: number of home runs
- **RBI**: number of runs batted in
- **SB**: number of stolen bases
- **CS**: number of times caught stealing
- **BB**: number of base on balls (walks)
- **SO**: number of strikeouts
- **IBB**: number of intentional walks
- **HBP**: number of hit by pitches
- **SH**: number of sacrifice hits
- **SF**: number of sacrifice flies
- **GIDP**: number of grounded into double plays
Details

Baseball fans are “statistics” crazy. They love to talk about things like RBIs, BAs and OBPs. In order to do so, they need the numbers. This data comes from the Lahman baseball database at http://www.baseball1.com/. The complete dataset includes data for all of baseball not just the year 2002 presented here.

Source


References

In addition to the data set above, the book Curve Ball, by Albert, J. and Bennett, J., Copernicus Books, gives an extensive statistical analysis of baseball.


Examples

data(batting)
attach(batting)
BA = H/AB # batting average
OBP = (H + BB + HBP) / (AB + BB + HBP + SF) # On base "percentage"

baycheck

Population estimate of type of Bay Checkerspot butterfly

Description

Estimates of the population of a type of Bay Checkerspot butterfly near San Francisco.

Usage

data(baycheck)

Format

A data frame with 27 observations on the following 2 variables.

year a numeric vector
Nt estimated number

Source

### Examples

```r
data(baycheck)
plot(Nt ~ year, baycheck)
## fit Ricker model \( N_{t+1} = N_t e^{-rt}\)
\( n = \text{length(baycheck}\$\text{year}) \)
\( yt = \text{with(baycheck, log(Nt[-1]/Nt[-n])} \)
\( nt = \text{with(baycheck, Nt[-n])} \)
\( \text{lm(yt ~ nt, baycheck)} \)
```

### best.times

<table>
<thead>
<tr>
<th>best.times</th>
<th>Best track and field times by age and distance</th>
</tr>
</thead>
</table>

### Description

A dataset giving world records in track and field running events for various distances and different age groups.

### Usage

```r
data(best.times)
```

### Format

A data frame with 113 observations on the following 6 variables.

- **Dist**  Distance in meters (42195 is a marathon)
- **Name**  Name of record holder
- **Date**  Date of record
- **Time**  Time in seconds
- **Time.1**  Time as character
- **age**  Age at time of record

### Details

Age-graded race results allow competitors of different ages to compare their race performances. This data set allows one to see what the relationship is based on peak performances.

### Source

The data comes from [http://www.personal.rdg.ac.uk/~snsgrubb/athletics/agegroups.html](http://www.personal.rdg.ac.uk/~snsgrubb/athletics/agegroups.html) which includes a calculator to compare results.
Examples

```r
data(best.times)
attach(best.times)
by.dist = split(best.times,as.factor(Dist))
lm(scale(Time) ~ age, by.dist[['400']])
dists = names(by.dist)
lapply(dists, function(n) print(lm(scale(Time) ~ age, by.dist[[n]])))
```

---

**blood**

*blood pressure readings*

---

**Description**

Blood pressure of 15 males taken by machine and expert

**Usage**

```r
data(blood)
```

**Format**

This data frame contains the following columns:

- **Machine** a numeric vector
- **Expert** a numeric vector

**Source**

Taken from Kitchen’s Exploring Statistics.

**References**

~~ possibly secondary sources and usages ~~

**Examples**

```r
data(blood)
attach(blood)
t.test(Machine,Expert)
detach(blood)
```
**breakdown**  
*Time of insulating fluid to breakdown*

**Description**  
The time in minutes for an insulating fluid to break down under varying voltage loads

**Usage**  
```r
data(breakdown)
```

**Format**  
A data frame with 75 observations on the following 2 variables.

- **voltage**  Number of kV
- **time**  time in minutes

**Details**  
An example from industry where a linear model is used with replication and transformation of variables.

**Source**  

**Examples**  
```r
data(breakdown)
plot(log(time) ~ voltage, data = breakdown)
```

---

**bright.stars**  
*List of bright stars with Hipparcos catalog number*

**Description**  
List of bright stars with Hipparcos catalog number.

**Usage**  
```r
data(bright.stars)
```
Format

A data frame with 96 observations on the following 2 variables.

name  Common name of star

hip  HIP number for identification

Details

The source of star names goes back to the Greeks and Arabs. Few are modern. This is a list of 96 common stars.

Source


Examples

data(bright.stars)
all.names = paste(bright.stars$name, sep="", collapse="")
x = unlist(strsplit(tolower(all.names), ""))
letter.dist = sapply(letters, function(i) sum(x == i))
data(scrabble) # for frequency info
p = scrabble$frequency[1:26]; p = p/sum(p)
chisq.test(letter.dist, p=p) # compare with English

brightness 

Brightness of 966 stars

Description

The Hipparcos Catalogue has information on over 100,000 stars. Listed in this dataset are brightness measurements for 966 stars from a given sector of the sky.

Usage

data(brightness)

Format

A univariate dataset of 966 numbers.

Details

This is field H5 in the catalog measuring the magnitude, V, in the Johnson UBV photometric system. The smaller numbers are for brighter stars.

Source

http://astro.estec.esa.nl/hipparcos
### Bumpers

**Description**

Bumper repair costs for various automobiles

**Usage**

`data(bumpers)`

**Format**

Price in dollars to repair a bumper.

**Source**


**Examples**

```r
data(bumpers)
stem(bumpers)
```

### BushApproval

**Description**

U.S. President George Bush approval ratings

**Usage**

`data(BushApproval)`

**Format**

A data frame with 323 observations on the following 3 variables.

- **date**: The date poll was begun (some take a few days)
- **approval**: a numeric number between 0 and 100
- **who**: a factor with levels fox, gallup, newsweek, time, cnn, upenn, zogby
Details

A data set of approval ratings of George Bush over the time of his presidency, as reported by several agencies. Most polls were of size approximately 1,000 so the margin of error is about 3 percentage points.

Source


Examples

data(BushApproval)
attach(BushApproval)

```r
## Plot data with confidence intervals. Each poll gets different line type
## no points at first
plot(strptime(date,"%m/%d/%y"),approval,type="n",
ylab = "Approval Rating",xlab="Date",
ylim=c(30,100)
)

## plot line for CI. Margin or error about 3
## matlines has trouble with dates from strftime()
colors = rainbow(6)

for(i in 1:nrow(BushApproval)) {
    lines(rep(strptime(date[i],"%m/%d/%y"),2),
c(approval[i]-3,approval[i]+3),
lty=as.numeric(who[i]),
col=colors[as.numeric(who[i])]
}

## plot points
points(strptime(date,"%m/%d/%y"),approval,pch=as.numeric(who))

## add legend
legend((2003-1970)*365+24*60*60,90,levels(who),lty=1:6,col=1:6)
detach(BushApproval)
```

bycatch

<table>
<thead>
<tr>
<th>bycatch</th>
<th>Number of Albatrosses accidentally caught during a fishing haul</th>
</tr>
</thead>
</table>

Description

This data set from Hillborn and Mangel contains data on the number of Albatrosses accidentally caught while fishing by commercial fisheries.
Usage
data(bycatch)

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

no.albatross  The number of albatross caught
no.hauls  Number of hauls with this many albatross caught

Details
During fishing operations non-target species are often captured. These are called “incidental catch”. In some cases, large-scale observer programs are used to monitor this incidental catch. When fishing for squid, albatrosses are caught while feeding on the squid at the time of fishing. This feeding is encouraged while the net is being hauled in, as the squid are clustered making it an opportunistic time for the albatross to eat.

Source

Examples
data(bycatch)
hauls = with(bycatch,rep(no.albatross,no.hauls))

---
cabinet  Estimated tax savings for US President Bush’s cabinet

description
Estimated savings from a repeal of the tax on capital gains and dividends for Bush’s cabinet members.

Usage
data(cabinet)

Format
A data frame with 19 observations on the following 4 variables.

name  Name of individual
position  Position of individual
est.dividend.cg  Estimated amount of dividend and capital gain income
est.tax.savings  Estimated tax savings
Details


“On May 22, 2003, the House of Representatives and the Senate passed tax legislation that included $320 billion in tax cuts. The final tax cut bill was signed into law by President Bush on May 28, 2003. The largest component of the new tax law is the reduction of tax rates on both capital gains and dividend income. The law also includes the acceleration of future tax cuts, as well as new tax reductions for businesses.

This capital gains and dividend tax cut will have virtually no impact on the average American. The vast majority of Americans (88 no capital gains on their tax returns. These taxpayers will receive no tax savings at all from the reduction in taxes on capital gains. Similarly, most Americans (75 from the reduction of taxes on dividends.

While the average American will derive little, if any, benefit from the cuts in dividend and capital gains taxes, the law offers significant benefits to the wealthy. For example, the top 1 receive an average tax cut of almost $21,000 each. In particular, some of the major beneficiaries of this plan will be Vice President Cheney, President Bush, and other members of the cabinet. Based on 2001 and 2002 dividends and capital gains income, Vice President Cheney, President Bush, and the cabinet are estimated to receive an average tax cut of at least $42,000 per year. Their average tax savings equals the median household income in the United States.”

Source


Examples

data(cabinet)
attach(cabinet)
median(est.dividend.cg)
mean(est.dividend.cg)
detach(cabinet)

---

camp

*Mount Campito Yearly Treering Data, -3435-1969.*

Description

Contains annual tree-ring measurements from Mount Campito from 3426 BC through 1969 AD.

Usage

data(camp)

Format

A univariate time series with 5405 observations. The object is of class ""ts"".
Details

This series is a standard example for the concept of long memory time series.
The data was produced and assembled at the Tree Ring Laboratory at the University of Arizona, Tuscon.

Source


References

This data set is in the tseries package. It is repackaged here for convenience only.

Examples

data(camp)
acf(camp)

---

cancer  cancer survival times

description

cancer survival times

Usage

data(cancer)

Format

The format is: List of 5 numeric components stomach, bronchus, colon, ovary and breast

Source


Examples

data(cancer)
boxplot(cancer)
**carbon**  
*Carbon Monoxide levels at different sites*

**Description**
Carbon Monoxide levels at different sites

**Usage**
data(carbon)

**Format**
This data frame contains the following columns:
- **Monoxide** a numeric vector
- **Site** a numeric vector

**Source**
Borrowed from Kitchen’s Exploring Statistics

**Examples**
data(carbon)  
boxplot(Monoxide ~ Site, data=carbon)

---

**carsafety**  
*Fatality information in U.S. for several popular cars*

**Description**
Safety statistics appearing in a January 12th, 2004 issue of the *New Yorker* showing fatality rates per million vehicles both for drivers of a car, and drivers of other cars that are hit.

**Usage**
data(carsafety)

**Format**
A data frame with 33 observations on the following 4 variables.
- **Make.model** The make and model of the car
- **type** Type of car
- **Driver.deaths** Number of drivers deaths per year if 1,000,000 cars were on the road
- **Other.deaths** Number of deaths in other vehicle caused by accidents involving these cars per year if 1,000,000 cars were on the road
Details

The article this data came from wishes to make the case that SUVs are not safer despite a perception among the U.S. public that they are.

Source


Examples

```r
data(carsafety)
plot(Driver.deaths + Other.deaths ~ type, data = carsafety)
plot(Driver.deaths + Other.deaths ~ type, data = carsafety)
```

---

**central.park**

**Weather in Central Park NY in May 2003**

---

Description

A listing of various weather measurements made at Central Park in New York City during the month of May 2003.

Usage

```r
data(central.park)
```

Format

A data frame with 31 observations on the following 19 variables.

- **DY** the day
- **MAX** maximum temperature (temperatures in Farenheit)
- **MIN** minimum temperature
- **AVG** average temperature
- **DEP** departure from normal
- **HDD** heating degree days
- **CDD** cooling degree days
- **WTR** Water fall. A factor as "T" is a trace.
- **SNW** Amount of snowfall
- **DPH** Depth of snow
- **SPD** Average wind speed
- **SPD1** Max wind speed
central.park.cloud

DIR  2 minimum direction
MIN2 Sunshine measurement a factor with two levels 0 M
PSBL Sunshine measurement a factor with levels 0 M
S.S Sunshine measurement. 0-3 = Clear, 4-7 partly cloudy, 8-10 is cloudy
WX (This is not as documented in the data source. Ignore this variable. It should be: 1 = FOG, 2 =
FOG REDUCING VISIBILITY TO 1/4 MILE OR LESS, 3 = THUNDER, 4 = ICE PELLETS,
5 = HAIL, 6 = GLAZE OR RIME, 7 = BLOWING DUST OR SAND; VS BY 1/2 MILE OR
LESS, 8 = SMOKE OR HAZE, 9 = BLOWING SNOW, X = TORNADO)
SPD3 peak wind speed
DR direction of peak wind

Details
This datasets summarizes the weather in New York City during the merry month of May 2003. This
data set comes from the daily climate report issued by the National Weather Service Office.

Source
This data is published by http://www.noah.gov.

Examples

data(central.park)
attach(central.park)
barplot(rbind(MIN,M=MIN),ylim=c(0,80))

central.park.cloud  Type of day in Central Park, NY May 2003

Description
The type of day in May 2003 in Central Park, NY

Usage
data(central.park.cloud)

Format
A factor with levels clear, partly, cloudy and cloudy.

Source
This type of data, and much more, is available from http://www.noaa.gov.

Examples

data(central.park.cloud)
table(central.park.cloud)
ceo2013  

*CEO compensation in 2013*

**Description**

Data on top 200 CEO compensations in the year 2013

**Usage**

`data(ceo2013)`

**Format**

A data frame.

**Source**


**Examples**

`data(ceo2013)`

---

**cfb**  

*Bootstrap sample from the Survey of Consumer Finances*

**Description**

A bootstrap sample from the “Survey of Consumer Finances”.

**Usage**

`data(cfb)`

**Format**

A data frame with 1000 observations on the following 14 variables.

- **WGT**  Weights to compensate for undersampling. Not applicable
- **AGE**  Age of participants
- **EDUC**  Education level (number of years) of participant
- **INCOME**  Income in year 2001 of participant
- **CHECKING**  Amount in checking account for participant
- **SAVING**  Amount in savings accounts
NMMF  Total directly-held mutual funds
STOCKS  Amount held in stocks
FIN  Total financial assets
VEHIC  Value of all vehicles (includes autos, motor homes, RVs, airplanes, boats)
HOMEEQ  Total home equity
OTHNFIN  Other financial assets
DEBT  Total debt
NETWORTH  Total net worth

Details
The data is oversampled to compensate for low response in the upper brackets. To compensate, weights are assigned. By bootstrapping the data with the weights, we get a “better” version of a random sample from the population.

Source

Examples
data(chicken)
attach(chicken)
mean(INCOME)

| chicken | weight gain of chickens fed 3 different rations |

Description
weight gain of chickens fed 3 different rations

Usage
data(chicken)

Format
This data frame contains the following columns:
Ration1  a numeric vector
Ration2  a numeric vector
Ration3  a numeric vector
Source

From Kitchens’ Exploring Statistics.

Examples

data(chips)
boxplot(chips)

---

chips Measurements of chip wafers

Description

The chips data frame has 30 rows and 8 columns.

Usage

data(chips)

Format

This data frame contains the following columns:

<table>
<thead>
<tr>
<th>wafer11</th>
<th>a numeric vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>wafer12</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer13</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer14</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer21</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer22</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer23</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>wafer24</td>
<td>a numeric vector</td>
</tr>
</tbody>
</table>

Source

From Kitchens’ Exploring Statistics

Examples

data(chips)
boxplot(chips)
**co2emiss**

*Carbon Dioxide Emissions from the U.S.A. from fossil fuel*

**Description**

Carbon Dioxide Emissions from the U.S.A. from fossil fuel

**Usage**

data(co2emiss)

**Format**

The format is: Time-Series [1:276] from 1981 to 2004: -30.5 -30.4 -30.3 -29.8 -29.6 ...

**Details**

Monthly estimates of 13C/12C in fossil-fuel CO2 emissions. For write up see [http://cdiac.esd.ornl.gov/trends/emis_mon/emis_mon_co2.html](http://cdiac.esd.ornl.gov/trends/emis_mon/emis_mon_co2.html). In particular find:

"An annual cycle, peaking during the winter months and reflecting natural gas consumption, and a semi-annual cycle of lesser amplitude, peaking in summer and winter and reflecting coal consumption, comprise the dominant features of the annual pattern. The relatively constant emissions until 1987, followed by an increase from 1987-1989, a decrease in 1990-1991 and record highs during the late 1990s, are also evident in the annual data of Marland et al. However, emissions have declined somewhat since 2000."

**Source**


**Examples**

data(co2emiss)
monthplot(co2emiss)
stl(co2emiss, s.window="periodic")

**coins**

*The coins in my change bin*

**Description**

The coins in author’s change bin with year and value.

**Usage**

data(coins)
Format

A data frame with 371 observations on the following 2 variables.

year  Year of coin
value  Value of coin: quarter, dime, nickel, or penny

Examples

data(coins)
years = cut(coins$year,seq(1920,2010,by=10),include.lowest=TRUE,
        labels = paste(192:200,"",sep=""))
table(years)

coldvermont  Daily minimum temperature in Woodstock Vermont

Description


Usage

data(coldvermont)

Format

A ts object with daily frequency

Source

Extracted from http://www.ce.washington.edu/pub/HYDRO/edm/met_thru_97/vttmin.dly.gz. Errors were possibly introduced.

Examples

data(coldvermont)
plot(coldvermont)
confint.h.test

Produce confidence interval for objects of class h.test

Description
Simple means to output a confidence interval for an h.test object.

Usage
```r
## S3 method for class 'h.test'
confint(object, parm, level, ...)  
```

Arguments
- `object`: A object of class htest, such as output from `t.test`.
- `parm`: ignored
- `level`: ignored
- `...`: can pass in function to transform via `transform` argument.

Value
No return value, outputs interval through `cat`.

Examples
```r
confint(t.test(rnorm(10)))
```

corn

Comparison of corn for new and standard variety

Description
Comparison of corn for new and standard variety

Usage
```r
data(corn)
```

Format
This data frame contains the following columns:
- `New`: a numeric vector
- `Standard`: a numeric vector
Source

From Kitchens’ Exploring Statistics

Examples

data(corn)
t.test(corn)

crime

violent crime rates in 50 states of US

Description

crime rates for 50 states in 1983 and 1993

Usage

data(corn)

Format

This data frame contains the following columns:

y1983 a numeric vector
y1993 a numeric vector

Source

from Kitchens’ Exploring Statistics

Examples

data(corn)
boxplot(corn)
t.test(corn[,1],corn[,2],paired=TRUE)
---

**Description**

The data collected in a calibration experiment consisting of a known load, applied to the load cell, and the corresponding deflection of the cell from its nominal position.

**Usage**

data(deflection)

**Format**

A data frame with 40 observations on the following 2 variables.

- **Deflection**: a numeric vector
- **Load**: a numeric vector

**Source**


**Examples**

data(deflection)

```r
res = lm(Deflection ~ Load, data = deflection)
plot(Deflection ~ Load, data = deflection)
abline(res) # looks good?
plot(res)
```

---

**demos**

*Provide menu for possible shiny demonstrations*

**Description**

Provides a menu to open one of the provided demonstrations which use **shiny** for animation.

**Usage**

demos()

**Details**

User must have installed **shiny** prior to usage. As **shiny** has some dependencies that don’t always work, this package is not a dependency of **UsingR**.
Value

No return value, when called a web page opens. Use Ctrl-C (or equivalent) in terminal to return to an interactive session.

Examples

```r
## demos()
```

---

<table>
<thead>
<tr>
<th>DensityPlot</th>
<th>Plots densities of data</th>
</tr>
</thead>
</table>

Description

Allows one to compare empirical densities of different distributions in a simple manner. The density is used as graphs with multiple histograms are too crowded. The usage is similar to side-by-side boxplots.

Usage

```r
DensityPlot(x, ...)
```

Arguments

- `x`  
  x may be a sequence of data vectors (eg. x,y,z), a data frame with numeric column vectors or a model formula

- `...`  
  You can pass in a bandwidth argument such as bw="SJ". See `density` for details. A legend will be placed for you automatically. To override the positioning set do.legend="manual". To skip the legend, set do.legend=FALSE.

Value

Makes a plot

Author(s)

John Verzani

References

Basically a modified boxplot function. As well it should be as it serves the same utility: comparing distributions.

See Also

`boxplot`, `violinplot`, `density`
### diamond

**Price by size for diamond rings**

#### Description

A data set on 48 diamond rings containing price in Singapore dollars and size of diamond in carats.

#### Usage

```r
data(diamond)
```

#### Format

A data frame with 48 observations on the following 2 variables.

- **carat**: A measurement of a diamond’s size
- **price**: Price in Singapore dollars

#### Details

This data comes from a collection of the *Journal of Statistics Education*. The accompanying documentation says:

“Data presented in a newspaper advertisement suggest the use of simple linear regression to relate the prices of diamond rings to the weights of their diamond stones. The intercept of the resulting regression line is negative and significantly different from zero. This finding raises questions about an assumed pricing mechanism and motivates consideration of remedial actions.”

#### Source


#### Examples

```r
data(diamond)
plot(price ~ carat, diamond, pch=5)
```
divorce  Time until divorce for divorced women (by age)

Description
The divorce data frame has 25 rows and 6 columns.

Usage
data(divorce)

Format
This data frame contains the following columns:

time of divorce  a factor
all ages  a numeric vector
0-17  a numeric vector
18-19  a numeric vector
20-24  a numeric vector
25-100  a numeric vector

Source
Forgot source

Examples
data(divorce)
apply(divorce[,2:6],2,sum)  # percent divorced by age of marriage

DOTplot  Make big DOT plot likestripchart

Description
A variant of the stripchart using big dots as the default.

Usage
DOTplot(x, ...)

Arguments

May be a vector, data frame, matrix (each column a variable), list or model formula. Treats each variable or group as a univariate dataset and makes corresponding DOTplot.

arguments passed onto points.

Value

Returns the graphic only.

Author(s)

John Verzani

See Also

See also as stripchart, dotplot

Examples

```r
x = c(1, 1, 2, 3, 5, 8)
DOTplot(x, main = "Fibonacci", cex = 2)
```

---

dottodot  

*Dot-to-dot puzzle*

Description

A set of points to make a dot-to-dot puzzle

Usage

```r
data(dottodot)
```

Format

A data frame with 49 observations on the following 4 variables.

x x position

y y position

pos where to put label

ind number for label

Details

Points to make a dot to dot puzzle to illustrate, text, points, and the argument pos.
Source

Illustration by Noah Verzani.

Examples

data(dottodot)
# make a blank graph
plot(y~x,data=dottodot,type="n",bty="n",xaxt="n",yaxt="n",xlab="",ylab="")
# add the points
points(y~x,data=dottodot)
# add the labels using pos argument
with(dottodot, text(x,y,labels=ind,pos=pos))
# solve the puzzle
lines(y~x, data=dottodot)

dowdata

The Dow Jones average from Jan 1999 to October 2000

Description

The dowdata data frame has 443 rows and 5 columns.

Usage

data(dowdata)

Format

This data frame contains the following columns:

- **Open** a numeric vector
- **High** a numeric vector
- **Date** a numeric vector
- **Low** a numeric vector
- **Close** a numeric vector

Source

this data comes from the site [http://www.forecasts.org/](http://www.forecasts.org/)

Examples

data(dowdata)
the.close <- dowdata$Close
n <- length(the.close)
plot(log(the.close[2:n]/the.close[1:(n-1)]))
**dvdsales**

*Monthly DVD player sales since introduction to May 2004*

**Description**

Monthly DVD player sales since introduction of DVD format to May 2004

**Usage**

```r
data(dvdsales)
```

**Format**

Matrix with rows recording the year, and columns the month.

**Source**

[http://www.thedigitalbits.com/articles/cemadvdsales.html](http://www.thedigitalbits.com/articles/cemadvdsales.html)

**Examples**

```r
data(dvdsales)
barplot(t(dvdsales[7:1]),beside=TRUE)
```

---

**emissions**

*CO2 emissions data and gross domestic product for 26 countries*

**Description**

The emissions data frame has 26 rows and 3 columns.

A data set listing GDP, GDP per capita, and CO2 emissions for 1999.

**Usage**

```r
data(emissions)
```

**Format**

This data frame contains the following columns:

- **GDP** a numeric vector
- **perCapita** a numeric vector
- **CO2** a numeric vector
**Source**


Prompted by a plot appearing in a June 2001 issue of the *New York Times*.

**Examples**

```r
data(emissions)
plot(emissions)
```

---

**ewr**

*Taxi in and taxi out times at EWR (Newark) airport for 1999-2001*

---

**Description**

The `ewr` data frame has 46 rows and 11 columns.

Gives taxi in and taxi out times for 8 different airlines and several months at EWR airport.

Airline codes are: AA (American Airlines), AQ (Aloha Airlines), AS (Alaska Airlines), CO (Continental Airlines), DL (Delta Airlines), HP (America West Airlines), NW (Northwest Airlines), TW (Trans World Airlines), UA (United Airlines), US (US Airways), and WN (Southwest Airlines)

**Usage**

```r
data(ewr)
```

**Format**

This data frame contains the following columns:

- **Year**: a numeric vector
- **Month**: a factor for months
- **AA**: a numeric vector
- **CO**: a numeric vector
- **DL**: a numeric vector
- **HP**: a numeric vector
- **NW**: a numeric vector
- **TW**: a numeric vector
- **UA**: a numeric vector
- **US**: a numeric vector
- **inorout**: a factor with levels *in* or *out*

**Source**

Examples

data(exec.pay)
boxplot(exec.pay[3:10])

exec.pay  Direct compensation for 199 United States CEOs in the year 2000

Description

Direct compensation for 199 United States CEOs in the year 2000 in units of $10,000.

Usage

data(exec.pay)

Format

A numeric vector with 199 entries each measuring compensation in 10,000s of dollars.

Source


Examples

data(exec.pay)
hist(exec.pay)

fat  Body measurements to predict percentage of body fat in males

Description

A data set containing many physical measurements of 252 males. Most of the variables can be measured with a scale or tape measure. Can they be used to predict the percentage of body fat? If so, this offers an easy alternative to an underwater weighing technique.

Usage

data(fat)
Format

A data frame with 252 observations on the following 19 variables.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td>Case Number</td>
</tr>
<tr>
<td>body.fat</td>
<td>Percent body fat using Brozek's equation, 457/Density - 414.2</td>
</tr>
<tr>
<td>body.fat.siri</td>
<td>Percent body fat using Siri's equation, 495/Density - 450</td>
</tr>
<tr>
<td>density</td>
<td>Density (gm/cm^2)</td>
</tr>
<tr>
<td>age</td>
<td>Age (yrs)</td>
</tr>
<tr>
<td>weight</td>
<td>Weight (lbs)</td>
</tr>
<tr>
<td>height</td>
<td>Height (inches)</td>
</tr>
<tr>
<td>BMI</td>
<td>Adiposity index = Weight/Height^2 (kg/m^2)</td>
</tr>
<tr>
<td>ffweight</td>
<td>Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek’s formula (lbs)</td>
</tr>
<tr>
<td>neck</td>
<td>Neck circumference (cm)</td>
</tr>
<tr>
<td>chest</td>
<td>Chest circumference (cm)</td>
</tr>
<tr>
<td>abdomen</td>
<td>Abdomen circumference (cm) “at the umbilicus and level with the iliac crest”</td>
</tr>
<tr>
<td>hip</td>
<td>Hip circumference (cm)</td>
</tr>
<tr>
<td>thigh</td>
<td>Thigh circumference (cm)</td>
</tr>
<tr>
<td>knee</td>
<td>Knee circumference (cm)</td>
</tr>
<tr>
<td>ankle</td>
<td>Ankle circumference (cm)</td>
</tr>
<tr>
<td>bicep</td>
<td>Extended biceps circumference (cm)</td>
</tr>
<tr>
<td>forearm</td>
<td>Forearm circumference (cm)</td>
</tr>
<tr>
<td>wrist</td>
<td>Wrist circumference (cm) &quot;distal to the styloid processes&quot;</td>
</tr>
</tbody>
</table>

Details

From the source:

“The data are as received from Dr. Fisher. Note, however, that there are a few errors. The body densities for cases 48, 76, and 96, for instance, each seem to have one digit in error as can be seen from the two body fat percentage values. Also note the presence of a man (case 42) over 200 pounds in weight who is less than 3 feet tall (the height should presumably be 69.5 inches, not 29.5 inches)! The percent body fat estimates are truncated to zero when negative (case 182).”

Source

This data set comes from the collection of the Journal of Statistics Education at http://www.amstat.org/publications/jse/datasets/fat.txt. The data set was contributed by Roger W. Johnson.

References

The source of the data is attributed to Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, Provo, Utah 84602,
Examples

```r
data(fat)
f = body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
    hip + thigh + knee + ankle + bicep + forearm + wrist
res = lm(f, data=fat)
summary(res)
```

father.son

*Pearson's data set on heights of fathers and their sons*

Description

1078 measurements of a father’s height and his son’s height.

Usage

```r
data(father.son)
```

Format

A data frame with 1078 observations on the following 2 variables.

- **fheight**  Father’s height in inches
- **sheight**  Son’s height in inches

Details

Data set used by Pearson to investigate regression. See data set galton for data set used by Galton.

Source

Read into R by the command

```r
read.table("http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat",sep=" ")[-1],
```
as mentioned by Chuck Cleland on the r-help mailing list.

Examples

```r
data(father.son)
## like cover of Freedman, Pisani, and Purves
plot(sheight ~ fheight, data=father.son,bty="1",pch=20)
abline(a=0,b=1,lty=2,lwd=2)
abline(lm(sheight ~ fheight, data=father.son),lty=1,lwd=2)
```
female.inc

Description

A data set containing incomes for 1,000 females along with race information. The data is sampled from data provided by the United States Census Bureau.

Usage

data(female.inc)

Format

A data frame with 1,000 observations on the following 2 variables.

income  Income for 2001 in dollars
race    a factor with levels black, hispanic or white

Details

The United States Census Bureau provides a lot of data on income distributions. This data comes from the Current Population Survey (CPS) for the year 2001. The raw data appears in table format. This data is sampled from the data in that table.

Source

The original table is found at http://ferret.bls.census.gov/macro/032002/perinc/new11_002.htm.

Examples

data(female.inc)
boxplot(income ~ race, female.inc)
boxplot(log(income,10) ~ race, female.inc)
sapply(with(female.inc,split(income,race)),median)
**firstchi**  
*Age of mother at birth of first child*

**Description**  
Age of mother at birth of first child

**Usage**  
data(firstchi)

**Format**  
The format is: num [1:87] 30 18 35 22 23 22 36 24 23 28 ...

**Source**  

**Examples**  
data(firstchi)  
hist(firstchi)

---

**five.yr.temperature**  
*Five years of weather in New York City*

**Description**  
Five years of maximum temperatures in New York City

**Usage**  
data(five.yr.temperature)

**Format**  
A data frame with 2,439 observations on the following 3 variables.  

- **days** Which day of the year  
- **years** The year  
- **temps** Maximum temperature

**Source**  
Dataset found on the internet, but original source is lost.
Examples

```r
data(five.yr.temperature)
attach(five.yr.temperature)
scatter.smooth(temps ~ days, col=gray(.75))
lines(sMOOTH.spline(temps ~ days), lty=2)
lines(supsmu(days, temps), lty=3)
```

---

**florida**

*County-by-county results of year 2000 US presidential election in Florida*

---

**Description**

The `florida` data frame has 67 rows and 13 columns.

Gives a county by county accounting of the US elections in the state of Florida.

**Usage**

```r
data(florida)
```

**Format**

This data frame contains the following columns:

- **County** Name of county
- **GORE** Votes for Gore
- **BUSH** Votes for Bush
- **BUCHANAN** Votes for Buchanan
- **NADER** Votes for Nader
- **BROWN** a numeric vector
- **HAGELIN** a numeric vector
- **HARRIS** a numeric vector
- **MCREYNOLDS** a numeric vector
- **MOOREHEAD** a numeric vector
- **PHILLIPS** a numeric vector
- **Total** a numeric vector

**Source**

**galileo**

**Galileo data on falling bodies**

**Description**

Data recorded by Galileo in 1609 during his investigations of the trajectory of a falling body.

**Usage**

```r
data(galileo)
```

**Format**

A data frame with 7 observations on the following 2 variables.

- **init.h** Initial height of ball
- **h.d** Horizontal distance traveled

**Details**

A simple ramp 500 punti above the ground was constructed. A ball was placed on the ramp at an indicated height from the ground and released. The horizontal distance traveled is recorded (in punti). (One punto is 169/180 millimeter, not a car by FIAT.)

**Source**

This data and example come from the *Statistical Sleuth* by Ramsay and Schafer, Duxbury (2001), section 10.1.1. They attribute an article in *Scientific American* by Drake and MacLachlan.

**Examples**

```r
data(galileo)
polynomial = function(x, coefs) {
  sum = 0
  for(i in 0:(length(coefs)-1)) {
    sum = sum + coefs[i+1]*x^i
  }
  sum
}
res.lm = lm(h.d ~ init.h, data = galileo)
res.lm2 = update(res.lm, . ~ . + I(init.h^2), data=galileo)
res.lm3 = update(res.lm2, . ~ . + I(init.h^3), data=galileo)
```
Description

Data set from tabulated data set used by Galton in 1885 to study the relationship between a parent’s height and their children.

Usage

data(galton)

Format

A data frame with 928 observations on the following 2 variables.

- **child**: The child’s height
- **parent**: The “midparent” height

Details

The midparent’s height is an average of the father’s height and 1.08 times the mother’s. In the data there are 205 different parents and 928 children. The data here is truncated at the ends for both parents and children so that it can be treated as numeric data. The data were tabulated and consequently made discrete. The father.son data set is similar data used by Galton and is continuous.

Source

This data was found at [http://www.bun.kyoto-u.ac.jp/~suchii/galton86.html](http://www.bun.kyoto-u.ac.jp/~suchii/galton86.html).

See also the data set father.son which was found from [http://stat-www.berkeley.edu/users/juliah/141C/pearson.dat](http://stat-www.berkeley.edu/users/juliah/141C/pearson.dat).

Examples

data(galton)
plot(galton)
## or with some jitter.
plot(jitter(child,5) ~ jitter(parent,5),galton)
## sunflowerplot shows flowers for multiple plots (Thanks MM)
sunflowerplot(galton)
Sales data for the Gap

**Description**
Sales data for the Gap from Jan

**Usage**
data(gap)

**Format**
The format is a `ts` object storing data from June 2002 through June 2005.

**Source**
http://home.businesswire.com

**Examples**
data(gap)
monthplot(gap)

Monthly average gasoline prices in the United States

**Description**
Average retail gasoline prices per month in the United States from January 2000 through February 2006. The hurricane Katrina caused a percentage loss of refinery capability leading to rapidly increasing prices.

**Usage**
data(gasprices)

**Format**
The format is: Time-Series [1:74] from 2000 to 2006: 129 138 152 146 148 ...

**Source**
From the Department of Energy web site: http://tonto.eia.doe.gov/oog/info/gdu/gaspump.html
getAnswer

Display answers to selected problems

Description

Displays answers to selected problems in the system’s web browser.

Usage

getAnswer(chapter = NULL, problem = NULL)
errata()

Arguments

chapter The chapter number
problem The problems number Not all answers are available.

Details

Some selected answers from the problems in Using R for Introductory Statistics are available from the UsingR package. The getAnswer function will display them one-by-one in the browser. The errata function will display the list of errata.

Value

If available, opens web browser to the requested answer or errata page

Author(s)

John Verzani

Examples

getAnswer()
goalspergame

**Goals per game in NHL**

**Description**

Goals per game in NHL

**Usage**

data(goalspergame)

**Format**


**Source**

Off internet site. Forgot which.

**Examples**

data(goalspergame)

---

goalspergame  

**google**

**Google stock values during 2005-02-07 to 2005-07-07**

**Description**

Closing stock price of a share of Google stock during 2005-02-07 to 2005-07-07

**Usage**

data(google)

**Format**

A data vector of numeric values with names attribute giving the dates.

**Source**

finance.yahoo.com

**Examples**

data(google)
plot(google, type="l")
**grades**  
*Current and previous grades*

**Description**  
A dataframe of a student's grade and their grade in their previous class. Graded on American A-F scale.

**Usage**  
data(grades)

**Format**  
A dataframe of 122 rows with 2 columns
- **prev** The grade in the previous class in the subject matter
- **grade** The grade in the current class

**Examples**  
data(grades)  
```r
table(grades)
```  

---

**grip**  
*Effects of cross-country ski-pole grip*

**Description**  
Simulated data set investigating effects of cross-country ski-pole grip.

**Usage**  
data(grip)

**Format**  
A data frame with 36 observations on the following 4 variables.
- **UBP** Measurement of upper-body power
- **person** One of four skiers
- **grip.type** Either classic, modern, or integrated.
- **replicate** A numeric vector
Details

Based on a study done at http://www.montana.edu/wwwd/movementscilab/ mentioned at http://www.xsksiworld.com/. The study investigates the effect of grip type on upper body power. As this influences performance in races, presumably a skier would prefer the grip that provides the best power output.

Examples

```r
data(grip)
ftable(xtabs(UBP ~ person + replicate + grip.type,grip))
```

```
hall.fame               Data frame containing baseball statistics including Hall of Fame membership
```

Description

A data frame containing baseball statistics for several players.

Usage

```r
data(hall.fame)
```

Format

A data frame with 1340 observations on the following 28 variables.

- `first` first name
- `last` last name
- `seasons` Seasons played
- `games` Games played
- `AB` Official At Bats
- `runs` Runs scored
- `hits` hits
- `doubles` doubles
- `triples` triples numeric vector
- `HR` Home runs
- `RBI` Runs batted in
- `BB` Base on balls
- `SO` Strike outs
- `BA` Batting Average
- `OBP` On Base percentage
- `SP` Slugging Percentage
Adjusted productions

batting runs

adjusted batting runs

Runs created

Runs created

Stolen Bases

Caught stealing

Runs scored by stealing

Fielding average

Fielding runs

C = Catcher, 1 = First Base, 2 = Second Base, 3 = Third Base, S = Shortstop, O = Outfield, and D = Designated hitter

a numeric vector

Not a member, Elected by the BBWAA, or Chosen by the Old Timers Committee or Veterans Committee

The sport of baseball lends itself to the collection of data. This data set contains many variables used to assess a players career. The Hall of Fame is reserved for outstanding players as judged initially by the Baseball Writers Association and subsequently by the Veterans Committee.


data(hall.fame)

hist(hall.fame$OBP)

with(hall.fame,last[Hall.Fame.Membership != "not a member"])

headtail(x, k = 3)

helper function to shorten display of a data frame
Arguments

- x a data frame
- k number of rows at top and bottom to show.

Value

No return value. Uses cat to show data

Examples

```
headtail(mtcars)
```

<table>
<thead>
<tr>
<th>healthy</th>
<th>Healthy or not?</th>
</tr>
</thead>
</table>

Description

Data on whether a patient is healthy with two covariates.

Usage

```
data(healthy)
```

Format

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

- p One covariate
- g Another covariate
- healthy 0 is healthy, 1 is not

Details

Data on health with information from two unspecified covariates.

Examples

```
data(healthy)
library(MASS)
stepAIC(glm(healthy ~ p + g, healthy, family=binomial))
```
Description

Simulated data of age vs. max heart rate

Usage

data(heartrate)

Format

This data frame contains the following columns:

- **age**  a numeric vector
- **maxrate** a numeric vector

Details

Does this fit the workout room value of 220 - age?

Source


Examples

data(heartrate)
plot(heartrate)
abline(lm(maxrate ~ age, data=heartrate))

Description

The home data frame has 15 rows and 2 columns.

Usage

data(home)
**homedata**

**Format**

This data frame contains the following columns:

- **old** a numeric vector
- **new** a numeric vector

**Details**

See full dataset homedata

**Source**

See full dataset homedata

**Examples**

```r
data(homedata)

## compare on the same scale
boxplot(data.frame(scale(homedata)))
```

---

**homedata**  
*Maplewood NJ assessed values for years 1970 and 2000*

**Description**

The *homedata* data frame has 6841 rows and 2 columns.  
Data set containing assessed values of homes in Maplewood NJ for the years 1970 and 2000. The properties were not officially assessed during that time and it is interesting to see the change in percentage appreciation.

**Usage**

```r
data(homedata)
```

**Format**

This data frame contains the following columns:

- **y1970** a numeric vector
- **y2000** a numeric vector

**Source**

Maplewood Reval

**Examples**

```r
data(homedata)
plot(homedata)
```
Description
The homeprice data frame has 29 rows and 7 columns.

Usage
data(homeprice)

Format
This data frame contains the following columns:

- **list** list price of home (in thousands)
- **sale** actual sale price
- **full** Number of full bathrooms
- **half** number of half bathrooms
- **bedrooms** number of bedrooms
- **rooms** total number of rooms
- **neighborhood** Subjective assessment of neighborhood on scale of 1-5

Details
This dataset is a random sampling of the homes sold in Maplewood, NJ during the year 2001. Of course the prices will either seem incredibly high or fantastically cheap depending on where you live, and if you have recently purchased a home.

Source
Source Burgdorff Realty.

Examples
data(homeprice)
plot(homeprice$sale,homeprice$list)
abline(lm(homeprice$list~homeprice$sale))
Homework averages for Private and Public schools

**Description**

Homework averages for Private and Public schools

**Usage**

```r
data(homework)
```

**Format**

This data frame contains the following columns:

- **Private** a numeric vector
- **Public** a numeric vector

**Source**

This is from Kitchens Exploring Statistics

**Examples**

```r
data(homework)
boxplot(homework)
```

---

**HUMMER**

Delivery of new HUMMER vehicles

**Description**

Gives monthly delivery numbers for new HUMMER vehicles from June 2003 through February 2006. During July, August, and September 2005 there was an Employee Pricing Incentive.

**Usage**

```r
data(HUMMER)
```

**Format**

The format is: Time-Series [1:33] from 2003 to 2006: 2493 2654 2987 2837 3157 2837 3157 1927 2141 2334 ...
Source
Compiled from delivery data available at http://www.gm.com/company/investor_information/sales_prod/hist_sales.html

Examples
data(HUMMER)
plot(HUMMER)

---

income_percentiles  Top percentiles of U.S. income

Description
Top percentiles of U.S. income

Usage
data(income_percentiles)

Format
A data frame with Year and various percentile (90th, 95th, ...)

Source
Not available

Examples
data(income_percentiles)

---

iq  IQ scores

Description
simulated IQ scores

Usage
data(iq)

Format
The format is: num [1:100] 72 75 77 77 81 82 83 84 84 86 ...
Source

From Kitchens Exploring Statistics

Examples

data(iq)
qqnorm(iq)

kid.weights

Weight and height measurement for a sample of U.S. children

Description

A sample from the data presented in the NHANES III survey (http://www.cdc.gov/nchs/nhanes.htm). This survey is used to form the CDC Growth Charts (http://www.cdc.gov/growthcharts) for children.

Usage

data(kid.weights)

Format

A data frame with 250 observations on the following 4 variables.

age  Age in months
weight  weight in pounds
height  height in inches
gender  Male or Female

Source

This data is extracted from the NHANES III survey: http://www.cdc.gov/nchs/nhanes.htm.

Examples

data(kid.weights)
attach(kid.weights)
plot(weight, height, pch = as.character(gender))
## find the BMI -- body mass index
m.ht = height*2.54/100  # 2.54 cm per inch
m.wt = weight / 2.2046  # 2.2046 lbs. per kg
bmi = m.wt/m.ht^2
hist(bmi)
**KSI**

*Data set on automobile deaths and injuries in Great Britain*

**Description**

Data on car drivers killed, car drivers killed or seriously injured (KSI), and light goods drivers killed during the years 1969 to 1984 in Great Britain. In February 1982 a compulsory seat belt law was introduced.

**Usage**

```r
data(KSI)
```

**Format**

The data is stored as a multi-variate zoo object.

**Source**

Data copied from Appendix 2 "Forecasting, structural time series, models and the Kalman Filter" by Andrew Harvey. The lg.k data is also found in the vandriver's dataset contained in the sspir package.

**References**

Source: HMSO: Road Accidents in Great Britain 1984.

**Examples**

```r
data(KSI)
plot(KSI)
seatbelt = time(KSI) < 1983 + (2-1)/12
```

---

**last.tie**

*Last tie in 100 coin tosses*

**Description**

Toss a coin 100 times and keep a running count of the number of heads and the number of tails. Record the times when the number is tied and report the last one. The distribution will have an approximate “arc-sine” law or well-shaped distribution.

**Usage**

```r
data(last.tie)
```
lawsuits

Format

200 numbers between 0 and 100 indicating when the last tie was.

Details

This data comes from simulating the commands:

```r
x = cumsum(sample(c(-1,1),100,replace=T))
```

and then finding the last tie with

```r
last.tie[i]<-max(0,max(which(!sign(x) == sign(x[length(x)]))))
```

Examples

```r
data(last.tie)
hist(last.tie)
```
Examples

data(lawsuits)
mean(lawsuits)
median(lawsuits)

<table>
<thead>
<tr>
<th>lorem</th>
<th>Placeholder text</th>
</tr>
</thead>
</table>

Description

Lorem Ipsum is simply dummy text of the printing and typesetting industry.

Usage

lorem

Format

a character string

Source

http://www.lipsum.com/

Examples

```
table(unlist(strsplit(lorem, "")))
```

<table>
<thead>
<tr>
<th>malpract</th>
<th>malpractice settlements</th>
</tr>
</thead>
</table>

Description

malpractice settlements

Usage

data(malpract)

Format

The format is: num [1:17] 760 380 125 250 2800 450 100 150 2000 180 ...

Source

From Kitchens Exploring Statistics
Description

A bag of the candy M and M's has many different colors. Each large production batch is blended to the ratios given in this data set. The batches are thoroughly mixed and then the individual packages are filled by weight using high-speed equipment, not by count.

Usage

data(mandms)

Format

A data frame with 5 observations on the following 6 variables.

- **blue**  percentage of blue
- **brown**  percentage of brown
- **green**  percentage of green
- **orange**  percentage of orange
- **red**  percentage of red
- **yellow**  percentage of yellow

Source

This data is attributed to an email sent by Masterfoods USA, A Mars, Incorporated Company. This email was archived at the Math Forum, [http://www.mathforum.org](http://www.mathforum.org).

Examples

```r
data(mandms)
bagfull = c(15,34,7,19,29,24)
names(bagfull) = c("blue","brown","green","orange","red","yellow")
prop = function(x) x/sum(x)
chisq.test(bagfull,p = prop(mandms["milk chocolate",]))
chisq.test(bagfull,p = prop(mandms["Peanut",]))
```
**math**

*Standardized math scores*

**Description**

Standardized math scores

**Usage**

data(math)

**Format**

The format is: num [1:30] 44 49 62 45 51 59 57 55 70 64 ...

**Source**


**Examples**

data(math)
hist(math)

---

**maydow**

*Dow Jones industrial average and May maximum temperature*

**Description**

A data set of both the Dow Jones industrial average and the maximum daily temperature in New York City for May 2003.

**Usage**

data(maydow)

**Format**

A data frame with 21 observations on the following 3 variables.

- **Day**  Day of the month
- **DJA**  The daily close of the DJIQ
- **max.temp**  Daily maximum temperature in Central Park
Details

Are stock traders influenced by the weather? This dataset looks briefly at this question by comparing the daily close of the Dow Jones industrial average with the maximum daily temperature for the month of May 2003. This month was rainy and unseasonably cool weather wise, yet the DJIA did well.

Source


Examples

data(maydow)
attach(maydow)
plot(max.temp,DJA)
plot(max.temp[-1],diff(DJA))

Medicare Sample from "Medicare Provider Charge Data"

Description

Sample from "Medicare Provider Charge Data"

Usage

data(Medicare)

Format

A data frame with 10000 observations and data for on billings for procedures at many different hospitals.

Source


References


Examples

data(Medicare)
**midsize**

*Price of new and used of three mid-sized cars*

**Description**

New and used prices of three popular mid-sized cars.

**Usage**

data(midsize)

**Format**

A data frame with 15 observations on the following 4 variables.

- **Year**  2004 is new car price, others are for used car
- **Accord**  Honda Accord
- **Camry**  Toyota Camry
- **Taurus**  Ford Taurus

**Details**

The value of a car depreciates over time. This data gives the price of a new car and values of similar models for previous years as reported by [http://www.edmunds.com](http://www.edmunds.com).

**Examples**

data(midsize)
plot(Accord ~ I(2004-Year), data = midsize)

---

**MLBattend**

*Major league baseball attendance data*

**Description**

Data on home-game attendance in Major League Baseball for the years 1969-2000.

**Usage**

data(MLBattend)
movies

Format
A data frame with 838 observations on the following 10 variables.

franchise Which team
league American or National league
division Which division
year The year (the year 2000 is recorded as 0)
attendance Actual attendance
runs.scored Runs scored by the team during year
runs.allowed Runs allows by the team during year
wins Number of wins for season
losses Number of losses for season
games.behind A measure of how far from division winner the team was. Higher numbers are worse.

Source

Examples

data(MLBattend)
boxplot(attendance ~ franchise, MLBattend)
with(MLBattend, cor(attendance,wins))

movies Data frome on top 25 movies for some week, many weeks ago

Description
Data on 25 top movies

Usage
data(movies)

Format
A data frame with 26 observations on the following 5 variables.
title Titles
current Current week
previous Previous week
gross Total
**Source**
Some movie website, sorry lost the url.

**Examples**
```
data(movies)
boxplot(movies$previous)
```

---

**movie_data_2011**
*Movie data for 2011 by weekend*

**Description**
Movie data for 2011 by weekend

**Usage**
```
data(movie_data_2011)
```

**Format**
A data frame with variables *Previous* (previous weekend rank), *Movie* (title), *Distributor*, *Genre*, *Gross* (per current weekend), *Change* (change from previous week), *Theaters* (number of theaters), *TotalGross* (total gross to date), *Days* (days out), *weekend* (weekend of report)

**Source**

**Examples**
```
data(movie_data_2011)
```

---

**mw.ages**
*Age distribution in year 2000 in Maplewood New Jersey*

**Description**
Age distribution in Maplewood New Jersey, a suburb of New York City. Data is broken down by Male and Female.

**Usage**
```
data(mw.ages)
```
**Format**

A data frame with 103 observations on the following 2 variables.

- **Male** Counts per age group. Most groups are 1 year, except for 100-104, 105-110, 110+
- **Female** Same

**Source**


**Examples**

data(mw.ages)
barplot(mw.ages$Male + mw.ages$Female)

---

**nba.draft**  
*NBA draft lottery odds for 2002*

**Description**

The NBA draft in 2002 has a lottery

**Usage**

data(nba.draft)

**Format**

A data frame with 13 observations on the following 2 variables.

- **Team** Team name
- **Record** The team won-loss record
- **Balls** The number of balls (of 1000) that this team has in the lottery selection

**Details**

The NBA draft has a lottery to determing the top 13 placings. The odds in the lottery are determined by the won-loss record of the team, with poorer records having better odds of winning.

**Source**


**Examples**

data(nba.draft)
top.pick = sample(row.names(nba.draft),1,prob = nba.draft$Balls)
**nisdc**  

*Description*  

*Usage*

```r
data(nisdc)
```

*Format*

A data frame measuring daily sea-ice extent from 1978 until 2013

*Source*


*References*


---

**normtemp**  

*Description*  
A data set used to investigate the claim that “normal” temperature is 98.6 degrees.

*Usage*

```r
data(normtemp)
```

*Format*

A data frame with 130 observations on the following 3 variables.

- **temperature**  
  normal body temperature
- **gender**  
  Gender 1 = male, 2 = female
- **hr**  
  Resting heart rate
Details

Is normal body temperature 98.6 degrees Fahrenheit? This dataset was constructed to match data presented in an article intending to establish the true value of “normal” body temperature.

Source


References


Examples

data(normtemp)
hist(normtemp$temperature)
t.test(normtemp$temperature,mu=98.2)
summary(lm(temperature ~ factor(gender), normtemp))

npdb

National Practioner Data Bank

Description

Selected variables from the publicly available data from the National Practioner Data Bank (NPDB).

Usage

data(npdb)

Format

A data frame with 6797 observations on the following 6 variables.

- **state**: 2 digit abbreviation of state
- **field**: Field of practice
- **age**: Age of practitioner (rounded down to 10s digit)
- **year**: Year of claim
- **amount**: Dollar amount of reward
- **ID**: A practitioner ID, masked for anonymity

The variable names do not match the original. The codings for field come from a document on http://63.240.212.200/publicdata.html.
Details

This dataset excerpts some interesting variables from the NPDB for the years 2000-2003. The question of capping medical malpractice awards to lower insurance costs is currently being debated nationwide (U.S.). This data is a primary source for determining this debate.


“The legislation that led to the creation of the NPDB was enacted the U.S. Congress believed that the increasing occurrence of medical malpractice litigation and the need to improve the quality of medical care had become nationwide problems that warranted greater efforts than any individual State could undertake. The intent is to improve the quality of health care by encouraging State licensing boards, hospitals and other health care entities, and professional societies to identify and discipline those who engage in unprofessional behavior; and to restrict the ability of incompetent physicians, dentists, and other health care practitioners to move from State to State without disclosure or discovery of previous medical malpractice payment and adverse action history. Adverse actions can involve licensure, clinical privileges, professional society membership, and exclusions from Medicare and Medicaid.”

Source

This data came from [http://www.npdb-hipdb.com/npdb.html](http://www.npdb-hipdb.com/npdb.html)

Examples

data(npdb)
table(table(npdb$ID))  # big offenders
hist(log(npdb$amount))  # log normal?

---

**nym.2002**  
*Random sample of 2002 New York City Marathon finishers*

Description

A random sample of finishers from the New York City Marathon.

Usage

data(nym.2002)

Format

A data frame with 1000 observations on the following 5 variables.

- **place**  Place in the race
- **gender**  What gender
- **age**  Age on day of race
- **home**  Indicator of hometown or nation
- **time**  Time in minutes to finish
ObamaApproval

Details

Each year thousands of participants line up to run the New York City Marathon. This list is a random sample from the finishers.

Source

From the New York City Road Runners web site http://www.nyrrc.org

Examples

data(nym.2002)
with(nym.2002, cor(time, age))

ObamaApproval Approval ratings for President Obama

Description

A collection of approval ratings for President Obama spanning a duration from early 2010 to the summer of 2013.

Usage

data(ObamaApproval)

Format

A data frame 7 variables.

Source


Examples

data(ObamaApproval)
**OBP**

*On base percentage for 2002 major league baseball season*

**Description**

The on base percentage, OBP, is a measure of how often a player gets on base. It differs from the more familiar batting average, as it includes bases on balls (BB) and hit by pitches (HBP). The exact formula is $\text{OBP} = \frac{\text{H} + \text{BB} + \text{HBP}}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}}$.

**Usage**

```r
data(OBP)
```

**Format**

438 numbers between 0 and 1 corresponding to the on base “percentage” for the 438 players who had 100 or more at bats in the 2002 baseball season. The "outlier" is Barry Bonds.

**Source**

This data came from the interesting Lahman baseball data base [http://www.baseball1.com/](http://www.baseball1.com/). The `names` attribute uses the playerID from this database. Unfortunately there were some errors in the extraction from the original data set. Consult the original for accurate numbers.

**Examples**

```r
data(OBP)
hist(OBP)
OBP[OBP>.5] # who is better than 50%? (only Barry Bonds)
```

---

**oral.lesion**

*Oral lesion location by town*

**Description**

A data set on oral lesion location for three Indian towns.

**Usage**

```r
data(oral.lesion)
```

**Format**

A data frame with 9 observations on the following 3 variables.

- **Kerala** a numeric vector
- **Gujarat** a numeric vector
- **Andhra** a numeric vector
ozonemonthly

Source


Examples

data(oral.lesion)
chisq.test(oral.lesion)$p.value
chisq.test(oral.lesion,simulate.p.value=TRUE)$p.value ## exact is 0.0269

ozonemonthly Monthly mean ozone values at Halley Bay Antartica

Description

A time series showing ozone values at Halley Bay Antartica

Usage

data(ozonemonthly)

Format

The format is: Time-Series [1:590] from 1957 to 2006: 313 311 370 359 334 296 288 274 NA NA ...
- attr(*, "names")= chr [1:590] "V5" "V6" "V7" "V8" ...

Details


Source

http://www.antarctica.ac.uk/met/jds/ozone/data/ZN0Z.DAT

References


Examples

data(ozonemonthly)
## notice decay in the 80s
plot(ozonemonthly)
## October plot shows dramatic swing
monthplot(ozonemonthly)
paradise

*Annual snowfall at Paradise Ranger Station, Mount Ranier*

**Description**

Annual snowfall (from July 1 to June 30th) measured at Paradise ranger station at Mount Ranier Washington.

**Usage**

data(paradise)

**Format**

The data is stored as a zoo class object. The time index refers to the year the snowfall begins.

**Details**

Due to its rapid elevation gain, and proximity to the warm moist air of the Pacific Northwest record amounts of snow can fall on Mount Ranier. This data set shows the fluctuations.

**Source**

[http://www.nps.gov/mora/current/weather.htm](http://www.nps.gov/mora/current/weather.htm)

**Examples**

```r
require(zoo)
data(paradise)
range(paradise, na.rm=TRUE)
plot(paradise)
```

---

pi2000

*first 2000 digits of pi*

**Description**

first 2000 digits of pi

**Usage**

data(pi2000)

**Format**

The format is: `num [1:2000] 3 1 4 1 5 9 2 6 5 3 ...`
primes

Source

Examples
data(pi2000)
chisq.test(table(pi2000))

primes Primes numbers less than 2003

Description
Prime numbers between 1 and 2003.

Usage
data(primes)

Format
The format is: num [1:304] 2 3 5 7 11 13 17 19 23 29 ...

Source

Examples
data(primes)
diff(primes)

puerto Incomes for Puerto Rican immigrants to Miami

Description
Incomes for Puerto Rican immigrants to Miami.

Usage
data(puerto)

Format
The format is: num [1:50] 150 280 175 190 305 380 290 300 170 315 ...
QQplot

**Source**

From Kitchens Exploring Statistics

**Examples**

```r
data(puerto)
hist(puerto)
```

---

**QQplot**

*Creates a qqplot with shaded density estimate*

**Description**

Creates a qqplot of two variables along with graphs of their densities, shaded so that the corresponding percentiles are clearly matched up.

**Usage**

```r
QQplot(x, y, n = 20, xsf = 4, ysf = 4, main = "qqplot", xlab = deparse(substitute(x)),
ylab = deparse(substitute(y)), pch = 16, pcol = "black", shade = "gray", ...)
```

**Arguments**

- `x` The x variable
- `y` The y variable
- `n` number of points to plot in qqplot.
- `xsf` scale factor to adjust size of x density graph
- `ysf` scale factor to adjust size of y density graph
- `main` title
- `xlab` label for x axis
- `ylab` label for y axis
- `pch` plot character for points in qqplot
- `pcol` color of plot character
- `shade` shading color
- `...` extra arguments passed to `plot.window`

**Details**

Shows density estimates for the two samples in a qqplot. Meant to make this useful plot more transparent to first-time users of quantile-quantile plots.

This function has some limitations: the scale factor may need to be adjusted; the code to shade only shaded trapezoids, and does not completely follow the density.
Value

Produces a graphic

Author(s)

John Verzani

See Also

qqplot, qqnorm

Examples

```r
x = rnorm(100)
y = rt(100, df=3)
qqplot(x, y)
```

---

**rat**

*Survival times of 20 rats exposed to radiation*

Description

Survival times of 20 rats exposed to radiation

Usage

data(rat)

Format

The format is: num [1:20] 152 152 115 109 137 88 94 77 160 165 ...

Source

From Kitchens Exploring Statistics

Examples

data(rat)
hist(rat)
Description

A simulated dataset on reaction time to an external event for subject using cell phones.

Usage

data(reaction.time)

Format

A data frame with 60 observations on the following 4 variables.

- **age**  Age of participant coded as 16-24 or 25+
- **gender**  Male or Female
- **control**  Code to indicate if subject is using a cell phone "T" or is in the control group "C"
- **time**  Time in seconds to react to external event

Details

Several studies indicate that cell phone usage while driving can effect reaction times to external events. This dataset uses simulated data based on values from the NHTSA study "The Influence of the Use of Mobile Phones on Driver Situation Awareness".

Source

The NHTSA study may be found at [http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/2.PDF](http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/2.PDF)

References

This study and others are linked from the web page [http://www.accidentreconstruction.com/research/cellphones/](http://www.accidentreconstruction.com/research/cellphones/).

Examples

data(reaction.time)
boxplot(time ~ control, data = reaction.time)
redrum

**Description**

Simulated length-at-age data for the red drum.

**Usage**

data(redrum)

**Format**

A data frame with 100 observations on the following 2 variables.

- **age**  age
- **length**  a numeric vector

**Details**

This data is simulated from values reported in a paper by Porch, Wilson and Nieland titled "A new growth model for red drum (Sciaenops ocellaus) that accommodates seasonal and ontogenic changes in growth rates" which appeared in *Fishery Bulletin* 100(1) ([http://fishbull.noaa.gov/1001/por.pdf](http://fishbull.noaa.gov/1001/por.pdf)). They attribute the data to Beckman et. al and say it comes from measurements in the Northern Gulf of Mexico, between September 1985 and October 1998.

**Examples**

data(redrum)
plot(length ~ age, redrum)

---

salmon.rate

**Simulated Data on Rate of Recruitment for Salmon**

**Description**

The Ricker model is used to model the relationship of recruitment of a salmon species versus the number of spawners. The model has two parameters, a rate of growth at small numbers and a decay rate at large numbers. This data set is simulated data for 83 different recordings using parameters found in a paper by Chen and Holtby.

**Usage**

data(salmon.rate)
Format

The format is: 83 numbers on decay rates.

Details

The Ricker model for recruitment modeled by spawner count

\[ R_t = S_t e^{a-bS_t} \]

The parameter \( b \) is a decay rate for large values of \( S \). In the paper by Chen and Holtby, they studied 83 datasets and found that \( b \) is log-normally distributed. The data is simulated from their values to illustrate a log normal distribution.

Source

These values are from D.G. Chen and L. Blair Holtby, “A regional meta-model for stock recruitment analysis using an empirical Bayesian approach”, found at http://www.iphc.washington.edu/.

Examples

data(salmon.rate)
hist(log(salmon.rate))

salmonharvest  Salmon harvest in Alaska from 1980 to 1998

Description

A data set of unofficial tallies of salmon harvested in Alaska between the years 1980 and 1998. The units are in thousands of fish.

Usage

data(salmonharvest)

Format

A multiple time series object with yearly sampling for the five species Chinook, Sockeye, Coho, Pink, and Chum.

Source

This data was found at http://seamarkets.alaska.edu/ak_harv_fish.htm

Examples

data(salmonharvest)
acf(salmonharvest)
Description

A data frame containing data on health behaviour for school-aged children.

Usage

data(samhda)

Format

A data frame with 600 observations on the following 9 variables.

wt  A numeric weight used in sampling

gender  1=Male, 2=Female, 7=not recorded

grade  1 = 6th, 2 = 8th, 3 = 10th

live.with.father  1 = Y, 2 = N

amt.smoke  Amount of days you smoked cigarettes in last 30. 1 = all 30, 2 = 20-29, 3 = 10-19, 4 = 6-9, 5 = 3-5, 6 = 1-2, 7=0

alcohol  Have you ever drank alcohol, 1 = Y, 2 = N

amt.alcohol  Number of days in last 30 in which you drank alcohol

marijuana  Ever smoke marijuana. 1 = Y, 2= N

amt.marijuana  Number of days in lst 30 that marijuana was used. 1 = Never used, 2 = all 30, 3 = 20-29, 4 = 10-19, 5 = 6-9, 6 = 3-5, 7 = 1-2, 8 =Used, but not in last 30 days

Details

A data frame containing data on health behaviour for school-aged children.

Source

This data is sampled from the data set "Health Behavior in School-Aged Children, 1996: [United States]" collected by the World Health Organization, http://www.icpsr.umich.edu/. It is available at the Substance Abuse and Mental Health Data Archive (SAMHDA). Only complete cases are given.

Examples

data(samhda)
attach(samhda)
table(amt.smoke)
SAT data with expenditures

Description

This dataset contains variables that address the relationship between public school expenditures and academic performance, as measured by the SAT.

Usage

data(SAT)

Format

A data frame with variables state, expend (expenditure per pupil), ratio (pupil/teacher ratio); salary (average teacher salary; percentage of SAT takers; verbal (verbal score); math (math score); total (average total).

Source

http://www.amstat.org/publications/jse/datasets/sat.txt

References

This data comes from http://www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm. It is also included in the mosaic package and commented on at http://sas-and-r.blogspot.com/2012/02/example-928-visualizing-simpsons.html. The variables are described at http://www.amstat.org/publications/jse/datasets/sat.txt.

The author references the original source: The variables in this dataset, all aggregated to the state level, were extracted from the 1997 Digest of Education Statistics, an annual publication of the U.S. Department of Education. Data from a number of different tables were downloaded from the National Center for Education Statistics (NCES) website (Available at: http://nces01.ed.gov/pubs/digest97/index.html) and merged into a single data file.

Examples

data(SAT)
scatter.with.hist  Scatterplot with histograms

Description

Draws a scatterplot of the data, and histogram in the margins. A trend line can be added, if desired.

Usage

scatter.with.hist(x, y,
    hist.col = gray(0.95),
    trend.line = "lm",
    ...
)

Arguments

x         numeric predictor
y         numeric response variables
hist.col  color for histogram
trend.line Draw a trend line using lm, supsmu or lowess. Use NULL for none.
...       Passed to plot command for scatterplot

Value

Draws the graphic. No return value.

Author(s)

John Verzani

References

This example comes from the help page for layout.

See Also

layout

Examples

data(emissions)
attach(emissions)
scatter.with.hist(perCapita, CO2)
Description

Distribution and point values of letters in Scrabble.

Usage

data(scrabble)

Format

A data frame with 27 observations on the following 3 variables.

- **piece**: Which piece
- **points**: point value
- **frequency**: Number of pieces

Details

Scrabble is a popular board game based on forming words from the players’ pieces. These consist of letters drawn from a pile at random. The game has a certain frequency of letters given by this data. These match fairly well with the letter distribution of the English language.

Examples

data(scrabble)

```r
## perform chi-squared analysis on long string. Is it in English?
quote = " R is a language and environment for statistical computing \\nand graphics. It is a GNU project which is similar to the S language \\nand environment which was developed at Bell Laboratories (formerly \\nAT&T, now Lucent Technologies) by John Chambers and colleagues. R \\can be considered as a different implementation of S. There are \\nsome important differences, but much code written for S runs \\unaltered under R."
quote.lc = tolower(quote)
quote = unlist(strsplit(quote.lc,""))
ltr.dist = sapply(c(letters,""), function(x) sum(quote == x))
chisq.test(ltr.dist,,scrabble$freq)
```
simple.chutes

simulate a chutes and ladder game

Description

This function will simulate a chutes and ladder game. It returns a trajectory for a single player. Optionally it can return the transition matrix which can be used to speed up the simulation.

Usage

simple.chutes(sim=FALSE, return.cl=FALSE, cl=make.cl())

Arguments

sim

Set to TRUE to return a trajectory.

return.cl

Set to TRUE to return a transition matrix

cl

set to the chutes and ladders transition matrix

Details

To make a chutes and ladders trajectory

simple.chutes(sim=TRUE)

To return the game board

simple.chutes(return.cl=TRUE)

when doing a lot of simulations, it may be best to pass in the game board

cl <- simple.chutes(return.cl=TRUE) simple.chutes(sim=TRUE,cl)

Value

returns a trajectory as a vector, or a matrix if asked to return the transition matrix

Author(s)

John Verzani

References

board from http://www.ahs.uwaterloo.ca/~musuem/vexhibit/Whitehill/snakes/snakes.gif

Examples

plot(simple.chutes(sim=TRUE))
simple.densityplot  Plots densities of data

Description

Allows one to compare empirical densities of different distributions in a simple manner. The density is used as graphs with multiple histograms are too crowded. The usage is similar to side-by-side boxplots.

Usage

simple.densityplot(x, ...)

Arguments

x  

x may be a sequence of data vectors (eg. x,y,z), a data frame with numeric column vectors or a model formula

...  

You can pass in a bandwidth argument such as bw="SJ". See density for details. A legend will be placed for you automatically. To override the positioning set do.legend="manual". To skip the legend, set do.legend=FALSE.

Value

Makes a plot

Author(s)

John Verzani

References

Basically a modified boxplot function. As well it should be as it serves the same utility: comparing distributions.

See Also

boxplot, simple.violinplot, density

Examples

## taken from boxplot
## using a formula
data(InsectSprays)
simple.densityplot(count ~ spray, data = InsectSprays)
## on a matrix (data frame)
mat <- cbind(Uni05 = (1:100)/21, Norm = rnorm(100),
    T5 = rt(100, df = 5), Gam2 = rgamma(100, shape = 2))
simple.densityplot(data.frame(mat))
Description

Simply plots histogram, boxplot and normal plot for experimental data analysis.

Usage

```
simple.eda(x)
```

Arguments

- `x` a vector of data

Value

Just does the plots. No return value

Author(s)

John Verzani

References

Inspired by S-Plus documentation

See Also

`hist, boxplot, qnorm`

Examples

```
x <- rnorm(100, 5, 10)
simple.eda(x)
```
**Description**

This makes 3 graphs to check for serial correlation in data. The graphs are a sequential plot (i vs \(X_i\)), a lag plot (plotting \(X_i\) vs \(X_i\) where \(k=1\) by default) and an autocorrelation plot from the times series ("ts") package.

**Usage**

```r
simple.eda.ts(x, lag=1)
```

**Arguments**

- `x` a univariate vector of data
- `lag` a lag to give to the lag plot

**Value**

Makes the graph with 1 row, 3 columns

**Author(s)**

John Verzani

**References**

see [http://www.itl.nist.gov/div898/handbook/eda/section3/eda34.htm](http://www.itl.nist.gov/div898/handbook/eda/section3/eda34.htm) for more information on these and other plots for exploratory data analysis.

**Examples**

```r
## The function is currently defined as

## look for no correlation
x <- rnorm(100); simple.eda.ts(x)
## you will find correlation here
simple.eda.ts(cumsum(x))
```
simple.fancy.stripchart

Makes a fancier strip chart: plots means and a line

Description
Not much, just hides some ugly code

Usage
simple.fancy.stripchart(l)

Arguments
1 A list with each element to be plotted with a stripchart

Value
Creates the plot

Author(s)
John Verzani

See Also
stripchart

Examples
x = rnorm(10); y = rnorm(10, 1)
simple.fancy.stripchart(list(x=x, y=y))

simple.freqpoly

Simply plot histogram and frequency polygon

Description
Simply plot histogram and frequency polygon. Students do not need to know how to add lines to a histogram, and how to extract values.

Usage
simple.freqpoly(x, ...)

simple.hist.and.boxplot

**Arguments**

- `x` a vector of data
- `...` arguments passed onto histogram

**Value**

returns just the plot

**Author(s)**

John Verzani

**See Also**

hist,density

**Examples**

```r
x <- rt(100,4)
simple.freqpoly(x)
```

---

**simple.hist.and.boxplot**

*A function to plot both a histogram and a boxplot*

**Description**

Simple function to plot both histogram and boxplot to compare

**Usage**

```r
simple.hist.and.boxplot(x, ...)
```

**Arguments**

- `x` vector of univariate data
- `...` Arguments passed to the hist function

**Value**

Just prints the two graphs

**Author(s)**

John Verzani
simple.lag

See Also

hist, boxplot, layout

Examples

x <- rnorm(100)
simple.hist.and.boxplot(x)

simple.lag  applies function to moving subsets of a data vector

Description

Used to apply a function to subsets of a data vector. In particular, it is used to find moving averages over a certain "lag" period.

Usage

simple.lag(x, lag, FUN = mean)

Arguments

x a data vector
lag the lag amount to use.
FUN a function to apply to the lagged data. Defaults to mean

Details

The function FUN is applied to the data x[(i-lag):i] and assigned to the (i-lag)th component of the return vector. Useful for finding moving averages.

Value

returns a vector.

Author(s)

Provided to R help list by Martyn Plummer

See Also

filter
Examples

```r
## find a moving average of the dow daily High
data(dowdata)
lag = 50; n = length(dowdata$High)
plot(simple.lag(dowdata$High,lag),type="l")
lines(dowdata$High[lag:n])
```

---

**simple.lm**  
_Simplify usage of lm_

Description

Simplify usage of lm by avoiding model notation, drawing plot, drawing regression line, drawing confidence intervals.

Usage

```r
simple.lm(x, y, show.residuals=FALSE, show.ci=FALSE, conf.level=0.95,pred=)
```

Arguments

- `x` The predictor variable
- `y` The response variable
- `show.residuals` set to TRUE to plot residuals
- `show.ci` set to TRUE to plot confidence intervals
- `conf.level` if `show.ci=TRUE` will plot these CI's at this level
- `pred` values of the x-variable for prediction

Value

returns plots and an instance of lm, as though it were called `lm(y ~ x)`

Author(s)

John Verzani

See Also

lm
### Examples

```r
## on simulated data
x <- 1:10
y <- 5*x + rnorm(10, 0, 1)
tmp <- simple.lm(x, y)
summary(tmp)

## predict values
simple.lm(x, y, pred=c(5, 6, 7))
```

---

**Description**

Do simple sign test like wilcox.test without ranking. Just computes two-sided p-value, no confidence interval is given.

**Usage**

```r
simple.median.test(x, median=NA)
```

**Arguments**

- `x`: A data vector
- `median`: The value of median under the null hypothesis

**Details**

Unlike wilcox.test, this tests the null hypothesis that the median is specified against the two-sided alternative. For illustration purposes only.

**Value**

Returns the p value.

**Author(s)**

John Verzani

**See Also**

wilcox.test

**Examples**

```r
x <- c(12, 2, 17, 25, 52, 8, 1, 12)
simple.median.test(x, 20)
```
simple.scatterplot

Simple scatter plot of x versus y with histograms of each

Description

Shows scatterplot of x vs y with histograms of each on sides of graph. As in the example from layout.

Usage

```r
simple.scatterplot(x, y, …)
```

Arguments

- `x` data vector
- `y` data vector
- `…` passed to plot command

Value

Returns the plot

Author(s)

John Verzani

See Also

layout

Examples

```r
x<-sort(rnorm(100))
y<-sort(rt(100,3))
simple.scatterplot(x,y)
```
**simple.sim**  
*Simplify the process of simulation*

**Description**

'simple.sim' is intended to make it a little easier to do simulations with R. Instead of writing a for loop, or dealing with column or row sums, a student can use this "simpler" interface.

**Usage**

```r
simple.sim(no.samples, f, ...)
```

**Arguments**

- `no.samples` How many samples do you wish to generate
- `f` A function which generates a single random number from some distributions. simple.sim generates the rest.
- `...` parameters passed to `f`. It does not like named parameters.

**Details**

This is simply a wrapper for a for loop that uses the function `f` to create random numbers from some distribution.

**Value**

returns a vector of size `no.samples`

**Note**

There must be a 1000 better ways to do this. See `replicate` or `sapply` for example.

**Author(s)**

John Verzani

**Examples**

```r
## First shows trivial (and very unnecessary usage)
## define a function `f` and then simulate
f <- function() rnorm(1)  # create a single random real number
sim <- simple.sim(100, f)  # create 100 random normal numbers
hist(sim)

## what does range look like?
f <- function(n, mu=0, sigma=1) {
  tmp <- rnorm(n, mu, sigma)
  max(tmp) - min(tmp)
}
simple.violinplot

Plots violinplots instead of boxplots

Description

This function serves the same utility as side-by-side boxplots, only it provides more detail about the different distribution. It plots violinplots instead of boxplots. That is, instead of a box, it uses the density function to plot the density. For skewed distributions, the results look like "violins". Hence the name.

Usage

simple.violinplot(x, ...)

Arguments

  x

  Either a sequence of variable names, or a data frame, or a model formula

  ...  

  You can pass arguments to polygon with this. Notably, you can set the color to red with col='red', and a border color with border='blue'

Value

  Returns a plot.

Author(s)

  John Verzani

References

  This is really the boxplot function from R/base with some minor adjustments

See Also

  boxplot, simple.densityplot
**simple.z.test**  
Implement basic z-test for illustrative purposes

**Description**  
Implements a z-test similar to the t.test function

**Usage**  
simple.z.test(x, sigma, conf.level=0.95)

**Arguments**  
- **x**: A data vector  
- **sigma**: the known variance  
- **conf.level**: Confidence level for confidence interval

**Value**  
Returns a confidence interval for the mean

**Author(s)**  
Joh Verzani

**See Also**  
t.test, prop.test

**Examples**  
```r  
x <- rnorm(100) ; x[101:150] <- rnorm(50,5)  
simple.violinplot(x,col="brown")  
f <- factor(rep(1:5,30))  
## make a quintet. Note also choice of bandwidth  
simple.violinplot(x-f,col="brown",bw="SJ")
```
Description

Judges scores from the disputed ice skating competition at the 2002 Winter olympics

Usage

data(skateranks)

Format

A data frame with 20 observations on the following 11 variables.

Name a factor with levels Berankova/Diabola Berezhnaya/Sikharulidze Bestnadigova/Bestandif Chuvaeva/Palamarchuk Cobisi/DePra Ina/Zimmerman Kautz/Jeschke Krasitseva/Znachkov Langlois/Archetto Lariviere/Faustino Pang/Tong Petrova/Tikhonov Ponomareva/SWviridov Savchenko/Morozov Scott/Dulebohn Sele/Pelletier Shen/Zhao Totmianina/Marinin Zagorska/Siudek Zhang/Zhang

Country a factor with levels Armenia Canada China Czech Germany Italy Poland Russia Slovakia US Ukraine Uzbekistan

Russia a numeric vector

China a numeric vector

US a numeric vector

France a numeric vector

Poland a numeric vector

Canada a numeric vector

Ukraine a numeric vector

Germany a numeric vector

Japan a numeric vector

Examples

data(skateranks)
**slc**  
*Sodium-Lithium countertransport*

**Description**  
Sodium-Lithium countertransport

**Usage**  
data(sl)

**Format**  
The format is: num [1:190] 0.467 0.430 0.192 0.192 0.293 ...

**Source**  
From Kitchens’ Exploring Statistics

**Examples**

data(sl)
hist(sl)

---

**smokyph**  
*Water pH levels at 75 water samples in the Great Smoky Mountains*

**Description**  
Water pH levels at 75 water samples in the Great Smoky Mountains

**Usage**

data(smokyph)

**Format**

This data frame contains the following columns:

- **waterph** a numeric vector
- **elev** a numeric vector
- **code** a numeric vector

**Source**

From Kitchens’ Exploring Statistics
Examples

```
data(smokyph)
plot(smokyph$eleve,smokyph$waterph)
```

---

snacks

*Snack data from the USDA*

**Description**

Subset of SR26 data on nutrients compiled by the USDA.

**Usage**

```
data(snacks)
```

**Format**

A data frame with some nutrition variables

**Source**

This data came from the SR26 data set found at [http://www.ars.usda.gov/Services/docs.htm?docid=8964](http://www.ars.usda.gov/Services/docs.htm?docid=8964).

**Examples**

```
data(snacks)
```

---

south

*Murder rates for 30 Southern US cities*

**Description**

Murder rates for 30 Southern US cities

**Usage**

```
data(south)
```

**Format**

The format is: num [1:30] 12 10 10 13 12 14 7 16 18 ... 

**Source**

From Kitchens’ Exploring Statistics
**Examples**

```r
data(south)
hist(south)
```

---

**southernosc**  
Southern Oscillations

**Description**

The southern oscillation is defined as the barometric pressure difference between Tahiti and the Darwin Islands at sea level. The southern oscillation is a predictor of el nino which in turn is thought to be a driver of world-wide weather. Specifically, repeated southern oscillation values less than -1 typically defines an el nino.

**Usage**

```r
data(southernosc)
```

**Format**

The format is: Time-Series [1:456] from 1952 to 1990: -0.7 1.3 0.1 -0.9 0.8 1.6 1.7 1.4 1.4 1.5 ...

**Source**


**References**


**Examples**

```r
data(southernosc)
plot(southernosc)
```
**sp500.excess**  
*Excess returns of S&P 500*

---

**Description**

Excess returns of S&P 500. These are defined as the difference between the series and some riskless asset.

**Usage**

```r
data(sp500.excess)
```

**Format**

The format is: Time-Series [1:792] from 1929 to 1995: 0.0225 -0.044 -0.0591 0.0227 0.0077 0.0432 0.0455 0.0171 0.0229 -0.0313 ...

**Source**

This data set is used in Tsay, Analysis of Financial Time Series. It was downloaded from www.gsb.uchicago.edu/fac/ruey.tsay/teaching/fts. The fSeries package may also contain this data set.

**Examples**

```r
data(sp500.excess)
plot(sp500.excess)
```

---

**Split.zoo**  
*Add split method for zoo objects*

---

**Description**

Splits zoo objects by a grouping variable ala split(). Each univariate series is turned into a multivariate zoo object. If the original series is multivariate, the output is a list of multivariate zoo objects.

**Usage**

```r
Split.zoo(x, f)
```

**Arguments**

- `x`  
an univariate or multivariate zoo object

- `f`  
A grouping variable of the same length of `x`. A warning is given is `length(f)` is not the same as index size of `x`
squareplot

Create a squareplot alternative to a segmented barplot

Description

Create a squareplot as an alternative to a segmented barplot. Useful when the viewer is interested in exact counts in the categories. A squareplot is often used by the New York Times. A grid of squares is presented with each color representing a different category. The colors appear contiguously reading top to bottom, left to right. The colors segment the graph as a segmented bargraph, but the squares allow an interested reader to easily tally the counts.

Usage

```r
squareplot(x, col = gray(seq(0.5, 1, length = length(x))), border = NULL, nrows = ceiling(sqrt(sum(x))), ncols = ceiling(sum(x)/nrows), ...)
```

Arguments

- `x`: a vector of counts
- `col`: a vector of colors
- `border`: border color passed to `polygon`
- `nrows`: number of rows
- `ncols`: number of columns
- `...`: passed to `title`
Value

Creates the graph, but has no return value.

Author(s)

John Verzani

References


Examples

```r
## A Roger Clemens Cy Young year -- roids?
squareplot(c(21,7,6),col=c("blue","green","white"))
```

---

### stud.recs

**Student records**

Description

A simulation of student records used for placement purposes

Usage

```r
data(stud.recs)
```

Format

A data frame with 160 observations on the following 6 variables.

- **seq.1** Score on sequential 1 test
- **seq.2** Score on sequential 2 test
- **seq.3** Score on sequential 3 test
- **sat.v** SAT verbal score
- **sat.m** SAT math score
- **num.grade** grade on first math class
- **letter.grade** grade on first math class

Details

Some simulated student records for placement purposes

Examples

```r
data(stud.recs)
hist(stud.recs$sat.v)
with(stud.recs,cor(sat.v,sat.m))
```
**student.expenses**    

*Some simulated data on student expenses*

---

**Description**

Some data for possible student expenses

**Usage**

```r
data(student.expenses)
```

**Format**

A data frame of 5 variables for 10 students. All answers are coded "Y" for yes, "N" for no.

- **cell.phone** Does student have cell phone.
- **cable.tv** Does student have cable TV.
- **dial.up** Does student pay for dial-up internet access.
- **cable.modem** Does student pay for high-speed or cable modem access to internet.
- **car** Does student own a car.

**Details**

Sample dataset of students expenses.

**Examples**

```r
data(student.expenses)
attach(student.expenses)
table(dial.up,cable.modem)
```

---

**superbarplot**  

*super segmented barplot*

---

**Description**

Plot a barplot, with bars nested and ranging from a max to a minimum value. A similar graphic is used on the weather page of the *New York Times*.

**Usage**

```r
superbarplot(x, names = 1:dim(x)[2], names_height = NULL,
            col = gray(seq(0.8, 0.5, length = dim(x)[1]/2)), ...)
```

Arguments

- **x**: A matrix with each pair of rows representing a min and max for the bar.
- **names**: Place a name in each bar.
- **names_height**: Where the names should go
- **col**: What colors to use for the bars. There should be half as many specified as rows of x

... passed to `plot.window`.

Details

A similar graphic on the weather page of the *New York Times* shows bars for record highs and lows, normal highs and lows and actual (or predicted) highs or lows for 10 days of weather. This graphic succinctly and elegantly displays a wealth of information. Intended as an illustration of the `polygon` function.

Value

Returns a plot, but no other values.

Author(s)

John Verzani

References

The weather page of the *New York Times*

See Also

`squareplot`

Examples

```r
record.high <- c(95, 95, 93, 96, 98, 96, 97, 96, 95, 97)
record.low <- c(49, 47, 48, 51, 49, 48, 52, 51, 49, 52)
normal.high <- c(78, 78, 79, 79, 79, 79, 79, 80, 80, 80)
normal.low <- c(62, 62, 63, 63, 63, 64, 64, 64, 64, 64)
actual.high <- c(80, 78, 80, 68, 83, 83, 73, 75, 77, 81)
actual.low <- c(62, 65, 66, 58, 69, 63, 59, 58, 59, 60)
x <- rbind(record.low, record.high, normal.low, normal.high, actual.low, actual.high)
superbarplot(x, names = the.names)
```
tastesgreat

Does new goo taste great?

Description

Fictitious data on taste test for new goo

Usage

data(tastesgreat)

Format

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

- gender  a factor with levels Female Male
- age  a numeric vector
- enjoyed  1 if enjoyed, 0 otherwise

Details

Fictitious data on a taste test with gender and age as covariates.

Examples

data(tastesgreat)
summary(glm(enjoyed ~ gender + age, data=tastesgreat, family=binomial))

tcm1y

One-year treasury security values

Description

The yields at constant fixed maturity have been constructed by the Treasury Department, based on the most actively traded marketable treasury securities.

Usage

data(tcm1y)

Format

The format is: Time-Series [1:558] from 1953 to 2000: 2.36 2.48 2.45 2.38 2.28 2.2 1.79 1.67 1.66 1.41 ...
Source

From the tcm data set in the tseries package. Given here for convenience only. They reference http://www.federalreserve.gov/Releases/H15/data.htm

Examples

data(tcmly)
ar(diff(log(tcmly)))

Description

Simulated measurements of temperature and salinity in the center of 'Eddy Juggernaut', a huge anticyclone (clockwise rotating) Loop Current Ring in the Gulf of Mexico. The start date is October 18, 1999.

Usage

data(tempsalinity)

Format

The data is stored as multivariate zooreg object with variables longitude, latitude, temperature (Celsius), and salinity (psu - practical salinity units http://www.toptotop.org/climate/psu.php).

Details

The temperature salinity profile of body of water can be characteristic. This data shows a change in the profile in time as the eddy accumulates new water.

Source

Data from simulation by Andrew Poje.

Examples

data(tempsalinity)
if(require(zoo)) {
plot(tempsalinity[,3:4])
## override plot.zoo method
plot.default(tempsalinity[,3:4])
abline(lm(salinity ~ temperature, tempsalinity, subset = 1:67))
abline(lm(salinity ~ temperature, tempsalinity, subset = -1:67))
}
What age is too young for a male to date a female?

Description

In U.S. culture, an older man dating a younger woman is not uncommon, but when the age difference becomes too great it may seem to some to be unacceptable. This data set is a survey of 10 people with their minimum age for an acceptable partner for a range of ages for the male. A surprising rule of thumb (in the sense that someone took the time to figure this out) for the minimum is half the age plus seven. Does this rule hold for this data set?

Usage

```r
data(too.young)
```

Format

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

- **Male**  a numeric vector
- **Female** a numeric vector

Examples

```r
data(too.young)
lm(Female ~ Male, data=too.young)
```

---

Burt’s IQ data for twins

Description

IQ data of Burt on identical twins that were separated near birth.

Usage

```r
data(twins)
```

Format

A data frame with 27 observations on the following 3 variables.

- **Foster**  IQ for twin raised with foster parents
- **Biological** IQ for twin raised with biological parents
- **Social**  Social status of biological parents
Source

This data comes from the R package that accompanies Julian Faraway’s notes *Practical Regression and Anova in R* (now a book).

Examples

data(twins)
plot(Foster ~ Biological, twins)

u2

*Song and lengths for U2 albums*

Description

Song titles and lengths of U2 albums from 1980 to 1997.

Usage

data(u2)

Format

The data is stored as a list with names. Each list entry correspond to an album stored as a vector. The values of the vector are the song lengths in seconds and the names are the track titles.

Source

http://www.u2station.com/u2ography.html

Examples

data(u2)
sapply(u2,mean) # average track length
max(sapply(u2,max)) # longest track length
sort(unlist(u2)) # lengths in sorted order
urchin.growth

Data on growth of sea urchins

Description
Data on growth of sea urchins.

Usage
data(urchin.growth)

Format
A data frame with 250 observations on the following 2 variables.

age  Estimated age of sea urchin
size  Measurement of size

Details
Data is sampled from a data set that accompanies the thesis of P. Grosjean.

Source
Thesis was found at http://www.sciviews.org/_pgrosjean.

Examples
data(urchin.growth)
plot(jitter(size) ~ jitter(age), data=urchin.growth)

vacation

vacation days

Description
vacation days

Usage
data(vacation)

Format
The format is: num [1:35] 23 12 10 34 25 16 27 18 28 13 ...
**Source**

From Kitchens’ Exploring Statistics

**Examples**

```r
data(vacation)
hist(vacation)
```

---

**violinplot**  
*Plots violinplots instead of boxplots*

---

**Description**

This function serves the same utility as side-by-side boxplots, only it provides more detail about the different distribution. It plots violinplots instead of boxplots. That is, instead of a box, it uses the density function to plot the density. For skewed distributions, the results look like “violins”. Hence the name.

**Usage**

```r
violinplot(x, ...)
```

**Arguments**

- `x`  
  Either a sequence of variable names, or a data frame, or a model formula
- `...`  
  You can pass arguments to polygon with this. Notably, you can set the color to red with `col='red'`, and a border color with `border='blue'`

**Value**

Returns a plot.

**Author(s)**

John Verzani

**References**

This is really the boxplot function from R/base with some minor adjustments

**See Also**

`boxplot`, `densityplot`
Examples

```r
## make a "violin"
x <- rnorm(100) ; x[101:150] <- rnorm(50,5)
violinplot(x,col="brown")
f <- factor(rep(1:5,30))
## make a quintet. Note also choice of bandwidth
violinplot(x-f,col="brown",bw="SJ")
```

---

**watertemp**

*Temperature measurement of water at 85m depth*

---

**Description**

Water temperature measurements at 10 minute intervals at a site off the East coast of the United States in the summer of 1974.

**Usage**

```r
data(watertemp)
```

**Format**

A zoo class object with index stored as POSIXct elements. The measurements are in Celsius.

**Source**


**Examples**

```r
if(require(zoo)) {
data(watertemp)
plot(watertemp)
acf(watertemp)
acf(diff(watertemp))
}
```
wellbeing

A random sample of Wake County, North Carolina residential real estate plots

Description

This data set comes from a JSE article http://www.amstat.org/publications/jse/v16n3/woodard.pdf by Roger Woodard. The data is described by: The information for this data set was taken from a Wake County, North Carolina real estate database. Wake County is home to the capital of North Carolina, Raleigh, and to Cary. These cities are the fifteenth and eighth fastest growing cities in the USA respectively, helping Wake County become the ninth fastest growing county in the country. Wake County boasts a 31.18 of approximately 823,345 residents. This data includes 100 randomly selected residential properties in the Wake County registry denoted by their real estate ID number. For each selected property, 11 variables are recorded. These variables include year built, square feet, adjusted land value, address, et al.

Usage

data(wchomes)

Format

a data frame

Source

http://www.amstat.org/publications/jse/v16n3/woodard.xls

References


Examples

data(wchomes)

wellbeing

What makes us happy?

Description

Correlated data on what makes us happy

Usage

data(wellbeing)
Format

A data frame with data about what makes people happy (well being) along with several other co-
variates

Source

Found from http://prcweb.co.uk/lab/what-makes-us-happy/.

References

org/

Examples

data(wellbeing)

---

yahoo.get.hist.quote  Download stock data from Yahoo!

Description

Downloads stock data from Yahoo!

Usage

yahoo.get.hist.quote(instrument = "gspc",
destfile = paste(instrument, ".csv", sep = ""),
start, end, quote = c("Open", "High", "Low", "Close"),
adjusted = TRUE, download = TRUE,
origin = "1970-01-01", compression = "d")

Arguments

instrument  Ticker symbol as character string.
destfile  Temporary file for storage
start  Date to start. Specified as "2005-12-31"
end  Date to end
quote  Any/All of "Open", "High", "Low", "Close"
adjusted  Adjust for stock splits, dividends. Defaults to TRUE
download  Download the data
origin  Dates are recorded in the number of days since the origin. A value of "1970-01-
01" is the default. This was changed from "1899-12-30".
compression  Passed to yahoo
Details

Goes to chart.yahoo.com and downloads the stock data. By default returns a multiple time series of class mts with missing days padded by NAs.

Value

A multiple time series with time measureing the number of days since the value specified to origin.

Author(s)

Daniel Herlemont <dherlemont@yats.com>

References

This function was found on the mailling list for R-SIG finance

See Also

yahoo.get.hist.quote in the tseries package

yellowfin

Yellow fin tuna catch rate in Tropical Indian Ocean

Description

Mean catch rate of yellow fin tuna in Tropical Indian Ocean for the given years.

Usage

data(yellowfin)

Format

A data frame with 49 observations on the following 2 variables.

  year  The year
  count  Mean number of fish per 100 hooks cast

Details

Estimates for the mean number of fish caught per 100 hooks are given for a number of years. This can be used to give an estimate for the size, or biomass, of the species during these years assuming the more abundant the fish, the larger the mean. In practice this assumption is viewed with a wide range of attitudes.

Source

References

See also http://www.soest.hawaii.edu/PFRP/large_pelagic_predators.html for rebuttals to the Myers and Worm article.

Examples

data(yellowfin)
plot(yellowfin)
Index

*Topic aplot
  simple.hist.and.boxplot, 94
*Topic datagen
  simple.sim, 99
*Topic datasets
  age.universe, 5
  aid, 6
  alaska.pipeline, 6
  alltime.movies, 7
  aosat, 8
  arctic.oscillations, 9
  babies, 10
  babyboom, 11
  batting, 12
  baycheck, 13
  best.times, 14
  blood, 15
  breakdown, 16
  bright.stars, 16
  brightness, 17
  bumpers, 18
  BushApproval, 18
  bycatch, 19
  cabinet, 20
  camp, 21
  cancer, 22
  carbon, 23
  carsafety, 23
  central.park, 24
  central.park.cloud, 25
  ceo2013, 26
  cfb, 26
  chicken, 27
  chips, 28
  co2emiss, 29
  coins, 29
  coldvermont, 30
  corn, 31
  crime, 32
  deflection, 33
  diamond, 35
  divorce, 36
  dottodot, 37
  dowdata, 38
  dvdsales, 39
  emissions, 39
  ewr, 40
  exec.pay, 41
  fat, 41
  father.son, 43
  female.inc, 44
  firstchi, 45
  five.yr.temperature, 45
  florida, 46
  galileo, 47
  galton, 48
  gap, 49
  gasprices, 49
  goalspergame, 51
  google, 51
  grades, 52
  grip, 52
  hall.fame, 53
  healthy, 55
  heartrate, 56
  home, 56
  homedata, 57
  homeprice, 58
  homework, 59
  HUMMER, 59
  income_percentiles, 60
  iq, 60
  kid.weights, 61
  KSI, 62
  last.tie, 62
  lawsuits, 63
  lorem, 64
  malpract, 64
INDEX

mandms, 65
math, 66
maydow, 66
Medicare, 67
midsize, 68
MLBAttend, 68
movie_data_2011, 70
movies, 69
mw.ages, 70
nba.draft, 71
nisdc, 72
normtemp, 72
npdb, 73
nym.NRPPR, 74
ObamaApproval, 75
OBP, 76
oral.lesion, 76
ozonemonthly, 77
paradise, 78
pi2000, 78
primes, 79
puerto, 79
rat, 81
reaction.time, 82
reddrum, 83
salmon.rate, 83
salmon.harvest, 84
samhda, 85
SAT, 86
scrabble, 88
skateranks, 102
slc, 103
smokyph, 103
snacks, 104
south, 104
southernosc, 105
sp500.excess, 106
stud.recs, 108
student.expenses, 109
tastesgreat, 111
tcmly, 111
tempsalinity, 112
too.young, 113
twins, 113
u2, 114
urchin.growth, 115
vacation, 115
watertemp, 117

wchomes, 118
wellbeing, 118
yellowfin, 120
*Topic dynamic
demos, 33
*Topic hplot
DOTTplot, 36
scatter.with.hist, 87
simple.eda, 91
simple.fancy.stripchart, 93
simple.freqpoly, 93
simple.scatterplot, 98
squareplot, 107
superbarplot, 109
*Topic htest
simple.median.test, 97
*Topic misc
QQplot, 80
Split.zoo, 106
yahoo.get.hist.quote, 119
*Topic multivariate
DensityPlot, 34
scatter.with.hist, 87
simple.densityplot, 90
simple.violinplot, 100
superbarplot, 109
violinplot, 116
*Topic regression
simple.lm, 96
*Topic ts
simple.lag, 95
*Topic univar
simple.chutes, 89
simple.eda, 91
simple.eda.ts, 92
simple.freqpoly, 93
simple.hist.and.boxplot, 94
simple.lag, 95
simple.sim, 99
simple.z.test, 101
squareplot, 107
*Topic utilities
answers, 8
confint.htest, 31
getAnswer, 50
headtail, 54

age.universe, 5
aid, 6
alaska.pipeline, 6
alltime.movies, 7
answers, 8
aosat, 8
arctic.oscillations, 9
babies, 10
babyboom, 11
batting, 12
baycheck, 13
best.times, 14
blood, 15
boxplot, 34, 90
breakdown, 16
bright.stars, 16
brightness, 17
bumpers, 18
BushApproval, 18
bycatch, 19

cabinet, 20
camp, 21
cancer, 22
carbon, 23
carsafety, 23
central.park, 24
central.park.cloud, 25
ceo2013, 26
cfb, 26
chicken, 27
chips, 28
c02emiss, 29
coins, 29
coldvermont, 30
confint.htest, 31
corn, 31
crime, 32
deflection, 33
demos, 33
density, 34, 90
DensityPlot, 34
diamond, 35
divorce, 36
dnstypplt (simple.densityplot), 90
DOTplot, 36
dotplot, 37
DOTplt (DOTplot), 36
dottodot, 37
dowdata, 38
dvdsales, 39
emissions, 39
errata (getAnswer), 50
ewr, 40
exec.pay, 41
fat, 41
father.son, 43, 48
female.inc, 44
firstchi, 45
five.yr.temperature, 45
florida, 46
galileo, 47
galton, 48
gap, 49
gasprices, 49
getAnswer, 50
goalspergame, 51
googles, 51
grades, 52
grip, 52

hall.fame, 53
headtail, 54
healthy, 55
heartrate, 56
home, 56
homedata, 57
homeprice, 58
homework, 59
HUMMER, 59
income_percentiles, 60
iq, 60
kid.weights, 61
KSI, 62
last.tie, 62
lawsuits, 63
layout, 87
lm, 87
lorem, 64
lowess, 87
malpract, 64
mandms, 65
INDEX

math, 66
maydow, 66
Medicare, 67
midsite, 68
MLBattend, 68
movie_data_2011, 70
movies, 69
mw.ages, 70
nba.draft, 71
nsdc, 72
normtemp, 72
npdb, 73
nym.2002, 74
ObamaApproval, 75
OBP, 76
oral.lesion, 76
ozonemonthly, 77
paradise, 78
pi2000, 78
plot, 87
plot.window, 110
points, 37
polygon, 107, 110
primes, 79
puerto, 79
qqnorm, 81
QQplot, 80
qqplot, 81
rat, 81
reaction.time, 82
reddrum, 83
replicate, 99
salmon.rate, 83
salmonharvest, 84
samhda, 85
sapply, 99
SAT, 86
scatter.with.hist, 87
scrabble, 88
simple.chutes, 89
simple.densityplot, 90
simple.eda, 91
simple.eda.ts, 92
simple.fancy.stripchart, 93
simple.freqpoly, 93
simple.hist.and.boxplot, 94
simple.lag, 95
simple.lm, 96
simple.median.test, 97
simple.plot.hist.and.box
(simple.hist.and.boxplot), 94
simple.scatterplot, 98
simple.sim, 99
simple.violinplot, 90, 100
simple.z.test, 101
skateranks, 102
slc, 103
smokypth, 103
snacks, 104
south, 104
southernosc, 105
sp500.excess, 106
split, 107
Split.zoo, 106
squareplot, 107, 110
stripchart, 37
stud.rec., 108
student.expenses, 109
superbarplot, 109
supsmu, 87
tastesgreat, 111
tcmly, 111
tempsalinity, 112
text, 37
too.young, 113
twins, 113
u2, 114
urchin.growth, 115
vacation, 115
violinplot, 34, 116
vlnplt(simple.violinplot), 100
watertemp, 117
wchomes, 118
wellbeing, 118
yahoo.get.hist.quote, 119
yellowfin, 120