

# Package ‘HistData’

January 2, 2012

**Type** Package

**Title** Data sets from the history of statistics and data visualization

**Version** 0.6-12

**Date** 2011-10-12

**Author** Michael Friendly, with contributions by Stephane Dray, Hadley Wickham, James Hanley, Dennis Murphy

**Maintainer** Michael Friendly <friendly@yorku.ca>

**Description** The HistData package provides a collection of small data sets that are interesting and important in the history of statistics and data visualization. The goal of the package is to make these available, both for instructional use and for historical research. Some of these present interesting challenges for graphics or analysis in R.

**Suggests** gtools, KernSmooth, maps, ggplot2, proto, grid, reshape,plyr, lattice, ReadImages, car, gplots, sp, heplots, lattice

**License** GPL

**LazyLoad** yes

**LazyData** yes

**Repository** CRAN

**Repository/R-Forge/Project** histdata

**Repository/R-Forge/Revision** 53

**Date/Publication** 2011-10-26 18:09:33

**Depends** R (>= 2.10)

## R topics documented:

HistData-package . . . . .	2
Arbuthnot . . . . .	4
Bowley . . . . .	6
Cavendish . . . . .	7
ChestSizes . . . . .	8
CushnyPebbles . . . . .	9
Dactyl . . . . .	11
DrinksWages . . . . .	12
Fingerprints . . . . .	14
Galton . . . . .	15
Guerry . . . . .	16
Jevons . . . . .	19
Langren . . . . .	20
Macdonell . . . . .	23
Michelson . . . . .	28
Minard . . . . .	30
Nightingale . . . . .	31
OldMaps . . . . .	34
PearsonLee . . . . .	35
PolioTrials . . . . .	38
Prostitutes . . . . .	39
Pyx . . . . .	40
Quarrels . . . . .	41
Snow . . . . .	44
Wheat . . . . .	48
Yeast . . . . .	49
ZeaMays . . . . .	50
<b>Index</b>	<b>53</b>

---

HistData-package	<i>Data sets from the history of statistics and data visualization</i>
------------------	--

---

### Description

The HistData package provides a collection of data sets that are interesting and important in the history of statistics and data visualization. The goal of the package is to make these available, both for instructional use and for historical research.

### Details

Package:	HistData
Type:	Package
Version:	0.6-12
Date:	2011-10-02
License:	GPL

LazyLoad: yes  
LazyData: yes

Some of the data sets have examples which reproduce an historical graph or analysis. These are meant mainly as starters for more extensive re-analysis or graphical elaboration. Some of these present graphical challenges to reproduce in R.

Descriptions of each data set can be found using `help(DataSet)`; `example(DataSet)` will likely show applications similar to the historical use.

Data sets included in the HistData package are:

[Arbuthnot](#) Arbuthnot's data on male and female birth ratios in London from 1629-1710

[Bowley](#) Bowley's data on values of British and Irish trade, 1855-1899

[Cavendish](#) Cavendish's 1798 determinations of the density of the earth

[ChestSizes](#) Quetelet's data on chest measurements of Scottish militiamen

[CushnyPebbles](#) Cushny-Pebbles data: Soporific effects of scopolamine derivatives

[Dactyl](#) Edgeworth's counts of dactyls in Virgil's Aeneid

[DrinksWages](#) Elderton and Pearson's (1910) data on drinking and wages

[Fingerprints](#) Waite's data on Patterns in Fingerprints

[Galton](#) Galton's data on the heights of parents and their children

[Guerry](#) Data from A.-M. Guerry, "Essay on the Moral Statistics of France"

[Jevons](#) W. Stanley Jevons' data on numerical discrimination

[Langren](#) van Langren's data on longitude distance between Toledo and Rome

[Macdonell](#) Macdonell's data on height and finger length of criminals, used by Gosset (1908)

[Michelson](#) Michelson's 1879 determinations of the velocity of light

[Minard](#) Data from Minard's famous graphic map of Napoleon's march on Moscow

[Nightingale](#) Florence Nightingale's data on deaths from various causes in the Crimean War

[OldMaps](#) Latitudes and Longitudes of 39 Points in 11 Old Maps

[PearsonLee](#) Pearson and Lee's 1896 data on the heights of parents and children classified by gender

[PolioTrials](#) Polio Field Trials Data on the Salk vaccine

[Prostitutes](#) Parent-Duchatelet's time-series data on the number of prostitutes in Paris

[Pyx](#) Trial of the Pyx

[Quarrels](#) Statistics of Deadly Quarrels

[Snow](#) John Snow's map and data on the 1854 London Cholera outbreak

[Wheat](#) Playfair's data on wages and the price of wheat

[Yeast](#) Student's (1906) Yeast Cell Counts

[ZeaMays](#) Darwin's Heights of Cross- and Self-fertilized Zea May Pairs

**Author(s)**

Michael Friendly

Maintainer: Michael Friendly <friendly@yorku.ca>

**References**

Friendly, M. (2007). A Brief History of Data Visualization. In Chen, C., Hardle, W. & Unwin, A. (eds.) *Handbook of Computational Statistics: Data Visualization*, Springer-Verlag, III, Ch. 1, 1-34.

Friendly, M. & Denis, D. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. <http://datavis.ca/milestones/>

Friendly, M. & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41, 103-130.

**See Also**

[Arbuthnot](#), [Bowley](#), [Cavendish](#), [ChestSizes](#), [CushnyPebbles](#),

[Dactyl](#), [DrinksWages](#), [Fingerprints](#), [Galton](#), [Guerry](#),

[Jevons](#), [Langren](#), [Macdonell](#), [Michelson](#), [Minard](#), [Nightingale](#),

[OldMaps](#), [PearsonLee](#), [PolioTrials](#), [Prostitutes](#), [Pyx](#), [Quarrels](#), [Snow](#), [Wheat](#), [Yeast](#), [ZeaMays](#)

Other packages containing data sets of historical interest include:

The [Guerry-package](#), containing maps and other data sets related to Guerry's (1833) *Moral Statistics of France*.

morsecodes from the (defunct) [xgobi](#) package for data from Rothkopf (1957) on errors in learning morse code, a classical example for MDS.

The [psych](#) package, containing Galton's [peas](#) data. The same data set is contained in [alr3](#) as [galtonpeas](#).

**Examples**

```
# see examples for the separate data sets
```

---

Arbuthnot

*Arbuthnot's data on male and female birth ratios in London from 1629-1710.*

---

**Description**

John Arbuthnot (1710) used these time series data on the ratios of male to female births in London from 1629-1710 to carry out the first known significance test, comparing observed data to a null hypothesis. The data for these 81 years showed that in every year there were more male than female christenings.

On the assumption that male and female births were equally likely, he showed that the probability of observing 82 years with more males than females was vanishingly small ( $\sim 4.14 \times 10^{-25}$ ). He used this to argue that a nearly constant birth ratio  $> 1$  could be interpreted to show the guiding hand of a devine being. The data set adds variables of deaths from the plague and total mortality obtained by Campbell and from Creighton (1965).

**Usage**

```
data(Arbuthnot)
```

**Format**

A data frame with 82 observations on the following 7 variables.

Year a numeric vector, 1629-1710

Males a numeric vector, number of male christenings

Females a numeric vector, number of female christenings

Plague a numeric vector, number of deaths from plague

Mortality a numeric vector, total mortality

Ratio a numeric vector, ratio of Males/Females

Total a numeric vector, total christenings in London (000s)

**Details**

Sandy Zabell (1976) pointed out several errors and inconsistencies in the Arbuthnot data. In particular, the values for 1674 and 1704 are identical, suggesting that the latter were copied erroneously from the former.

**Source**

Arbuthnot, John (1710). "An argument for Devine Providence, taken from the constant Regularity observ'd in the Births of both Sexes," *Philosophical transactions*, 27, 186-190. Published in 1711.

**References**

Campbell, R. B., Arbuthnot and the Human Sex Ratio, <http://www.math.uni.edu/~campbell/arbuth.html>

Creighton, C. (1965). A History of Epidemics in Britain, 2nd edition, vol. 1 and 2. NY: Barnes and Noble.

S. Zabell (1976). Arbuthnot, Heberden, and the *Bills of Mortality*. Technical Report No. 40, Department of Statistics, University of Chicago.

**Examples**

```
data(Arbuthnot)
# plot the sex ratios
with(Arbuthnot, plot(Year,Ratio, type='b', ylim=c(1, 1.20), ylab="Sex Ratio (M/F)"))
abline(h=1, col="red")
# add loess smooth
Arb.smooth <- with(Arbuthnot, loess.smooth(Year,Ratio))
lines(Arb.smooth$x, Arb.smooth$y, col="blue", lwd=2)

# plot the total christenings to observe the anomalie in 1704
with(Arbuthnot, plot(Year>Total, type='b', ylab="Total Christenings"))
```

---

 Bowley

*Bowley's data on values of British and Irish trade, 1855-1899*


---

### Description

In one of the first statistical textbooks, Arthur Bowley (1901) used these data to illustrate an arithmetic and graphical analysis of time-series data using the total value of British and Irish exports from 1855-1899. He presented a line graph of the time-series data, supplemented by overlaid line graphs of 3-, 5- and 10-year moving averages. His goal was to show that while the initial series showed wide variability, moving averages made the series progressively smoother.

### Usage

```
data(Bowley)
```

### Format

A data frame with 45 observations on the following 2 variables.

Year Year, from 1855-1899

Value total value of British and Irish exports (millions of Pounds)

### Source

Bowley, A. L. (1901). *Elements of Statistics*. London: P. S. King and Son, p. 151-154.

Digitized from Bowley's graph.

### Examples

```
data(Bowley)

# plot the data
with(Bowley,plot(Year, Value, type='b',
  ylab="Value of British and Irish Exports",
  main="Bowley's example of the method of smoothing curves"))

# find moving averages-- use center alignment (requires width=ODD)
require(gtools, warn.conflicts=FALSE)
mav3<-running(Bowley$Value, width=3, align="center")
mav5<-running(Bowley$Value, width=5, align="center")
mav9<-running(Bowley$Value, width=9, align="center")
lines(Bowley$Year[2:44], mav3, col='blue', lty=2)
lines(Bowley$Year[3:43], mav5, col='green3', lty=3)
lines(Bowley$Year[5:41], mav9, col='brown', lty=4)

# add lowess smooth
lines(lowess(Bowley), col='red', lwd=2)

require(ggplot2, warn.conflicts=FALSE)
qplot(Year,Value, data=Bowley)+geom_smooth()
```

Cavendish

*Cavendish's Determinations of the Density of the Earth***Description**

Henry Cavendish carried out a series of experiments in 1798 to determine the mean density of the earth, as an indirect means to calculate the gravitational constant,  $G$ , in Newton's formula for the force ( $f$ ) of gravitational attraction,  $f = GmM/r^2$  between two bodies of mass  $m$  and  $M$ .

Stigler (1977) used these data to illustrate properties of robust estimators with real, historical data. For these data sets, he found that trimmed means performed as well or better than more elaborate robust estimators.

**Usage**

data(Cavendish)

**Format**

A data frame with 29 observations on the following 3 variables.

density Cavendish's 29 determinations of the mean density of the earth

density2 same as density, with the third value (4.88) replaced by 5.88

density3 same as density, omitting the the first 6 observations

**Details**

Density values ( $D$ ) of the earth are given as relative to that of water. If the earth is regarded as a sphere of radius  $R$ , Newton's law can be expressed as  $GD = 3g/(4\pi R)$ , where  $g = 9.806m/s^2$  is the acceleration due to gravity; so  $G$  is proportional to  $1/D$ .

density contains Cavendish's measurements as analyzed, where he treated the value 4.88 as if it were 5.88. density2 corrects this. Cavendish also changed his experimental apparatus after the sixth determination, using a stiffer wire in the torsion balance. density3 replaces the first 6 values with NA.

The modern "true" value of  $D$  is taken as 5.517. The gravitational constant can be expressed as  $G = 6.674 * 10^{-11}m^3/kg/s^2$ .

**Source**

Kyle Siegrist, "Virtual Laboratories in Probability and Statistics", <http://www.math.uah.edu/stat/data/Cavendish.xhtml>

Stephen M. Stigler (1977), "Do robust estimators work with *real* data?", *Annals of Statistics*, 5, 1055-1098

## References

Cavendish, H. (1798). Experiments to determine the density of the earth. *Philosophical Transactions of the Royal Society of London*, 88 (Part II), 469-527. Reprinted in A. S. Mackenzie (ed.), *The Laws of Gravitation*, 1900, New York: American.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*, NY: Wiley, p. 520.

## Examples

```
data(Cavendish)
summary(Cavendish)
boxplot(Cavendish, ylab='Density', xlab='Data set')
abline(h=5.517, col="red", lwd=2)

# trimmed means
sapply(Cavendish, mean, trim=.1, na.rm=TRUE)

# express in terms of G
G <- function(D, g=9.806, R=6371) 3*g / (4 * pi * R * D)

boxplot(10^5 * G(Cavendish), ylab='~ Gravitational constant (G)', xlab='Data set')
abline(h=10^5 * G(5.517), col="red", lwd=2)
```

---

ChestSizes

*Chest measurements of 5738 Scottish Militiamen*

---

## Description

Quetelet's data on chest measurements of 5738 Scottish Militiamen. Quetelet (1846) used this data as a demonstration of the normal distribution of physical characteristics.

## Usage

```
data(ChestSizes)
```

## Format

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

chest Chest size (in inches)

count Number of soldiers with this chest size

## Source

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont. CA: Wadsworth.

Statlib: <http://lib.stat.cmu.edu/DASL/Datafiles/MilitiamenChests.html>

## References

A. Quetelet, *Lettres a S.A.R. le Duc Regnant de Saxe-Cobourg et Gotha, sur la Theorie des Probabilites, Appliquee aux Sciences Morales et Politiques*. Brussels: M. Hayes, 1846, p. 400.

## Examples

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

# frequency polygon
plot(ChestSizes, type='b')
#barplot
barplot(ChestSizes[,2], ylab="Frequency", xlab="Chest size")
```

---

CushnyPebbles

*Cushny-Pebbles Data: Soporific Effects of Scopolamine Derivatives*

---

## Description

Cushny and Peebles (1905) studied the effects of hydrobromides related to scopolamine and atropine in producing sleep. The sleep of mental patients was measured without hypnotic (Control) and after treatment with one of three drugs: L. hyoscyamine hydrobromide (L\_hyoscyamine), L. hyoscine hydrobromide (L\_hyoscyine), and a mixture (racemic) form, DL\_hyoscine, called atropine. The L (levo) and D (detro) form of a given molecule are optical isomers (mirror images).

The drugs were given on alternate evenings, and the hours of sleep were compared with the intervening control night. Each of the drugs was tested in this manner a varying number of times in each subject. The average number of hours of sleep for each treatment is the response.

Student (1908) used these data to illustrate the paired-sample t-test in small samples, testing the hypothesis that the mean difference between a given drug and the control condition was zero. This data set became well known when used by Fisher (1925). Both Student and Fisher had problems labeling the drugs correctly (see Senn & Richardson (1994)), and consequently came to wrong conclusions.

But as well, the sample sizes (number of nights) for each mean differed widely, ranging from 3-9, and this was not taken into account in their analyses. To allow weighted analyses, the number of observations for each mean is contained in the data frame `CushnyPebblesN`.

## Usage

```
data(CushnyPebbles)
data(CushnyPebblesN)
```

**Format**

CushnyPebbles: A data frame with 11 observations on the following 4 variables.

Control a numeric vector: mean hours of sleep

L\_hyoscyamine a numeric vector: mean hours of sleep

L\_hyoscine a numeric vector: mean hours of sleep

D\_hyoscine a numeric vector: mean hours of sleep

CushnyPebblesN: A data frame with 11 observations on the following 4 variables.

Control a numeric vector: number of observations

L\_hyoscyamine a numeric vector: number of observations

L\_hyoscine a numeric vector: number of observations

DL\_hyoscine a numeric vector: number of observations

**Details**

The last patient (11) has no Control observations, and so is often excluded in analyses or other versions of this data set.

**Source**

Cushny, A. R., and Peebles, A. R. (1905), "The Action of Optical Isomers. II: Hyoscines," *Journal of Physiology*, 32, 501-510.

Senn, Stephen, Data from Cushny and Peebles, <http://www.senns.demon.co.uk/Data/Cushny.xls>

**References**

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh and London: Oliver & Boyd.

Student (1908), "The Probable Error of a Mean," *Biometrika*, 6, 1-25.

Senn, S.J. and Richardson, W. (1994), "The first t-test", *Statistics in Medicine*, 13, 785-803.

**See Also**

[sleep](#) for an alternative form of this data set.

**Examples**

```
data(CushnyPebbles)
# quick looks at the data
plot(CushnyPebbles)
boxplot(CushnyPebbles, ylab="Hours of Sleep", xlab="Treatment")

#####
# Repeated measures MANOVA
require(car)
```

```

CPmod <- lm(cbind(Control, L_hyoscyamine, L_hyoscine, DL_hyoscine) ~ 1, data=CushnyPebbles)

# Assign within-S factor and contrasts
Treatment <- factor(colnames(CushnyPebbles), levels=colnames(CushnyPebbles))
contrasts(Treatment) <- matrix(
c(-3, 1, 1, 1,
  0,-2, 1, 1,
  0, 0,-1, 1), ncol=3)
colnames(contrasts(Treatment)) <- c("Control.Drug", "L.DL", "L_hy.DL_hy")

Treats <- data.frame(Treatment)
(CPaov <- Anova(CPmod, idata=Treats, idesign= ~Treatment))
summary(CPaov, univariate=FALSE)

if (require(heplots)) {
  heplot(CPmod, idata=Treats, idesign= ~Treatment, iterm="Treatment",
xlab="Control vs Drugs", ylab="L vs DL drug")
  pairs(CPmod, idata=Treats, idesign= ~Treatment, iterm="Treatment")
}

#####
# reshape to long format, add Ns

CPlong <- stack(CushnyPebbles)[,2:1]
colnames(CPlong) <- c("treatment", "sleep")
CPN <- stack(CushnyPebblesN)
CPlong <- data.frame(patient=rep(1:11,4), CPlong, n=CPN$values)
str(CPlong)

```

---

## Description

Edgeworth (1885) took the first 75 lines in Book XI of Virgil's *Aeneid* and classified each of the first four "feet" of the line as a dactyl (one long syllable followed by two short ones) or not.

Grouping the lines in blocks of five gave a 4 x 25 table of counts, represented here as a data frame with ordered factors, Foot and Lines. Edgeworth used this table in what was among the first examples of analysis of variance applied to a two-way classification.

## Usage

```
data(Dactyl)
```

**Format**

A data frame with 60 observations on the following 3 variables.

Foot an ordered factor with levels 1 < 2 < 3 < 4

Lines an ordered factor with levels 1:5 < 6:10 < 11:15 < 16:20 < 21:25 < 26:30 < 31:35 < 36:40 < 41:45 < 46:50 < 51:55 < 56:60 < 61:65 < 66:70 < 71:75

count number of dactyls

**Source**

Stigler, S. (1999) *Statistics on the Table* Cambridge, MA: Harvard University Press, table 5.1.

**References**

Edgeworth, F. Y. (1885). On methods of ascertaining variations in the rate of births, deaths and marriages. *Journal of the [Royal] Statistical Society*, 48, 628-649.

**Examples**

```
data(Dactyl)

# display the basic table
xtabs(count ~ Foot+Lines, data=Dactyl)

# simple two-way anova
anova(dact.lm <- lm(count ~ Foot+Lines, data=Dactyl))

# plot the lm-quartet
op <- par(mfrow=c(2,2))
plot(dact.lm)
par(op)

# show table as a simple mosaicplot
mosaicplot(xtabs(count ~ Foot+Lines, data=Dactyl), shade=TRUE)
```

---

DrinksWages

---

*Elderton and Pearson's (1910) data on drinking and wages*


---

**Description**

In 1910, Karl Pearson weighed in on the debate, fostered by the temperance movement, on the evils done by alcohol not only to drinkers, but to their families. The report "A first study of the influence of parental alcoholism on the physique and ability of their offspring" was an ambitious attempt to use the new methods of statistics to bear on an important question of social policy, to see if the hypothesis that children were damaged by parental alcoholism would stand up to statistical scrutiny.

Working with his assistant, Ethel M. Elderton, Pearson collected voluminous data in Edinburgh and Manchester on many aspects of health, stature, intelligence, etc. of children classified according to

the drinking habits of their parents. His conclusions were almost invariably negative: the tendency of parents to drink appeared unrelated to any thing he had measured.

The firestorm that this report set off is well described by Stigler (1999), Chapter 1. The data set `DrinksWages` is just one of Pearson's many tables, that he published in a letter to *The Times*, August 10, 1910.

### Usage

```
data(DrinksWages)
```

### Format

A data frame with 70 observations on the following 6 variables, giving the number of non-drinkers (`sober`) and drinkers (`drinks`) in various occupational categories (`trade`).

`class` wage class: a factor with levels A B C

`trade` a factor with levels baker barman billposter ... wellsinker wireworker

`sober` the number of non-drinkers, a numeric vector

`drinks` the number of drinkers, a numeric vector

`wage` weekly wage (in shillings), a numeric vector

`n` total number, a numeric vector

### Details

The data give Karl Pearson's tabulation of the father's trades from an Edinburgh sample, classified by whether they drink or are sober, and giving average weekly wage.

The wages are averages of the individuals' nominal wages. Class A is those with wages under 2.5s.; B: those with wages 2.5s. to 30s.; C: wages over 30s.

### Source

Pearson, K. (1910). *The Times*, August 10, 1910.

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Table 1.1

### References

M. E. Elderton & K. Pearson (1910). A first study of the influence of parental alcoholism on the physique and ability of their offspring, *Eugenics Laboratory Memoirs*, 10.

### Examples

```
data(DrinksWages)
plot(DrinksWages)
```

```
# plot proportion sober vs. wage | class
with(DrinksWages, plot(wage, sober/n, col=c("blue", "red", "green")[class]))
```

```
# fit logistic regression model of sober on wage
mod.sober <- glm(cbind(sober, n) ~ wage, family=binomial, data=DrinksWages)
summary(mod.sober)
op <- par(mfrow=c(2,2))
plot(mod.sober)
par(op)

# TODO: plot fitted model
```

---

Fingerprints

*Waite's data on Patterns in Fingerprints*


---

### Description

Waite (1915) was interested in analyzing the association of patterns in fingerprints, and produced a table of counts for 2000 right hands, classified by the number of fingers describable as a "whorl", a "small loop" (or neither). Because each hand contributes five fingers, the number of Whorls + Loops cannot exceed 5, so the contingency table is necessarily triangular.

Karl Pearson (1904) introduced the test for independence in contingency tables, and by 1913 had developed methods for "restricted contingency tables," such as the triangular table analyzed by Waite. The general formulation of such tests for association in restricted tables is now referred to as models for quasi-independence.

### Usage

```
data(Fingerprints)
```

### Format

A frequency data frame with 36 observations on the following 3 variables, representing a 6 x 6 table giving the cross-classification of the fingers on 2000 right hands as a whorl, small loop or neither.

Whorls Number of whorls, an ordered factor with levels 0 < 1 < 2 < 3 < 4 < 5

Loops Number of small loops, an ordered factor with levels 0 < 1 < 2 < 3 < 4 < 5

count Number of hands

### Details

Cells for which Whorls + Loops > 5 have NA for count

### Source

Stigler, S. M. (1999). *Statistics on the Table*. Cambridge, MA: Harvard University Press, table 19.4.

## References

Pearson, K. (1904). Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation. Reprinted in *Karl Pearson's Early Statistical Papers*, Cambridge: Cambridge University Press, 1948, 443-475.

Waite, H. (1915). The analysis of fingerprints, *Biometrika*, 10, 421-478.

## Examples

```
data(Fingerprints)
xtabs(count ~ Whorls + Loops, data=Fingerprints)
```

---

Galton

*Galton's data on the heights of parents and their children*

---

## Description

Galton (1886) presented these data in a table, showing a cross-tabulation of 928 adult children born to 205 fathers and mothers, by their height and their mid-parent's height. He visually smoothed the bivariate frequency distribution and showed that the contours formed concentric and similar ellipses, thus setting the stage for correlation, regression and the bivariate normal distribution.

## Usage

```
data(Galton)
```

## Format

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

parent a numeric vector: height of the mid-parent (average of father and mother)

child a numeric vector: height of the child

## Details

The data are recorded in class intervals of width 1.0 in. He used non-integer values for the center of each class interval because of the strong bias toward integral inches.

All of the heights of female children were multiplied by 1.08 before tabulation to compensate for sex differences. See Hanley (2004) for a reanalysis of Galton's raw data questioning whether this was appropriate.

## Source

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature *Journal of the Anthropological Institute*, 15, 246-263

## References

Friendly, M. & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41, 103-130.

Galton, F. (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences*. London: Macmillan.

Hanley, J. A. (2004). "Transmuting" Women into Men: Galton's Family Data on Human Stature. *The American Statistician*, 58, 237-243. See: <http://www.medicine.mcgill.ca/epidemiology/hanley/galton/> for source materials.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press, Table 8.1

Wachsmuth, A. W., Wilkinson L., Dallal G. E. (2003). Galton's bend: A previously undiscovered nonlinearity in Galton's family stature regression data. *The American Statistician*, 57, 190-192. <http://www.cs.uic.edu/~wilkinson/Publications/galton.pdf>

## See Also

[PearsonLee, galton](#)

## Examples

```
data(Galton)

#####
# sunflower plot with regression line and data ellipses and lowess smooth
#####

with(Galton,
{
sunflowerplot(parent,child, xlim=c(62,74), ylim=c(62,74))
reg <- lm(child ~ parent)
abline(reg)
lines(lowess(parent, child), col="blue", lwd=2)
if(require(car)) {
dataEllipse(parent,child, xlim=c(62,74), ylim=c(62,74), plot.points=FALSE)
}
})
```

## Description

Andre-Michel Guerry (1833) was the first to systematically collect and analyze social data on such things as crime, literacy and suicide with the view to determining social laws and the relations among these variables.

The Guerry data frame comprises a collection of 'moral variables' on the 86 departments of France around 1830. A few additional variables have been added from other sources.

## Usage

```
data(Guerry)
```

## Format

A data frame with 86 observations (the departments of France) on the following 23 variables.

**dept** Department ID: Standard numbers for the departments, except for Corsica (200)

**Region** Region of France ('N'='North', 'S'='South', 'E'='East', 'W'='West', 'C'='Central'). Corsica is coded as NA

**Department** Department name: Departments are named according to usage in 1830, but without accents. A factor with levels Ain Aisne Allier ... Vosges Yonne

**Crime\_pers** Population per Crime against persons. Source: A2 (Compte général, 1825-1830)

**Crime\_prop** Population per Crime against property. Source: A2 (Compte général, 1825-1830)

**Literacy** Percent Read & Write: Percent of military conscripts who can read and write. Source: A2

**Donations** Donations to the poor. Source: A2 (Bulletin des lois)

**Infants** Population per illegitimate birth. Source: A2 (Bureau des Longitudes, 1817-1821)

**Suicides** Population per suicide. Source: A2 (Compte général, 1827-1830)

**MainCity** Size of principal city ('1:Sm', '2:Med', '3:Lg'), used as a surrogate for population density. Large refers to the top 10, small to the bottom 10; all the rest are classed Medium. Source: A1. An ordered factor with levels 1:Sm < 2:Med < 3:Lg

**Wealth** Per capita tax on personal property. A ranked index based on taxes on personal and movable property per inhabitant. Source: A1

**Commerce** Commerce and Industry, measured by the rank of the number of patents / population. Source: A1

**Clergy** Distribution of clergy, measured by the rank of the number of Catholic priests in active service / population. Source: A1 (Almanach officiel du clergy, 1829)

**Crime\_parents** Crimes against parents, measured by the rank of the ratio of crimes against parents to all crimes– Average for the years 1825-1830. Source: A1 (Compte général)

**Infanticide** Infanticides per capita. A ranked ratio of number of infanticides to population– Average for the years 1825-1830. Source: A1 (Compte général)

**Donation\_clergy** Donations to the clergy. A ranked ratio of the number of bequests and donations inter vivos to population– Average for the years 1815-1824. Source: A1 (Bull. des lois, ordonn. d'autorisation)

- Lottery** Per capita wager on Royal Lottery. Ranked ratio of the proceeds bet on the royal lottery to population— Average for the years 1822-1826. Source: A1 (Compte rendus par le ministre des finances)
- Desertion** Military desertion, ratio of the number of young soldiers accused of desertion to the force of the military contingent, minus the deficit produced by the insufficiency of available billets— Average of the years 1825-1827. Source: A1 (Compte du ministere du guerre, 1829 etat V)
- Instruction** Instruction. Ranks recorded from Guerry's map of Instruction. Note: this is inversely related to Literacy (as defined here)
- Prostitutes** Prostitutes in Paris. Number of prostitutes registered in Paris from 1816 to 1834, classified by the department of their birth Source: Parent-Duchatelet (1836), *De la prostitution en Paris*
- Distance** Distance to Paris (km). Distance of each department centroid to the centroid of the Seine (Paris) Source: calculated from department centroids
- Area** Area (1000 km<sup>2</sup>). Source: Angeville (1836)
- Pop1831** 1831 population. Population in 1831, taken from Angeville (1836), *Essai sur la Statistique de la Population française*, in 1000s

## Details

Note that most of the variables (e.g., Crime\_pers) are scaled so that 'more is better' morally.

Values for the quantitative variables displayed on Guerry's maps were taken from Table A2 in the English translation of Guerry (1833) by Whitt and Reinking. Values for the ranked variables were taken from Table A1, with some corrections applied. The maximum is indicated by rank 1, and the minimum by rank 86.

## Source

Angeville, A. (1836). *Essai sur la Statistique de la Population française* Paris: F. Doufour.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France* Paris: Crochard. English translation: Hugh P. Whitt and Victor W. Reinking, Lewiston, N.Y. : Edwin Mellen Press, 2002.

Parent-Duchatelet, A. (1836). *De la prostitution dans la ville de Paris*, 3rd ed, 1857, p. 32, 36

## References

Dray, S. and Jombart, T. (2009). A Revisit Of Guerry's Data: Introducing Spatial Constraints In Multivariate Analysis. Unpublished manuscript.

Brunsdon, C. and Dykes, J. (2007). Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis. Geographical Information Science Research Conference (GISRUK 07), NUI Maynooth, Ireland, April, 2007.

Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis. *Statistical Science*, 22, 368-399.

Friendly, M. (2007). Data from A.-M. Guerry, Essay on the Moral Statistics of France (1833), <http://www.math.yorku.ca/SCS/Gallery/guerry/guerrydat.html>.

**See Also**

The **Guerry** package for maps of France: [gfrance](#) and related data.

**Examples**

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

---

Jevons

*W. Stanley Jevons' data on numerical discrimination*

---

**Description**

In a remarkable brief note in *Nature*, 1871, W. Stanley Jevons described the results of an experiment he had conducted on himself to determine the limits of the number of objects an observer could comprehend immediately without counting them. This was an important philosophical question: How many objects can the mind embrace at once?

He carried out 1027 trials in which he tossed an "uncertain number" of uniform black beans into a box and immediately attempted to estimate the number "without the least hesitation". His questions, procedure and analysis anticipated by 75 years one of the most influential papers in modern cognitive psychology by George Miller (1956), "The magical number 7 plus or minus 2: Some limits on ..." For Jevons, the magical number was 4.5, representing an empirical law of complete accuracy.

**Usage**

```
data(Jevons)
```

**Format**

A frequency data frame with 50 observations on the following 4 variables.

actual Actual number: a numeric vector

estimated Estimated number: a numeric vector

frequency Frequency of this combination of (actual, estimated): a numeric vector

error actual-estimated: a numeric vector

**Details**

The original data were presented in a two-way, 13 x 13 frequency table, estimated (3:15) x actual (3:15).

**Source**

Jevons, W. S. (1871). The Power of Numerical Discrimination, *Nature*, 1871, III (281-282)

## References

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, *Psychological Review*, 63, 81-97, <http://www.musanim.com/miller1956/>

## Examples

```
data(Jevons)
# show as tables
xtabs(frequency ~ estimated+actual, data=Jevons)
xtabs(frequency ~ error+actual, data=Jevons)

# show as sunflowerplot with regression line
with(Jevons, sunflowerplot(actual, estimated, frequency, main="Jevons data on numerical estimation"))
Jmod <- lm(estimated ~ actual, data=Jevons, weights=frequency)
abline(Jmod)

# show as balloonplots
if (require(gplots)) {

with(Jevons, balloonplot(actual, estimated, frequency, xlab="actual", ylab="estimated",
  main="Jevons data on numerical estimation\nBubble area proportional to frequency", text.size=0.8))

with(Jevons, balloonplot(actual, error, frequency, xlab="actual", ylab="error",
  main="Jevons data on numerical estimation: Errors\nBubble area proportional to frequency", text.size=0.8))
}

# plot average error
if(require(reshape)) {
unJevons <- untable(Jevons, Jevons$frequency)
str(unJevons)

mean_error <- function(df) mean(df$error, na.rm=TRUE)
Jmean <- ddply(unJevons, .(actual), mean_error)
with(Jmean, plot(actual, V1, ylab='Mean error', xlab='Actual number', type='b', main='Jevons data'))
abline(h=0)
}
```

---

Langren

*van Langren's data on longitude distance between Toledo and Rome*

---

## Description

Michael Florent van Langren (1598-1675) was a Dutch mathematician and astronomer, who served as a royal mathematician to King Phillip IV of Spain, and who worked on one of the most significant problems of his time— the accurate determination of longitude, particularly for navigation at sea.

In order to convince the Spanish court of the seriousness of the problem (often resulting in great losses through ship wrecks), he prepared a 1-dimensional line graph, showing all the available

estimates of the distance in longitude between Toledo and Rome, which showed large errors, for even this modest distance. This 1D line graph, from Langren (1644), is believed to be the first known graph of statistical data (Friendly et al., 2009). It provides a compelling example of the notions of statistical variability and bias.

The data frame `Langren1644` gives the estimates and other information derived from the previously known 1644 graph. It turns out that van Langren produced other versions of this graph, as early as 1628. The data frame `Langren.all` gives the estimates derived from all known versions of this graph.

### Usage

```
data(Langren1644)
data(Langren.all)
```

### Format

`Langren1644`: A data frame with 12 observations on the following 9 variables, giving determinations of the distance in longitude between Toledo and Rome, from the 1644 graph.

`Name` The name of the person giving a determination, a factor with levels A. Argelius ... T. Brahe

`Longitude` Estimated value of the longitude distance between Toledo and Rome

`Year` Year associated with this determination

`Longname` A longer version of the `Name`, where appropriate; a factor with levels Andrea Argoli Christoph Clavius Tycho Brahe

`City` The principal city where this person worked; a factor with levels Alexandria Amsterdam Bamberg Bologna Frankfurt Hven Leuven Middelburg Nuremberg Padua Paris Rome

`Country` The country where this person worked; a factor with levels Belgium Denmark Egypt Flanders France Germany Italy Italy

`Latitude` Latitude of this `City`; a numeric vector

`Source` Likely source for this determination of `Longitude`; a factor with levels Astron Map

`Gap` A numeric vector indicating whether the `Longitude` value is below or above the median

`Langren.all`: A data frame with 61 observations on the following 4 variables, giving determinations of `Longitude` between Toledo and Rome from all known versions of van Langren's graph.

`Author` Author of the graph, a factor with levels Langren Lelewel

`Year` Year of publication

`Name` The name of the person giving a determination, a factor with levels Algonos1 Algonos2 Apianus ... Schonerus

`Longitude` Estimated value of the longitude distance between Toledo and Rome

### Details

In all the graphs, Toledo is implicitly at the origin and Rome is located relatively at the value of `Longitude`. To judge correspondence with an actual map, the positions in (lat, long) are

```
toledo <- c(39.86, -4.03); rome <- c(41.89, 12.5)
```

## Source

The longitude values were digitized from images of the various graphs, which may be found on the Supplementary materials page for Friendly et al. (2009).

## References

Friendly, M., Valero-Mora, P. and Ulargui, J. I. (2009). The First (Known) Statistical Graph: Michael Florent van Langren and the "Secret" of Longitude. Unpublished ms (submitted). Supplementary materials: <http://www.math.yorku.ca/SCS/Gallery/langren/>.

Langren, M. F. van. (1644). *La Verdadera Longitud por Mar y Tierra*. Antwerp: (n.p.), 1644. English translation available at <http://www.math.yorku.ca/SCS/Gallery/langren/verdadera.pdf>.

Lelewel, J. (1851). *Géographie du Moyen Âge*. Paris: Pilliet, 1851.

## Examples

```
data(Langren1644)

## Not run:
require(maps)
require(ggplot2)
require(reshape)
require(plyr)

# set latitude to that of Toledo
Langren1644$Latitude <- 39.68

#           x/long   y/lat
bbox <- c( 38.186, -9.184,
          43.692, 28.674 )
bbox <- matrix(bbox, 2, 2, byrow=TRUE)

borders <- as.data.frame(map("world", plot = FALSE,
                             xlim = expand_range(bbox[,2], 0.2),
                             ylim = expand_range(bbox[,1], 0.2))[c("x", "y")])

data(world.cities)
# get actual locations of Toledo & Rome
cities <- subset(world.cities,
                 name %in% c("Rome", "Toledo") & country.etc %in% c("Spain", "Italy"))
colnames(cities)[4:5]<-c("Latitude", "Longitude")

mplot <- ggplot(Langren1644, aes(Longitude, Latitude) ) +
  geom_path(aes(x, y), borders, colour = "grey60") +
  geom_point(y = 40) +
  geom_text(aes(label = Name), y = 40.1, angle = 90, hjust = 0, size = 3)
mplot <- mplot +
  geom_segment(aes(x=-4.03, y=40, xend=30, yend=40))

mplot <- mplot +
  geom_point(data = cities, colour = "red", size = 2) +
```

```

geom_text(data=cities, aes(label=name), color="red", size=3, vjust=-0.5) +
  coord_cartesian(xlim=bbox[,2], ylim=bbox[,1])

# make the plot have approximately aspect ratio = 1
windows(width=10, height=2)
mplot

## End(Not run)

if (require(ReadImages)) {
  gimage <- read.jpeg(system.file("images", "google-toledo-rome3.jpg", package="HistData"))
  plot(gimage)

  # pixel coordinates of Toledo and Rome in the image, measured from the bottom left corner
  toledo.map <- c(130, 59)
  rome.map <- c(505, 119)

  # confirm locations of Toledo and Rome
  points(rbind(toledo.map, rome.map), cex=2)

  # set a scale for translation of lat,long to pixel x,y
  scale <- data.frame(x=c(130, 856), y=c(52,52))
  rownames(scale)=c(0,30)
  lines(scale)

  xlate <- function(x) {
    130+x*726/30
  }
  points(x=xlate(Langren1644$Longitude), y=rep(57, nrow(Langren1644)), pch=25, col="blue")
  text(x=xlate(Langren1644$Longitude), y=rep(57, nrow(Langren1644)), labels=Langren1644$Name, srt=90, adj=c(0, 0.5))
}

# show variation in estimates across graphs
library(lattice)
graph <- paste(Langren.all$Author, Langren.all$Year)
dotplot(Name ~ Longitude, data=Langren.all)

dotplot( as.factor(Year) ~ Longitude, data=Langren.all, groups=Name, type="o")

dotplot(Name ~ Longitude|graph, data=Langren.all, groups=graph)

# why the gap?
gap.mod <- glm(Gap ~ Year + Source + Latitude, family=binomial, data=Langren1644)
anova(gap.mod, test="Chisq")

```

## Description

In the second issue of *Biometrika*, W. R. Macdonell (1902) published an extensive paper, *On Criminal Anthropometry and the Identification of Criminals* in which he included numerous tables of physical characteristics 3000 non-habitual male criminals serving their sentences in England and Wales. His Table III (p. 216) recorded a bivariate frequency distribution of height by finger length. His main purpose was to show that Scotland Yard could have indexed their material more efficiently, and find a given profile more quickly.

W. S. Gosset (aka "Student") used these data in two classic papers in 1908, in which he derived various characteristics of the sampling distributions of the mean, standard deviation and Pearson's  $r$ . He said, "Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically." Among his experiments, he randomly shuffled the 3000 observations from Macdonell's table, and then grouped them into samples of size 4, 8, ..., calculating the sample means, standard deviations and correlations for each sample.

## Usage

```
data(Macdonell)
data(MacdonellDF)
```

## Format

Macdonell: A frequency data frame with 924 observations on the following 3 variables giving the bivariate frequency distribution of height and finger.

height lower class boundaries of height, in decimal ft.

finger length of the left middle finger, in mm.

frequency frequency of this combination of height and finger

MacdonellDF: A data frame with 3000 observations on the following 2 variables.

height a numeric vector

finger a numeric vector

## Details

Class intervals for height in Macdonell's table were given in 1 in. ranges, from (4' 7" 9/16 - 4' 8" 9/16), to (6' 4" 9/16 - 6' 5" 9/16). The values of height are taken as the lower class boundaries.

For convenience, the data frame MacdonellDF presents the same data, in expanded form, with each combination of height and finger replicated frequency times.

## Source

Macdonell, W. R. (1902). On Criminal Anthropometry and the Identification of Criminals. *Biometrika*, 1(2), 177-227. doi:10.1093/biomet/1.2.177 <http://www.jstor.org/stable/2331487>

The data used here were obtained from:

Hanley, J. (2008). Macdonell data used by Student. <http://www.medicine.mcgill.ca/epidemiology/hanley/Student/>

## References

Hanley, J. and Julien, M. and Moodie, E. (2008). Student's z, t, and s: What if Gosset had R? *The American Statistician*, 62(1), 64-69.

Gossett, W. S. [Student] (1908). Probable error of a mean. *Biometrika*, 6(1), 1-25. <http://www.york.ac.uk/depts/math/histstat/student.pdf>

Gossett, W. S. [Student] (1908). Probable error of a correlation coefficient. *Biometrika*, 6, 302-310.

## Examples

```
data(Macdonell)

# display the frequency table
xtabs(frequency ~ finger+round(height,3), data=Macdonell)

## Some examples by james.hanley@mcgill.ca   October 16, 2011
## http://www.biostat.mcgill.ca/hanley/
## See: http://www.biostat.mcgill.ca/hanley/Student/

#####
## naive contour plots of height and finger ##
#####

# make a 22 x 42 table
attach(Macdonell)
ht <- unique(height)
fi <- unique(finger)
fr <- t(matrix(frequency, nrow=42))
detach(Macdonell)

dev.new(width=10, height=5) # make plot double wide
op <- par(mfrow=c(1,2),mar=c(0.5,0.5,0.5,0.5),oma=c(2,2,0,0))

dx <- 0.5/12
dy <- 0.5/12

plot(ht,ht,xlim=c(min(ht)-dx,max(ht)+dx),
      ylim=c(min(fi)-dy,max(fi)+dy), xlab="", ylab="", type="n" )

# unpack 3000 heights while looping though the frequencies
heights <- c()
for(i in 1:22) {
  for (j in 1:42) {
    f <- fr[i,j]
    if(f>0) heights <- c(heights,rep(ht[i],f))
    if(f>0) text(ht[i], fi[j], toString(f), cex=0.4, col="grey40" )
  }
}
text(4.65,13.5, "Finger length (cm)",adj=c(0,1), col="black") ;
text(5.75,9.5, "Height (feet)", adj=c(0,1), col="black") ;
text(6.1,11, "Observed bin\nfrequencies", adj=c(0.5,1), col="grey40",cex=0.85) ;
```

```

# crude contour plot
contour(ht, fi, fr, add=TRUE, drawlabels=FALSE, col="grey60")

# smoother contour plot (Galton smoothed 2-D frequencies this way)
# [Galton had experience with plotting isobars for meteorological data]
# it was the smoothed plot that made him remember his 'conic sections'
# and ask a mathematician to work out for him the iso-density
# contours of a bivariate Gaussian distribution...

dx <- 0.5/12; dy <- 0.05 ; # shifts caused by averaging

plot(ht,ht,xlim=c(min(ht),max(ht)),ylim=c(min(fi),max(fi)), xlab="", ylab="", type="n" )

sm.fr <- matrix(rep(0,21*41),nrow <- 21)
for(i in 1:21) {
for (j in 1:41) {
  smooth.freq <- (1/4) * sum( fr[i:(i+1), j:(j+1)] )
  sm.fr[i,j] <- smooth.freq
  if(smooth.freq > 0 )
  text(ht[i]+dx, fi[j]+dy, sub("^0.", ".", toString(smooth.freq)), cex=0.4, col="grey40" )
}
}

contour(ht[1:21]+dx, fi[1:41]+dy, sm.fr, add=TRUE, drawlabels=FALSE, col="grey60")
text(6.05,11, "Smoothed bin\nfrequencies", adj=c(0.5,1), col="grey40", cex=0.85) ;
par(op)
dev.new() # new default device

#####
## bivariate kernel density estimate
#####

if(require(KernSmooth)) {
MDest <- bkde2D(MacdonellDF, bandwidth=c(1/8, 1/8))
contour(x=MDest$x1, y=MDest$x2, z=MDest$fhat,
xlab="Height (feet)", ylab="Finger length (cm)", col="red", lwd=2)
with(MacdonellDF, points(jitter(height), jitter(finger), cex=0.5))
}

#####
## sunflower plot of height and finger with data ellipses ##
#####

with(MacdonellDF,
{
sunflowerplot(height, finger, size=1/12, seg.col="green3",
xlab="Height (feet)", ylab="Finger length (cm)")
reg <- lm(finger ~ height)
abline(reg, lwd=2)
if(require(car)) {
dataEllipse(height, finger, plot.points=FALSE, levels=c(.40, .68, .95))
}
}

```

```
})
```

```
#####
## Sampling distributions of sample sd (s) and z=(ybar-mu)/s
#####

# note that Gosset used a divisor of n (not n-1) to get the sd.
# He also used Sheppard's correction for the 'binning' or grouping.
# with concatenated height measurements...

mu <- mean(heights) ; sigma <- sqrt( 3000 * var(heights)/2999 )
c(mu,sigma)

# 750 samples of size n=4 (as Gosset did)

# see Student's z, t, and s: What if Gosset had R?
# [Hanley J, Julien M, and Moodie E. The American Statistician, February 2008]

# see also the photographs from Student's notebook ('Original small sample data and notes")
# under the link "Gosset' 750 samples of size n=4"
# on website http://www.biostat.mcgill.ca/hanley/Student/
# and while there, look at the cover of the Notebook containing his yeast-cell counts
# http://www.medicine.mcgill.ca/epidemiology/hanley/Student/750samplesOf4/Covers.JPG
# (Biometrika 1907) and decide for yourself why Gosset, when forced to write under a
# pen-name, might have taken the name he did!

# PS: Can you figure out what the 750 pairs of numbers signify?
# hint: look again at the numbers of rows and columns in Macdonell's (frequency) Table III.

n <- 4
Nsamples <- 750

y.bar.values <- s.over.sigma.values <- z.values <- c()
for (samp in 1:Nsamples) {
  y <- sample(heights,n)
  y.bar <- mean(y)
  s <- sqrt( (n/(n-1))*var(y) )
  z <- (y.bar-mu)/s
  y.bar.values <- c(y.bar.values,y.bar)
  s.over.sigma.values <- c(s.over.sigma.values,s/sigma)
  z.values <- c(z.values,z)
}

op <- par(mfrow=c(2,2),mar=c(2.5,2.5,2.5,2.5),oma=c(2,2,0,0))
# sampling distributions
hist(heights,breaks=seq(4.5,6.5,1/12), main="Histogram of heights (N=3000)")
hist(y.bar.values, main=paste("Histogram of y.bar (n=",n,")",sep=""))

hist(s.over.sigma.values,breaks=seq(0,4,0.1),
main=paste("Histogram of s/sigma (n=",n,")",sep=""));
```

```

z=seq(-5,5,0.25)+0.125
hist(z.values,breaks=z-0.125, main="Histogram of z=(ybar-mu)/s")
# theoretical
lines(z, 750*0.25*sqrt(n-1)*dt(sqrt(n-1)*z,3), col="red", lwd=1)
par(op)

#####
## Chisquare probability plot for bivariate normality
#####

mu <- colMeans(MacdonellDF)
sigma <- var(MacdonellDF)
Dsqr <- mahalnobis(MacdonellDF, mu, sigma)

Q <- qchisq(1:3000/3000, 2)
plot(Q, sort(Dsqr), xlab="Chisquare (2) quantile", ylab="Squared distance")
abline(a=0, b=1, col="red", lwd=2)

```

---

Michelson

*Michelson's Determinations of the Velocity of Light*

---

## Description

The data frame `Michelson` gives Albert Michelson's measurements of the velocity of light in air, made from June 5 to July 2, 1879, reported in Michelson (1882). The given values + 299,000 are Michelson's measurements in km/sec. The number of cases is 100 and the "true" value on this scale is 734.5.

Stigler (1977) used these data to illustrate properties of robust estimators with real, historical data. For this purpose, he divided the 100 measurements into 5 sets of 20 each. These are contained in `MichelsonSets`.

## Usage

```

data(Michelson)
data(MichelsonSets)

```

## Format

`Michelson`: A data frame with 100 observations on the following variable, given in time order of data collection

`velocity` a numeric vector

`MichelsonSets`: A 20 x 5 matrix, with format int [1:20, 1:5] 850 850 1000 810 960 800 830 830 880 720 ... - attr(\*, "dimnames")=List of 2 ..\$: NULL ..\$: chr [1:5] "ds12" "ds13" "ds14" "ds15" ...

## Details

The "true" value is taken to be 734.5, arrived at by taking the "true" speed of light in a vacuum to be 299,792.5 km/sec, and adjusting for the velocity in air.

The data values are recorded in order, and so may also be taken as a time series.

## Source

Kyle Siegrist, "Virtual Laboratories in Probability and Statistics", <http://www.math.uah.edu/stat/data/Michelson.xhtml>

Stephen M. Stigler (1977), "Do robust estimators work with *real* data?", *Annals of Statistics*, 5, 1055-1098

## References

Michelson, A. A. (1882). "Experimental determination of the velocity of light made at the United States Naval Academy, Anapolis". *Astronomical Papers*, 1,109-145, U. S. Nautical Almanac Office.

## Examples

```
data(Michelson)

# density plot (default bandwidth & 0.6 * bw)
plot(density(Michelson$velocity), xlab="Speed of light - 299000 (km/s)",
     main="Density plots of Michelson data")
lines(density(Michelson$velocity), adjust=0.6, lty=2)
rug(jitter(Michelson$velocity))
abline(v=mean(Michelson$velocity), col="blue")
abline(v=734.5, col="red", lwd=2)
text(mean(Michelson$velocity), .004, "mean", srt=90, pos=2)
text(734.5, .004, "true", srt=90, pos=2)

# index / time series plot
plot(Michelson$velocity, type="b")
abline(h=734.5, col="red", lwd=2)
lines(lowess(Michelson$velocity), col="blue", lwd=2)

# examine lag=1 differences
plot(diff(Michelson$velocity), type="b")
lines(lowess(diff(Michelson$velocity)), col="blue", lwd=2)

# examine different data sets
boxplot(MichelsonSets, ylab="Velocity of light - 299000 (km/s)", xlab="Data set")
abline(h=734.5, col="red", lwd=2)

# means and trimmed means
(mn <- apply(MichelsonSets, 2, mean))
(tm <- apply(MichelsonSets, 2, mean, trim=.1))
points(1:5, mn)
points(1:5+.05, tm, pch=16, col="blue")
```

---

Minard	<i>Data from Minard's famous graphic map of Napoleon's march on Moscow</i>
--------	--

---

### Description

Charles Joseph Minard's graphic depiction of the fate of Napoleon's Grand Army in the Russian campaign of 1815 has been called the "greatest statistical graphic ever drawn" (Tufte, 1983). Friendly (2002) describes some background for this graphic, and presented it as Minard's Challenge: to reproduce it using modern statistical or graphic software, in a way that showed the elegance of some computer language to both describe and produce this graphic.

### Usage

```
data(Minard.troops)
data(Minard.cities)
data(Minard.temp)
```

### Format

`Minard.troops`: A data frame with 51 observations on the following 5 variables giving the number of surviving troops.

```
long Longitude
lat Latitude
survivors Number of surviving troops, a numeric vector
direction a factor with levels A ("Advance") R ("Retreat")
group a numeric vector
```

`Minard.cities`: A data frame with 20 observations on the following 3 variables giving the locations of various places along the path of Napoleon's army.

```
long Longitude
lat Latitude
city City name: a factor with levels Bobr Chjat ... Witebsk Wixma
```

`Minard.temp`: A data frame with 9 observations on the following 4 variables, giving the temperature at various places along the march of retreat from Moscow.

```
long Longitude
temp Temperature
days Number of days on the retreat march
date a factor with levels Dec01 Dec06 Dec07 Nov09 Nov14 Nov28 Oct18 Oct24
```

### Details

`date` in `Minard.temp` should be made a real date in 1815.

**Source**

<http://www.cs.uic.edu/~wilkinson/TheGrammarOfGraphics/minard.txt>

**References**

Friendly, M. (2002). Visions and Re-visions of Charles Joseph Minard, *Journal of Educational and Behavioral Statistics*, 27, No. 1, 31-51.

Friendly, M. (2003). Re-Visions of Minard. <http://www.math.yorku.ca/SCS/Gallery/re-minard.html>

**Examples**

```
data(Minard.troops); data(Minard.cities)

## Not run:
require(ggplot2)
plot_troops <- ggplot(Minard.troops, aes(long, lat)) +
  geom_path(aes(size = survivors, colour = direction, group = group))

plot_both <- plot_troops +
  geom_text(aes(label = city), size = 4, data = Minard.cities)

plot_polished <- plot_both +
  scale_size(to = c(1, 12),
    breaks = c(1, 2, 3) * 10^5, labels = comma(c(1, 2, 3) * 10^5)) +
  scale_colour_manual(values = c("grey50", "red")) +
  xlab(NULL) +
  ylab(NULL)

# re-scale the plot window to an aspect ratio of ~ 4 x 1
windows(width=12, height=3)
plot_polished

## TODO: add the plot of temperature below

## End(Not run)
```

---

Nightingale

*Florence Nightingale's data on deaths from various causes in the  
Crimean War*

---

**Description**

In the history of data visualization, Florence Nightingale is best remembered for her role as a social activist and her view that statistical data, presented in charts and diagrams, could be used as powerful arguments for medical reform.

After witnessing deplorable sanitary conditions in the Crimea, she wrote several influential texts (Nightingale, 1858, 1859), including polar-area graphs (sometimes called "Coxcombs" or rose diagrams), showing the number of deaths in the Crimean from battle compared to disease or preventable causes that could be reduced by better battlefield nursing care.

Her *Diagram of the Causes of Mortality in the Army in the East* showed that most of the British soldiers who died during the Crimean War died of sickness rather than of wounds or other causes. It also showed that the death rate was higher in the first year of the war, before a Sanitary Commissioners arrived in March 1855 to improve hygiene in the camps and hospitals.

### Usage

```
data(Nightingale)
```

### Format

A data frame with 24 observations on the following 10 variables.

Date a Date, composed as `as.Date(paste(Year, Month, 1, sep='-'), "%Y-%b-%d")`

Month Month of the Crimean War, an ordered factor

Year Year of the Crimean War

Army Estimated average monthly strength of the British army

Disease Number of deaths from preventable or mitagable zymotic diseases

Wounds Number of deaths directly from battle wounds

Other Number of deaths from other causes

Disease.rate Annual rate of deaths from preventable or mitagable zymotic diseases, per 1000

Wounds.rate Annual rate of deaths directly from battle wounds, per 1000

Other.rate Annual rate of deaths from other causes, per 1000

### Details

For a given cause of death, D, annual rates per 1000 are calculated as  $12 * 1000 * D / \text{Army}$ , rounded to 1 decimal.

The two panels of Nightingale's Coxcomb correspond to dates before and after March 1855

### Source

The data were obtained from:

Pearson, M. and Short, I. (2007). Understanding Uncertainty: Mathematics of the Coxcomb. <http://understandinguncertainty.org/node/214>.

## References

Nightingale, F. (1858) *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army* Harrison and Sons, 1858

Nightingale, F. (1859) *A Contribution to the Sanitary History of the British Army during the Late War with Russia* London: John W. Parker and Son.

Small, H. (1998) Florence Nightingale's statistical diagrams <http://www.florence-nightingale-avenging-angel.co.uk/GraphicsPaper/Graphics.htm>

Pearson, M. and Short, I. (2008) Nightingale's Rose (flash animation). <http://understandinguncertainty.org/files/animations/Nightingale11/Nightingale1.html>

## Examples

```
data(Nightingale)

# For some graphs, it is more convenient to reshape death rates to long format
# keep only Date and death rates
require(reshape)
Night<- Nightingale[,c(1,8:10)]
melted <- melt(Night, "Date")
names(melted) <- c("Date", "Cause", "Deaths")
melted$Cause <- sub("\\.rate", "", melted$Cause)
melted$Regime <- ordered( rep(c(rep('Before', 12), rep('After', 12)), 3), levels=c('Before', 'After'))
Night <- melted

# subsets, to facilitate separate plotting
Night1 <- subset(Night, Date < as.Date("1855-04-01"))
Night2 <- subset(Night, Date >= as.Date("1855-04-01"))

## Not run:
require(ggplot2)
# Before plot
cxc1 <- ggplot(Night1, aes(x = factor(Date), y=Deaths, fill = Cause)) +
# do it as a stacked bar chart first
  geom_bar(width = 1, position="identity", color="black") +
# set scale so area ~ Deaths
  scale_y_sqrt()
# A coxcomb plot = bar chart + polar coordinates
cxc1 + coord_polar(start=3*pi/2) +
opts(title="Causes of Mortality in the Army in the East") +
xlab("")

# After plot
cxc2 <- ggplot(Night2, aes(x = factor(Date), y=Deaths, fill = Cause)) +
  geom_bar(width = 1, position="identity", color="black") +
  scale_y_sqrt()
cxc2 + coord_polar(start=3*pi/2) +
opts(title="Causes of Mortality in the Army in the East") +
xlab("")

# do both together, with faceting
```

```

cxc <- ggplot(Night, aes(x = factor(Date), y=Deaths, fill = Cause)) +
  geom_bar(width = 1, position="identity", color="black") +
  scale_y_sqrt() +
  facet_grid(. ~ Regime, scales="free", labeller=label_both)
cxc + coord_polar(start=3*pi/2) +
opts(title="Causes of Mortality in the Army in the East") +
xlab("")

## End(Not run)

## What if she had made a set of line graphs?

colors <- c("blue", "red", "black")
with(Nightingale, {
plot(Date, Disease.rate, type="n", col="blue",
ylab="Annual Death Rate", xlab="Date", xaxt="n",
main="Causes of Mortality of the British Army in the East");
# background, to separate before, after
rect(as.Date("1854/4/1"), -10, as.Date("1855/3/1"),
1.02*max(Disease.rate), col="lightgray", border="transparent");
text( as.Date("1854/4/1"), .98*max(Disease.rate), "Before Sanitary\nCommission", pos=4);
text( as.Date("1855/4/1"), .98*max(Disease.rate), "After Sanitary\nCommission", pos=4);
# plot the data
points(Date, Disease.rate, type="b", col=colors[1]);
points(Date, Wounds.rate, type="b", col=colors[2]);
points(Date, Other.rate, type="b", col=colors[3])
}
)
# add custom Date axis and legend
axis.Date(1, at=seq(as.Date("1854/4/1"), as.Date("1856/3/1"), "4 months"), format="%b %Y")
legend(as.Date("1855/10/20"), 700, c("Disease", "Wounds", "Other"),
col=colors, fill=colors, title="Cause")

```

## Description

The data set is concerned with the problem of aligning the coordinates of points read from old maps (1688 - 1818) of the Great Lakes area. 39 easily identifiable points were selected in the Great Lakes area, and their (lat, long) coordinates were recorded using a grid overlaid on each of 11 old maps, and using linear interpolation.

It was conjectured that maps might be systematically in error in five key ways: (a) constant error in latitude; (b) constant error in longitude; (c) proportional error in latitude; (d) proportional error in longitude; (e) angular error from a non-zero difference between true North and the map's North.

One challenge from these data is to produce useful analyses and graphical displays that relate to these characteristics or to other aspects of the data.

**Usage**

```
data(OldMaps)
```

**Format**

A data frame with 468 observations on the following 6 variables, giving the latitude and longitude of 39 points recorded from 12 sources (Actual + 11 maps).

`point` a numeric vector

`col` Column in the table a numeric vector

`name` Name of the map maker, using Actual for the true coordinates of the points. A factor with levels Actual Arrowsmith Belin Cary Coronelli D'Anville} \code{Del'Isle Lattre Melish Mitchell Popple

`year` Year of the map

`lat` Latitude

`long` Longitude

**Details**

Some of the latitude and longitude values are inexplicably negative. It is probable that this is an error in type setting, because the table footnote says "\*" denotes that interpolation accuracy is not good," yet no "\*"s appear in the body of the table.

**Source**

Andrews, D. F., and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many fields for the Student and Research Worker*. New York: Springer, Table 10.1. The data were obtained from <http://www.stat.duke.edu/courses/Spring01/sta114/data/Andrews/T10.1>.

**Examples**

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

with(OldMaps, plot(abs(long),abs(lat), pch=col, col=colors()[point]))
```

**Description**

Wachsmuth et. al (2003) noticed that a loess smooth through Galton's data on heights of mid-parents and their offspring exhibited a slightly non-linear trend, and asked whether this might be due to Galton having pooled the heights of fathers and mothers and sons and daughters in constructing his tables and graphs.

To answer this question, they used analogous data from English families at about the same time, tabulated by Karl Pearson and Alice Lee (1896, 1903), but where the heights of parents and children were each classified by gender of the parent.

**Usage**

```
data(PearsonLee)
```

**Format**

A frequency data frame with 746 observations on the following 6 variables.

child child height in inches, a numeric vector

parent parent height in inches, a numeric vector

frequency a numeric vector

gp a factor with levels fd fs md ms

par a factor with levels Father Mother

chl a factor with levels Daughter Son

**Details**

The variables gp, par and chl are provided to allow stratifying the data according to the gender of the father/mother and son/daughter.

**Source**

Pearson, K. and Lee, A. (1896). Mathematical contributions to the theory of evolution. On telegony in man, etc. *Proceedings of the Royal Society of London*, 60 , 273-283.

Pearson, K. and Lee, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometika*, 2(4), 357-462. (Tables XXII, p. 415; XXV, p. 417; XXVIII, p. 419 and XXXI, p. 421.)

**References**

Wachsmuth, A.W., Wilkinson L., Dallal G.E. (2003). Galton's bend: A previously undiscovered nonlinearity in Galton's family stature regression data. *The American Statistician*, 57, 190-192. <http://www.cs.uic.edu/~wilkinson/Publications/galton.pdf>

**See Also**

[Galton](#)

**Examples**

```

data(PearsonLee)
str(PearsonLee)

with(PearsonLee,
{
lim <- c(55,80)
xv <- seq(55,80, .5)
sunflowerplot(parent,child, number=frequency, xlim=lim, ylim=lim, seg.col="gray", size=.1)
abline(lm(child ~ parent, weights=frequency), col="blue", lwd=2)
lines(xv, predict(loess(child ~ parent, weights=frequency), data.frame(parent=xv)), col="blue", lwd=2)
# NB: dataEllipse doesn't take frequency into account
if(require(car)) {
dataEllipse(parent,child, xlim=lim, ylim=lim, plot.points=FALSE)
}
})

## separate plots for combinations of (chl, par)

# this doesn't quite work, because xyplot can't handle weights
require(lattice)
xyplot(child ~ parent|par+chl, data=PearsonLee, type=c("p", "r", "smooth"), col.line="red")

# Using ggplot [thx: Dennis Murphy]
require(ggplot2)
ggplot(PearsonLee, aes(x = parent, y = child, weight=frequency)) +
  geom_point(size = 1.5, position = position_jitter(width = 0.2)) +
  geom_smooth(method = lm, aes(weight = PearsonLee$frequency,
    colour = 'Linear'), se = FALSE, size = 1.5) +
  geom_smooth(aes(weight = PearsonLee$frequency,
    colour = 'Loess'), se = FALSE, size = 1.5) +
  facet_grid(chl ~ par) +
  scale_colour_manual(breaks = c('Linear', 'Loess'),
    values = c('green', 'red')) +
  opts(legend.position = c(0.14, 0.885),
    legend.background = theme_rect(fill = 'white'))

# inverse regression, as in Wachmuth et al. (2003)

ggplot(PearsonLee, aes(x = child, y = parent, weight=frequency)) +
  geom_point(size = 1.5, position = position_jitter(width = 0.2)) +
  geom_smooth(method = lm, aes(weight = PearsonLee$frequency,
    colour = 'Linear'), se = FALSE, size = 1.5) +
  geom_smooth(aes(weight = PearsonLee$frequency,
    colour = 'Loess'), se = FALSE, size = 1.5) +
  facet_grid(chl ~ par) +
  scale_colour_manual(breaks = c('Linear', 'Loess'),
    values = c('green', 'red')) +
  opts(legend.position = c(0.14, 0.885),
    legend.background = theme_rect(fill = 'white'))

```

**Description**

The data frame `PolioTrials` gives the results of the 1954 field trials to test the Salk polio vaccine (named for the developer, Jonas Salk), conducted by the National Foundation for Infantile Paralysis (NFIP). It is adapted from data in the article by Francis et al. (1955). There were actually two clinical trials, corresponding to two statistical designs (`Experiment`), discussed by Brownlee (1955). The comparison of designs and results represented a milestone in the development of randomized clinical trials.

**Usage**

```
data(PolioTrials)
```

**Format**

A data frame with 8 observations on the following 6 variables.

`Experiment` a factor with levels `ObservedControl` `RandomizedControl`

`Group` a factor with levels `Controls` `Grade2NotInoculated` `IncompleteVaccinations` `NotInoculated` `Placebo` `Vaccinated`

`Population` the size of the population in each group in each experiment

`Paralytic` the number of cases of paralytic polio observed in that group

`NonParalytic` the number of cases of paralytic polio observed in that group

`FalseReports` the number of cases initially reported as polio, but later determined not to be polio in that group

**Details**

The data frame is in the form of a single table, but actually comprises the results of two separate field trials, given by `Experiment`. Each should be analyzed separately, because the designs differ markedly.

The original design (`Experiment == "ObservedControl"`) called for vaccination of second-graders at selected schools in selected areas of the country (with the consent of the children's parents, of course). The `Vaccinated` second-graders formed the treatment group. The first and third-graders at the schools were not given the vaccination, and formed the `Controls` group.

In the second design (`Experiment == "RandomizedControl"`) children were selected (again in various schools in various areas), all of whose parents consented to vaccination. The sample was randomly divided into treatment (`Group == "Vaccinated"`), given the real polio vaccination, and control groups (`Group == "Placebo"`), a placebo dose that looked just like the real vaccine. The experiment was also double blind: neither the parents of a child in the study nor the doctors treating the child knew which group the child belonged to.

In both experiments, `NotInoculated` refers to children who did not participate in the experiment. `IncompleteVaccinations` refers to children who received one or two, but not all three administrations of the vaccine.

**Source**

Kyle Siegrist, "Virtual Laboratories in Probability and Statistics", <http://www.math.uah.edu/stat/data/Polio.xhtml>

Thomas Francis, Robert Korn, et al. (1955). "An Evaluation of the 1954 Poliomyelitis Vaccine Trials", *American Journal of Public Health*, 45, (50 page supplement with a 63 page appendix).

**References**

K. A. Brownlee (1955). "Statistics of the 1954 Polio Vaccine Trials", *Journal of the American Statistical Association*, 50, 1005-1013.

**Examples**

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

---

Prostitutes	<i>Parent-Duchatelet's time-series data on the number of prostitutes in Paris</i>
-------------	---

---

**Description**

A table indicating month by month, for the years 1812-1854, the number of prostitutes on the registers of the administration of the city of Paris.

**Usage**

```
data(Prostitutes)
```

**Format**

A data frame with 516 observations on the following 5 variables.

Year a numeric vector

month a factor with levels Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep

count a numeric vector: number of prostitutes

mon a numeric vector: numeric month

date a Date

**Details**

The data table was digitally scanned with OCR, and errors were corrected by comparing the yearly totals recorded in the table to the row sums of the scanned data.

**Source**

Parent-Duchatelet, A. (1857), *De la prostitution dans la ville de Paris*, 3rd ed, p. 32, 36

**Examples**

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

---

Pyx

*Trial of the Pyx*


---

**Description**

Stigler (1997, 1999) recounts the history of one of the oldest continuous schemes of sampling inspection carried out by the Royal Mint in London for about eight centuries. The Trial of the Pyx was the final, ceremonial stage in a process designed to ensure that the weight and quality of gold and silver coins from the mint met the standards for coinage.

At regular intervals, coins would be taken from production and deposited into a box called the Pyx. When a Trial of the Pyx was called, the contents of the Pyx would be counted, weighed and assayed for content, and the results would be compared with the standard set for the Royal Mint.

The data frame Pyx gives the results for the year 1848 (Great Britain, 1848) in which 10,000 gold sovereigns were assayed. The coins in each bag were classified according to the deviation from the standard content of gold for each coin, called the Remedy,  $R = 123 * (12/5760) = .25625$ , in grains, for a single sovereign.

**Usage**

```
data(Pyx)
```

**Format**

A frequency data frame with 72 observations on the following 4 variables giving the distribution of 10,000 sovereigns, classified according to the Bags in which they were collected and the Deviation from the standard weight.

Bags an ordered factor with levels 1 and 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9 < 10

Group an ordered factor with levels below std < near std < above std

Deviation an ordered factor with levels Below  $-R < (-R \text{ to } -.2) < (-.2 \text{ to } -.1) < (-.1 \text{ to } 0) < (0 \text{ to } .1) < (.1 \text{ to } .2) < (.2 \text{ to } R) < \text{Above } R$

count number of sovereigns

**Details**

Bags 1-4 were selected as "near to standard", bags 5-7 as below standard, bags 8-10 as above standard. This classification is reflected in Group.

**Source**

Stigler, S. M. (1999). *Statistics on the Table*. Cambridge, MA: Harvard University Press, table 21.1.

## References

Great Britain (1848). "Report of the Commissioners Appointed to Inquire into the Constitution, Management and Expense of the Royal Mint." In Vol 28 of *House Documents for 1849*.

Stigler, S. M. (1997). Eight Centuries of Sampling Inspection: The Trial of the Pyx *Journal of the American Statistical Association*, 72(359), 493-500

## Examples

```
data(Pyx)
# display as table
xtabs(count ~ Bags+Deviation, data=Pyx)
```

---

Quarrels

*Statistics of Deadly Quarrels*

---

## Description

*The Statistics Of Deadly Quarrels* by Lewis Fry Richardson (1960) is one of the earlier attempts at quantification of historical conflict behavior.

The data set contains 779 dyadic deadly quarrels that cover a time period from 1809 to 1949. A quarrel consists of one pair of belligerents, and is identified by its beginning date and magnitude (log 10 of the number of deaths). Neither actor in a quarrel is identified by name.

Because Richardson took a dyad of belligerents as his unit, a given war, such as World War I or World War II comprises multiple observations, for all pairs of belligerents. For example, there are forty-four pairs of belligerents coded for World War I.

For each quarrel, the nominal variables include the type of quarrel, as well as political, cultural, and economic similarities and dissimilarities between the pair of combatants.

## Usage

```
data(Quarrels)
```

## Format

A data frame with 779 observations on the following 84 variables.

```
ID   V84: Id sequence
year  V1: Begin date of quarrel
international V2: Nation vs nation
colonial V3: Nation vs colony
revolution V4: Revolution or civil war
nat.grp V5: Nation vs gp in other nation
grp.grpSame V6: Grp vs grp (same nation)
grp.grpDif V7: Grp vs grp (between nations)
```

numGroups V8: Num grps agst which fighting  
months V9: Num months fighting  
pairs V10: Num pairs in whole matrix  
monthsPairs V11: Num mons for all in mtrx  
logDeaths V12: Log (killed) matrix  
deaths V13: Total killed for matrix  
exchangeGoods V14: Gp sent goods to other  
obstacleGoods V15: Gp puts obstacles to goods  
intermarriageOK V16: Present intermarriages  
intermarriageBan V17: Intermarriages banned  
simBody V18: Similar body characteristics  
difBody V19: Difference in body characteristics  
simDress V20: Similarity of customs (dress)  
difDress V21: Difference of customs (dress)  
eqWealth V22: Common level of wealth  
difWealth V23: Difference in wealth  
simMariagCust V24: Similar marriage cusomst  
difMariagCust V25: Different marriage customs  
simRelig V26: Similar religeon or philosophy of life  
difRelig V27: Religeon or philisophy felt different  
philanthropy V28: General philanthropy  
restrictMigration V29: Restricted immigrations  
sameLanguage V30: Common mother tongue  
difLanguage V31: Different languages  
simArtSci V32: Similar science, arts  
travel V33: Travel  
ignorance V34: Ignorant of other/both  
simPersLiberty V35: Personal liberty similar  
difPersLiberty V36: More personal liberty  
sameGov V37: Common government  
sameGovYrs V38: Years since common govt established  
prevConflict V39: Belligerents fought previously  
prevConflictYrs V40: Years since belligerents fought  
chronicFighting V41: Chronic figthing between belligerents  
persFriendship V42: Autocrats personal friends  
persResentment V43: Leaders personal resentment  
difLegal V44: Annoyingly different legal systems

nonintervention V45: Policy of nonintervention  
thirdParty V46: Led by 3rd group to conflict  
supportEnemy V47: Supported others enemy  
attackAlly V48: Attacked ally of other  
rivalsLand V49: Rivals territory concess  
rivalsTrade V50: Rivals trade  
churchPower V51: Church civil power  
noExtension V52: Policy not extending ter  
territory V53: Desired territory  
habitation V54: Wanted habitation  
minerals V55: Desired minerals  
StrongHold V56: Wanted strategic stronghold  
taxation V57: Taxed other  
loot V58: Wanted loot  
objectedWar V59: Objected to war  
enjoyFight V60: Enjoyed fighting  
pride V61: Elated by strong pride  
overpopulated V62: Insufficient land for population  
fightForPay V63: Fought only for pay  
joinWinner V64: Desired to join winners  
otherDesiredWar V65: Quarrel desired by other  
propaganda3rd V66: Issued of propaganda to third parties  
protection V67: Offered protection  
sympathy V68: Sympathized under control  
debt V69: Owed money to others  
prevAllies V70: Had fought as allies  
yearsAllies V71: Years since fought as allies  
intermingled V72: Had intermingled on territory  
interbreeding V73: Interbreeding between groups  
propadanda V74: Issued propaganda to other group  
orderedObey V75: Ordered other to obey  
commerceOther V76: Commercial enterprises  
feltStronger V77: Felt stronger  
competeIntellect V78: Competed succesfully intellectual occ  
insecureGovt V79: Government insecure  
prepWar V80: Preparations for war  
RegionalError V81: Regional error measure  
CasualtyError V82: Casualty error measure  
Auxiliaries V83: Auxiliaries in service of nation at war

## Details

In the original data set obtained from ICPSR, variables were named V1-V84. These were renamed to make them more meaningful. V84, renamed ID was moved to the first position, but otherwise the order of variables is the same.

In many of the factor variables, 0 is used to indicate "irrelevant to quarrel". This refers to those relations that Richardson found absent or irrelevant to the particular quarrel, and did not subsequently mention.

See the original codebook at [http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=5407&ds=1&file\\_id=652814](http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=5407&ds=1&file_id=652814) for details not contained here.

## Source

<http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/05407>

## References

Lewis F. Richardson, (1960). *The Statistics Of Deadly Quarrels*. (Edited by Q. Wright and C. C. Lienau). Pittsburgh: Boxwood Press.

Rummel, Rudolph J. (1967), "Dimensions of Dyadic War, 1820-1952." *Journal of Conflict Resolution*. 11, (2), 176 - 183.

## Examples

```
data(Quarrels)
str(Quarrels)
```

---

Snow

*John Snow's map and data on the 1854 London Cholera outbreak*

---

## Description

The Snow data consists of the relevant 1854 London streets, the location of 578 deaths from cholera, and the position of 13 water pumps (wells) that can be used to re-create John Snow's map showing deaths from cholera in the area surrounding Broad Street, London in the 1854 outbreak. Another data frame provides boundaries of a tessellation of the map into Thiessen (Voronoi) regions which include all cholera deaths nearer to a given pump than to any other.

The apochrophal story of the significance of Snow's map is that, by closing the Broad Street pump (by removing its handle), Dr. Snow stopped the epidemic, and demonstrated that cholera is a water borne disease. The method of contagion of cholera was not previously understood. Snow's map is the most famous and classical example in the field of medical cartography, even if it didn't happen exactly this way. At any rate, the map, together with various statistical annotations, is compelling because it points to the Broad Street pump as the source of the outbreak.

**Usage**

```
data(Snow.deaths)
data(Snow.pumps)
data(Snow.streets)
data(Snow.polygons)
```

**Format**

`Snow.deaths`: A data frame with 578 observations on the following 3 variables, giving the address of a person who died from cholera. When many points are associated with a single street address, they are "stacked" in a line away from the street so that they are more easily visualized. This is how they are displayed on John Snow's original map. The dates of the deaths are not recorded.

`case` Sequential case number, in some arbitrary, randomized order

`x` x coordinate

`y` y coordinate

`Snow.pumps`: A data frame with 13 observations on the following 4 variables, giving the locations of water pumps within the boundaries of the map.

`pump` pump number

`label` pump label: Briddle St Broad St ... Warwick

`x` x coordinate

`y` y coordinate

`Snow.streets`: A data frame with 1241 observations on the following 4 variables, giving coordinates used to draw the 528 street segment lines within the boundaries of the map. The map is created by drawing lines connecting the `n` points in each street segment.

`street` street segment number: 1:528

`n` number of points in this street line segment

`x` x coordinate

`y` y coordinate

`Snow.polygons`: A data frame with 54 observations on the following 3 variables, giving Thiessen (Voronoi) polygons containing each pump. Their boundaries define the area that is closest to each pump relative to all other pumps. They are mathematically defined by the perpendicular bisectors of the lines between all pumps. The outlines of these polygons can be drawn by connecting all points sequentially starting at each value of `start==0`. Here, each line segment consists of two sequential points.

`start` line start indicator. The value `start==0` indicates the start of a new line, including all following points having `start==1`

`x` x coordinate

`y` y coordinate

## Details

The scale of the source map is approx. 1:2000. The (x, y) coordinate units are 100 meters, with an arbitrary origin.

One limitation of these data sets is the lack of exact street addresses. Another is the lack of any data that would serve as a population denominator to allow for a comparison of mortality rates in the Broad Street pump area as opposed to others. A third is the lack of dates of death that could allow analysis of the time course of the outbreak. See Koch (2000), Koch (2004), Koch & Denike (2009) and Tufte (1999), p. 27-37, for further discussion.

## Source

Tobler, W. (1994). Snow's Cholera Map <http://www.ncgia.ucsb.edu/pubs/snow/snow.html>; data files were obtained from <http://ncgia.ucsb.edu/Publications/Software/cholera/>

The data in these files were first digitized in 1992 by Rusty Dodson of the NCGIA, Santa Barbara, from the map included in the book by John Snow: "Snow on Cholera...", London, Oxford University Press, 1936.

## References

Koch, T. (2000). *Cartographies of Disease: Maps, Mapping, and Medicine*. ESRI Press. ISBN: 9781589481206.

Koch, T. (2004). The Map as Intent: Variations on the Theme of John Snow *Cartographica*, 39 (4), 1-14.

Koch, T. and Denike, K. (2009). Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854. *Social Science & Medicine* 69, 1246-1251.

Tufte, E. (1997). *Visual Explanations*. Cheshire, CT: Graphics Press.

## Examples

```
data(Snow.deaths); data(Snow.pumps); data(Snow.streets); data(Snow.polygons)

## draw a rough approximation to Snow's map and data

# define some functions to make the pieces re-usable
Sdeaths <- function(col="red", pch=15, cex=0.6) {
# make sure that the plot limits include all the other stuff
  plot(Snow.deaths[,c("x","y")], col=col, pch=pch, cex=cex,
        xlab="", ylab="", xlim=c(3,20), ylim=c(3,20),
        main="Snow's Cholera Map of London")
}
# function to plot and label the pump locations
Spumps <- function(col="blue", pch=17, cex=1.5) {
  points(Snow.pumps[,c("x","y")], col=col, pch=pch, cex=cex)
  text(Snow.pumps[,c("x","y")], labels=Snow.pumps$label, pos=1, cex=0.8)
}

# function to draw the streets
Sstreets <- function(col="gray") {
slist <- split(Snow.streets[,c("x","y")],as.factor(Snow.streets[, "street"]))
```

```

invisible(lapply(slist, lines, col=col))
}

# draw a scale showing distance in meters in upper left
mapscale <- function(xs=3.5, ys=19.7) {
  scale <- matrix(c(0,0, 4,0, NA, NA), nrow=3, ncol=2, byrow=TRUE)
  colnames(scale)<- c("x","y")
# tick marks
  scale <- rbind(scale, expand.grid(y=c(-.1, .1, NA), x=0:4)[,2:1])
  lines(xs+scale[,1], ys+scale[,2])
# value and axis labels
  stext <- matrix(c(0,0, 2,0, 4,0, 4, 0.1), nrow=4, ncol=2, byrow=TRUE)
  text(xs+stext[,1], ys+stext[,2], labels=c("0", "2", "4", "100 m."), pos=c(1,1,1,4), cex=0.8)
}

# draw the map with the pieces
Sdeaths()
Spumps()
Sstreets()
mapscale()

# draw the Thiessen polygon boundaries
starts <- which(Snow.polygons$start==0)
for(i in 1:length(starts)) {
  this <- starts[i):(starts[i]+1)
  lines(Snow.polygons[this,2:3], col="blue", lwd=2, lty=2)
}

## overlay bivariate kernel density contours of deaths
Sdeaths()
Spumps()
Sstreets()
mapscale()
require(KernSmooth)
kde2d <- bkde2D(Snow.deaths[,2:3], bandwidth=c(0.5,0.5))
contour(x=kde2d$x1, y=kde2d$x2, z=kde2d$fhat, add=TRUE)

## re-do this the sp way... [thx: Stephane Dray]

library(sp)

# streets
slist <- split(Snow.streets[,c("x","y")], as.factor(Snow.streets[, "street"]))
Ll1 <- lapply(slist, Line)
Ls11 <- Lines(Ll1, "Street")
Snow.streets.sp <- SpatialLines(list(Ls11))
plot(Snow.streets.sp, col="gray")
title(main="Snow's Cholera Map of London (sp)")

# deaths
Snow.deaths.sp = SpatialPoints(Snow.deaths[,c("x","y")])

```

```
plot(Snow.deaths.sp, add=TRUE, col = 'red', pch=15, cex=0.6)

# pumps
spp <- SpatialPoints(Snow.pumps[,c("x", "y")])
Snow.pumps.sp <- SpatialPointsDataFrame(spp, Snow.pumps[,c("x", "y")])
plot(Snow.pumps.sp, add=TRUE, col='blue', pch=17, cex=1.5)
text(Snow.pumps[,c("x", "y")], labels=Snow.pumps$label, pos=1, cex=0.8)
```

---

Wheat

*Playfair's data on wages and the price of wheat*


---

### Description

Playfair (1821) used a graph, showing parallel time-series of the price of wheat and the typical weekly wage for a "good mechanic" from 1570 to 1820 to argue that working men had never been as well-off in terms of purchasing power as they had become toward the end of this period.

His graph is a classic in the history of data visualization, but commits the sin of showing two non-commensurable Y variables on different axes. Scatterplots of wages vs. price or plots of ratios (e.g., wages/price) are in some ways better, but both of these ideas were unknown in 1821.

### Usage

```
data(Wheat)
```

### Format

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

Year Year, in intervals of 10 from 1570 to 1820: a numeric vector

Wheat Price of Wheat (Shillings/Quarter bushel): a numeric vector

Wages Weekly wage (Shillings): a numeric vector

Monarch Reigning monarch: a factor with levels Anne Charles Cromwell Elizabet George I  
James I James II William

### Source

Playfair, W. (1821). *Letter on our Agricultural Distresses, Their Causes and Remedies*. London: W. Sams, 1821

Data values: digitized from <http://www.math.yorku.ca/SCS/Gallery/images/playfair-wheat1.gif>

### References

Friendly, M. & Denis, D. (2005). The early origins and development of the scatterplot *Journal of the History of the Behavioral Sciences*, 41, 103-130.

**Examples**

```
data(Wheat)

# trivial time series plot
Play.ts <- ts(Wheat[,2:3], start=1570, end=1820, deltat=10)
plot(Play.ts, plot.type="single", col=1:2)
```

---

Yeast

*Student's (1906) Yeast Cell Counts*


---

**Description**

Counts of the number of yeast cells were made each of 400 regions in a 20 x 20 grid on a microscope slide, comprising a 1 sq. mm. area. This experiment was repeated four times, giving samples A, B, C and D.

Student (1906) used these data to investigate the errors in random sampling. He says "there are two sources of error: (a) the drop taken may not be representative of the bulk of the liquid; (b) the distribution of the cells over the area which is examined is never exactly uniform, so that there is an 'error of random sampling.'"

The data in the paper are provided in the form of discrete frequency distributions for the four samples. Each shows the frequency distribution squares containing a count of 0, 1, 2, ... yeast cells. These are combined here in Yeast. In addition, he gives a table (Table I) showing the actual number of yeast cells counted in the 20 x 20 grid for sample D, given here as YeastD.mat.

**Usage**

```
data(Yeast)
data(YeastD.mat)
```

**Format**

Yeast: A frequency data frame with 36 observations on the following 3 variables, giving the frequencies of

sample Sample identifier, a factor with levels A B C D

count The number of yeast cells counted in a square

freq The number of squares with the given count

YeastD.mat: A 20 x 20 matrix containing the count of yeast cells in each square for sample D.

**Details**

Student considers the distribution of a total of  $Nm$  particles distributed over  $N$  unit areas with an average of  $m$  particles per unit area. With uniform mixing, for a given particle, the probability of it falling on any one area is  $p = 1/N$ , and not falling on that area is  $q = 1 - 1/N$ . He derives the probability distribution of 0, 1, 2, 3, ... particles on a single unit area from the binomial expansion of  $(p + q)^{mN}$ .

**Source**

D. J. Hand, F. Daly, D. Lunn, K. McConway and E. Ostrowski (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall. The data may be found at <http://www.stat.duke.edu/courses/Spring98/sta113/Data/Hand/yeast.dat>

**References**

"Student" (1906) On the error of counting with a haemocytometer. *Biometrika*, 5, 351-360. <http://www.jstor.org/stable/pdfplus/2331633.pdf>

**Examples**

```
data(Yeast)

require(lattice)
# basic bar charts
# TODO: frequencies should start at 0, not 1.
barchart(count~freq|sample, data=Yeast, ylab="Number of Cells", xlab="Frequency")
barchart(freq~count|sample, data=Yeast, xlab="Number of Cells", ylab="Frequency",
horizontal=FALSE, origin=0)

# same, using xyplot
xyplot(freq~count|sample, data=Yeast, xlab="Number of Cells", ylab="Frequency",
horizontal=FALSE, origin=0, type="h", lwd=10)
```

---

ZeaMays

---

*Darwin's Heights of Cross- and Self-fertilized Zea May Pairs*


---

**Description**

Darwin (1876) studied the growth of pairs of zea may (aka corn) seedlings, one produced by cross-fertilization and the other produced by self-fertilization, but otherwise grown under identical conditions. His goal was to demonstrate the greater vigour of the cross-fertilized plants. The data recorded are the final height (inches, to the nearest 1/8th) of the plants in each pair.

In the *Design of Experiments*, Fisher (1935) used these data to illustrate a paired t-test (well, a one-sample test on the mean difference, cross - self). Later in the book (section 21), he used this data to illustrate an early example of a non-parametric permutation test, treating each paired difference as having (randomly) either a positive or negative sign.

**Usage**

```
data(ZeaMays)
```

**Format**

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

pair pair number, a numeric vector  
 pot pot, a factor with levels 1 2 3 4  
 cross height of cross fertilized plant, a numeric vector  
 self height of cross fertilized plant, a numeric vector  
 diff cross - self for each pair

**Details**

In addition to the standard paired t-test, several types of non-parametric tests can be contemplated:

(a) Permutation test, where the values of, say self are permuted and  $\text{diff} = \text{cross} - \text{self}$  is calculated for each permutation. There are  $15!$  permutations, but a reasonably large number of random permutations would suffice. But this doesn't take the paired samples into account.

(b) Permutation test based on assigning each  $\text{abs}(\text{diff})$  a + or - sign, and calculating the mean(diff). There are  $2^{15}$  such possible values. This is essentially what Fisher proposed. The p-value for the test is the proportion of absolute mean differences under such randomization which exceed the observed mean difference.

(c) Wilcoxon signed rank test: tests the hypothesis that the median signed rank of the diff is zero, or that the distribution of diff is symmetric about 0, vs. a location shifted alternative.

**Source**

Darwin, C. (1876). *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom*, 2nd Ed. London: John Murray.

Andrews, D. and Herzberg, A. (1985) *Data: a collection of problems from many fields for the student and research worker*. New York: Springer. Data retrieved from: <http://lib.stat.cmu.edu/datasets/Andrews/>

**References**

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver & Boyd.

**See Also**

[wilcox.test](#)

[independence\\_test](#) in the coin package, a general framework for conditional inference procedures (permutation tests)

**Examples**

```
data(ZeaMays)

#####
## Some preliminary exploration ##
#####
```

```

boxplot(ZeaMays[,c("cross", "self")], ylab="Height (in)", xlab="Fertilization")

# examine large individual diff/ces
largediff <- subset(ZeaMays, abs(diff) > 2*sd(abs(diff)))
with(largediff, segments(1, cross, 2, self, col="red"))

# plot cross vs. self. NB: unusual trend and some unusual points
with(ZeaMays, plot(self, cross, pch=16, cex=1.5))
abline(lm(cross ~ self, data=ZeaMays), col="red", lwd=2)

# pot effects ?
anova(lm(diff ~ pot, data=ZeaMays))

#####
## Tests of mean difference ##
#####
# Wilcoxon signed rank test
# signed ranks:
with(ZeaMays, sign(diff) * rank(abs(diff)))
wilcox.test(ZeaMays$cross, ZeaMays$self, conf.int=TRUE, exact=FALSE)

# t-tests
with(ZeaMays, t.test(cross, self))
with(ZeaMays, t.test(diff))

mean(ZeaMays$diff)
# complete permutation distribution of diff, for all 2^15 ways of assigning
# one value to cross and the other to self (thx: Bert Gunter)
N <- nrow(ZeaMays)
allmeans <- as.matrix(expand.grid(as.data.frame(matrix(rep(c(-1,1),N), nr =2)))) %*% abs(ZeaMays$diff) / N

# upper-tail p-value
sum(allmeans > mean(ZeaMays$diff)) / 2^N
# two-tailed p-value
sum(abs(allmeans) > mean(ZeaMays$diff)) / 2^N

hist(allmeans, breaks=64, xlab="Mean difference, cross-self",
main="Histogram of all mean differences")
abline(v=c(1, -1)*mean(ZeaMays$diff), col="red", lwd=2, lty=1:2)

plot(density(allmeans), xlab="Mean difference, cross-self",
main="Density plot of all mean differences")
abline(v=c(1, -1)*mean(ZeaMays$diff), col="red", lwd=2, lty=1:2)

```

# Index

## \*Topic **datasets**

Arbuthnot, 4  
Bowley, 6  
Cavendish, 7  
ChestSizes, 8  
CushnyPebbles, 9  
Dactyl, 11  
DrinksWages, 12  
Fingerprints, 14  
Galton, 15  
Guerry, 16  
Jevons, 19  
Langren, 20  
Macdonell, 23  
Michelson, 28  
Minard, 30  
Nightingale, 31  
OldMaps, 34  
PearsonLee, 35  
PolioTrials, 38  
Prostitutes, 39  
Pyx, 40  
Quarrels, 41  
Snow, 44  
Wheat, 48  
Yeast, 49  
ZeaMays, 50

## \*Topic **nonparametric**

ZeaMays, 50

## \*Topic **package**

HistData-package, 2

## \*Topic **spatial**

Langren, 20  
Minard, 30  
OldMaps, 34  
Snow, 44

## \*Topic

Langren, 20  
Minard, 30

OldMaps, 34

Arbuthnot, 3, 4, 4

Bowley, 3, 4, 6

Cavendish, 3, 4, 7

ChestSizes, 3, 4, 8

CushnyPebbles, 3, 4, 9

CushnyPebblesN (CushnyPebbles), 9

Dactyl, 3, 4, 11

DrinksWages, 3, 4, 12

Fingerprints, 3, 4, 14

Galton, 3, 4, 15, 36

galton, 16

galtonpeas, 4

gfrance, 19

Guerry, 3, 4, 16

Guerry-package, 4

HistData (HistData-package), 2

HistData-package, 2

independence\_test, 51

Jevons, 3, 4, 19

Langren, 3, 4, 20

Langren1644 (Langren), 20

Macdonell, 3, 4, 23

MacdonellDF (Macdonell), 23

Michelson, 3, 4, 28

MichelsonSets (Michelson), 28

Minard, 3, 4, 30

Nightingale, 3, 4, 31

OldMaps, 3, 4, 34

PearsonLee, 3, 4, 16, 35

peas, 4

PolioTrials, 3, 4, 38

Prostitutes, 3, 4, 39

Pyx, 3, 4, 40

Quarrels, 3, 4, 41

sleep, 10

Snow, 3, 4, 44

Wheat, 3, 4, 48

wilcox.test, 51

Yeast, 3, 4, 49

YeastD.mat (Yeast), 49

ZeaMays, 3, 4, 50