Package 'Stat2Data'

December 27, 2012

Type Package

Title Datasets for Stat2

Version 1.6

Date 2012-09-12

Author Robin Lock

Maintainer Robin Lock <rlock@stlawu.edu>

Description Datasets for Stat2 textbook (by Cannon, et. al., published by WH Freeman)

License GPL-3

LazyLoad yes

R topics documented:

Stat2Data-package *DataSets for Stat2 Textbook*

Description

DataSets for Stat2 Textbook (by Cannon, et. al.)

Details

Package:	Stat2Data
Type:	Package
Version:	1.6
Date:	2012-09-11
License:	GPL-2
LazyLoad:	yes

Author(s)

Robin Lock

Maintainer: Robin Lock <rlock@stlawu.edu>

Alfalfa *Alfalfa*

Description

Growth of alfalfa sprouts in acidic conditions

Format

A dataset with 15 observations on the following 3 variables.

`Ht4`	Height of alfalfa sprouts after four days
`Acid`	Amount of acid: `1.5HCl`, `3.0HCl`, or `water`
`Row`	a= closest to window through e=farthest from window

Details

Some students were interested in how an acidic environment might affect the growth of plants. They planted alfalfa seeds in 15 cups and randomly chose five to get plain water, five to get a moderate amount of acid (1.5M HCl), and five to get a stronger acid solution (3.0M HCl). The plants were grown in an indoor room so the students assumed that the distance from the main source of daylight (a window) might have an affect on growth rates. For this reason, they arranged the cups in five rows of three with one cup from each Acid level in each row. These are labeled in the data set as

Row: a= farthest from the window through e=nearest to the window.

Source

Neumann, A., Richards, A.-L., and Randa, J. (2001). Effects of acid rain on alfalfa plants. Unpublished manuscript, Oberlin College.

Examples

```
data(Alfalfa)
```

`ArcheryData` *ArcheryData*

Description

Score results from an archery class

Format

A dataset with 18 observations on the following 7 variables.

`Attendance`	Number of days in class
`Average`	Average score over all days
`Sex`	Coded as `f` or `m`
`Day1`	Archery score on first day
`LastDay`	Archery score on last day
`Improvement`	Last day - first day score
`Improve`	1=improved or 0= did not improve

Details

In 2002, Heather Tollerud, a Saint Olaf College student, undertook a study of the archery scores of students at the college who were enrolled in an archery course. Students taking the course record a score for each day they attend class from the first until the last day. Hopefully the instruction they receive helps them to improve their game.

Source

Student project

`AutoPollution` *AutoPollution*

Description

AutoPollution

Format

A dataset with 36 observations on the following 4 variables.

Noise	Noise level (decibels)
Size	Vehicle size: 1=small, 2=medium, or 3=large
Type	1=standard filer or 2=new filter
Side	Side of vehicle: code1=right or 2=left

Details

In a 1973 testimony before the Air and Water Pollution Subcommittee of the Senate Public Works Committee, John McKinley, President of Texaco discussed a new filter that had been developed to reduce pollution. Questions were raised about the effects of this filter on other measures of vehicle performance. The data set AutoPollution gives the results of an experiment on 36 different cars. The cars were randomly assigned to get either this new filter or a standard filter and the noise level for each car was measured.

Source

Data explanation and link can be found at http://lib.stat.cmu.edu/DASL/Stories/airpollutionfilters.html.

References

A.Y. Lewin and M.F. Shakun, Policy Sciences: Methodology and Cases, Pergammon Press, 1976, p 313.

Backpack *Backpack*

Description

Backpack weights for a sample of college students

Format

A dataset with 100 observations on the following 9 variables.

BackpackWeight	Backpack weight (in pounds)
BodyWeight	Body weight (in pounds)
Ratio	BackpackWeight/BodyWeight
BackProblems	0=no or 1=yes
Major	Code for academic major
Year	Year in school
Sex	a factor with levels Female Male
Status	Graduate or undergraduate? G or U
Units	Number of credits taken that quarter

Details

A survey of students at California Polytechnic State University (San Luis Obispo) collected data to investigate the question of whether back aches might be due to carrying heavy backpacks,

Source

Mintz J., Mintz J., Moore K., and Schuh K., "Oh, My Aching Back! A Statistical Analysis of Backpack Weights," Stats: The Magazine for Students of Statistics, vol. 32, 2002, pp. 1719.

`BaseballTimes` *BaseballTimes*

Description

Game times and boxscore information for baseball games

Format

A dataset with 15 observations on the following 7 variables.

`Game`	Code for opposing teams
`League`	`AL`= American League or `NL`=National League
`Runs`	Total number of runs scored (both teams)
`Margin`	Margin of victory (Winner-Loser score)
`Pitchers`	Total number of pitchers used (both teams
`Attendance`	Number of spectators at the game
`Time`	Total time for the game (in minutes)

Details

Data were collected for 15 Major Leage Baseball (MLB) games played on August 26, 2008.

Source

Data from boxscores at www.baseball-reference.com

`BeeStings` *BeeStings*

Description

Does number of beestings depend on previous stings?

Format

A dataset with 18 observations on the following 3 variables.

`Occasion`	Trial: I to IX
`Treatment`	`Fresh` or `Stung`
`Stingers`	Number of stingers

Details

If you are stung by a bee, does that make you more likely to get stung again? Might bees leave behind a chemical message that tells other bees to attack you? To test this hypothesis, scientists dangled a 4x4 array of 16 muslin-wrapped cotton balls over a beehive. Eight of 16 balls had been previously stung; the other eight were fresh. The response was the total number of new stingers left behind by the bees. The process was repeated for a total of nine tials.

Source

Free, J.B. (1961) "The stinging response of honeybees," Animal Behavior, vol. 9, pp 193-196.

`BirdNest` *BirdNest*

Description

Data on bird nests

Format

A dataset with 84 observations on the following 12 variables.

`Species`	Latin species name
`Common`	Common species name
`Page`	Page in a bird manual describing the species
`Length`	Mean body length for the species (in cm)
`Nesttype`	Type of nest
`Location`	Location of nest
`No.eggs`	Number of eggs
`Color`	a numeric vector
`Incubate`	Mean length of time (in days) the species incubates eggs in the nest
`Nestling`	Mean length of time (in days) the species cares for babies in the nest until fledged
`Totcare`	Total care time = Incubate+Nestling
`Closed`	1=closed nest (pendant, spherical, cavity, crevice, burrow), 0=open nest (saucer, cup)

Details

Amy R. Moore, as a student at Grinnell College in 1999, wanted to study the relationship between species characteristics and the type of nest a bird builds, using data collected from available sources. For the study, she collected data by species for 83 separate species of North American passerines.

Source

Project by Amy Moore at Grinnell College

References

The Birders Handbook, by Ehrlich, et al. (1988)

Blood1 *Blood1*

Description

Systolic blood pressure for a sample of 500 adults

Format

A dataset with 500 observations on the following 3 variables.

SystolicBP	Systolic blood pressure (mm of Hg)
Smoke	Y=smoker or N=non-smoker
Overwt	1=normal, 2=overweight, or 3=obese

Details

Data on systolic blood pressure, along with smoker status and weight status, for a sample of 500 adults.

Source

Data are part of a larger case study for the 2003 Annual Meeting of the Statistical Society of Canada. http://www.ssc.ca/en/education/archived-case-studies/case-studies-for-the-2003-annual-meeting-blood-pressure.

BlueJays *Blue Jays*

Description

Body measurements for a sample of blue jays

Format

A dataset with 123 observations on the following 9 variables.

BirdID	ID tag for bird
KnownSex	Sex coded as F or M
BillDepth	Thickness of the bill measured at the nostril (in mm)
BillWidth	Width of the bill (in mm)
BillLength	Length of the bill (in mm)
Head	Distance from tip of bill to back of head (in mm)
Mass	Body mass (in grams)
Skull	Distance from base of bill to hack of skull (in mm)
Sex	Sex coded as 0=female or 1=male

Details

Body measurements for captured blue jays. Values are averaged for birds captured more than once.

Source

Data from Keith Tarvin, Department of Biology, Oberlin College

---

BritishUnions *BritishUnions*

Description

Poll attitudes towards British trade unions

Format

A dataset with 17 observations on the following 7 variables.

`Date`	Month of the poll `Aug-77` to `Sep-79`
`AgreePct`	Percent who agree (unions have too much power)
`DisagreePct`	Percent who disagree
`NetSupport`	DisagreePct-AgreePct
`Months`	Months since August 1975
`Late`	`1`=after 1986 or `0`=before 1986
`Unemployment`	Unemployment rate

Details

The British polling company Ipsos MORI conducted several opinion polls in the UK between 1975 and 1995 in which they asked whether people agree or disagree with the statement "Trade unions have too much power in Britain today".

Source

Data from the Ipsos MORI website at
http://www.ipsos-mori.com/researchpublications/researcharchive/poll.aspx?oItemID=94

CAFE *CAFE*

Description

Senate votes for Corporate Average Fuel Economy (CAFE) bill

Format

A dataset with 100 observations on the following 7 variables.

`Senator`	Senator's name

State	Code for senator's state
Party	party affiliation: D=Democrat, I=Independent, R=Republican
Contribution	Contributions from car manufactures (dollars)
LogContr	Log of (Contribution+1)
Dem	1=Democrat/Indpendent 0=Republican
Vote	1=yes or 0=no

Details

The Corporate Average Fuel Economy (CAFE) bill was proposed by Senators John McCain and John Kerry to improve the fuel economy of cars and light trucks sold in the United States. However a critical vote on an amendment in March of 2002 threatened to indefinitely postpone CAFE. The amendment charged the National Highway Traffic Safety Administration to develop a new standard, the effect being to put on indefinite hold the McCain-Kerry bill. It passed by a vote of 62-38. A political question of interest is whether there is evidence of monetary influence on a senator's vote. Scott Preston, a professor of statistics at SUNY, Oswego, collected data on this vote which includes the vote of each senator (1=Yes or 0=No) and monetary contributions that each of the 100 senators received over his or her lifetime from the car manufacturers.

Source

Thanks to Prof. Scott Preston for the data.

`CalciumBP` *CalciumBP*

Description

Calcium and blood pressure

Format

A dataset with 21 observations on the following 2 variables.

Treatment	`Calcium` or `Placebo`
Decrease	Beginning-ending blood pressure

Details

The purpose of this study was to see whether daily calcium supplements can lower blood pressure. The subjects were 21 men; each was randomly assigned either to a treatment group or to a control group. Those in the treatment group took a daily pill containing calcium. Those in the control group took a daily pill with no active ingredients. Each subject's blood pressure was measured at the beginning of the 12-week study, and again at the end. The decrease in blood pressure (begin-end) was recorded (so a negative value means blood pressure increased).

Source

Dataset downloaded from online data source Data and Story Library,
http://lib.stat.cmu.edu/DASL/Stories/CalciumandBloodPressure.html

CancerSurvival *CancerSurvival*

Description

Cancer survival with ascorbate supplement

Format

A dataset with 64 observations on the following 2 variables.

`Survival`	Survival time (in days)
`Organ`	`Breast`, `Bronchus`, `Colon`, `Ovary`, or `Stomach`

Details

In the 1970's doctors wondered if giving terminal cancer patients a supplement of ascorbate would prolong their lives. They designed an experiment to compare cancer patients who received ascorbate to cancer patients who did not receive the supplement. The result of that experiment was that, in fact, ascorbate did seem to prolong the lives of these patients. But then a second question arose. Was the effect of the ascorbate different when different organs were affected by the cancer? The researchers took a second look at the data. This time they concentrated only on those patients who received the ascorbate and divided the data up by which organ was affected by the cancer. They had 5 different organs represented among the patients (all of whom only had one organ affected): Stomach, bronchus, colon, ovary, and breast.

Source

From the article "Supplemental Ascorbate in the Supportive Treatment of Cancer: Reevaluation of Prolongation of Survival Times in Terminal Human Cancer" by Ewan Cameron and Linus Pauling, Proceedings of the National Academy of Sciences of the United States of America, Vol. 75, No. 9 (Sep., 1978), pp. 4538-4542.

Caterpillars *Caterpillars*

Description

Measurements on a sample of Manduca Sexta caterpillars

Format

A dataset with 267 observations on the following 18 variables.

`Instar`	Coded from 1 (smallest) to 5 (largest) indicating stage of the caterpillar's life
`ActiveFeeding`	Indicator (`Y` or `N`) of whether or not the animal is actively feeding
`Fgp`	Indicator (`Y` or `N`) of whether or not the animal is in a free growth period
`Mgp`	Indicator (`Y` or `N`) of whether or not the animal is in a maximum growth period
`Mass`	Body mass (in grams)

LogMass	Log (base 10) of body mass
Intake	Wet food intake (in grams/day)
LogIntake	Log (base 10) of Intake
WetFrass	Amount of frass (solid waste) produced (in grams/day)
LogWetFrass	Log (base 10) of WetFrass
DryFrass	Amount of frass, after drying, produced (in grams/day)
LogDryFrass	Log (base 10) of DryFrass
Cassim	CO2 assimilation (ingestion - excretion)
LogCassim	Log (base 10) of Cassim
Nfrass	Nitrogen in frass
LogNfrass	Log (base 10) of Nfrass
Nassim	Nitrogen assimilation (ingestion - excretion)
LogNassim	Log (base 10) of Nassim

Details

Student and faculty researchers at Kenyon College conducted numerous experiments with Manduca Sexta caterpillars to study biological growth.

Source

We thank Professors Harry Itagaki, Drew Kerkhoff, Chris Gillen, and Judy Holdener and their students for sharing this data from research supported by NSF InSTaRs grant #0827208.

Cereal *Cereal*

Description

Breakfast cereals

Format

A dataset with 36 observations on the following 4 variables.

Cereal	Brandname of cereal
Calories	Calories per serving
Sugar	Grams of sugar per serving
Fiber	Grams of fiber per serving

Details

Data give nutrition contents (per serving) for 36 breakfast cereals.

Source

These data were collected by Patricia Benedict, Ronald Brahler, and Kenneth Motz, who read the nutritional labels on the boxes, in an attempt to learn whether cereals high in fiber are also high in sugar and calories. The cereals are all of those that were sold at Russo Stop & Shop in University Heights, OH, in July, 1990.

ChemoTHC *ChemoTHC*

Description

Comparison of two treatments for nausea in chemotherapy

Format

A dataset with 2 observations on the following 4 variables.

Drug	Prochlorperazine or THC
Effective	Count of effectrive cases
NotEffective	Count of noneffective cases
Patients	Number of patients in the treatment

Details

An article in the New England Journal of Medicine described a study on the effectiveness of medications for combatting nausea in patients undergoing chemotherapy treatments for cancer. In the experiment, 157 patients were divided at random into two groups. One group of 78 patients were given a standard antinausea drug called prochlorperazine, while the other group of 79 patients received THC (the active ingredient in marijuana). Both medications were delivered orally and no patients were told which of the two drugs they were taking. The response measured was whether or not the patient experienced relief from nausea when undergoing chemotherapy. Dataset is a 2x2 table of counts.

Source

Sallan SE, Cronin C, Zelen M, Zinberg NE (1980), "Antiemetics in patients receiving chemotherapy for cancer: a randomized comparison of delta-9-tetrahydrocannabinol and prochlorperazine," New England Journal of Medicine, 302(3) p.135-138

ChildSpeaks *ChildSpeaks*

Description

Age at first speaking and aptitude test scores

Format

A dataset with 21 observations on the following 3 variables.

Child	ID for each child
Age	Age at first speaking (in months)
Gesell	Gesell Aptitude Test Score

Details

The data are from a study about whether there is a relationship between the age at which a child first speaks (in months) and his or her score on a Gesell Aptitude Test taken later in childhood.

Source

These data were originally collected by L.M. Linde of UCLA but were first published by M.R. Mickey, O.J. Dunn, and V. Clark, "Note on the use of stepwise regression in detecting outliers," Computers and Biomedical Research, 1 (1967), pp. 105-111. The data have been used by several authors. We found them in David Moore's Basic Practice of Statistics, WH Freeman (2004)

Clothing *Clothing*

Description

Clothing retailer

Format

A dataset with 60 observations on the following 8 variables.

ID	Case ID
Amount	Net dollar amount spent by customers in their latest purchase from this retailer
Recency	Number of months since the last purchase
Freq12	Number of purchases in the last 12 months
Dollar12	Dollar amount of purchases in the last 12 months
Freq24	Number of purchases in the last 24 months
Dollar24	Dollar amount of purchases in the last 24 months
Card	1 for customers who have a private-label credit card with the retailer, 0 if not

Details

This dataset represents a random sample of 60 customers from a large clothing retailer. The manager of the store is interested in predicting how much a customer will spend on his or her next purchase based on one or more of the available explanatory variables.

Source

Personal communication from clothing retailer David Cameron

CloudSeeding *Cloud Seeding*

Description

Rainfall amounts from a cloud seeding experiment (winter only)

Format

A dataset with 28 observations on the following 7 variables.

`Seeded`	Treatment coded as `S`=seded or `U`=unseeded
`Season`	All in `Winter`
`TE`	Rainfall in East (treatment)
`TW`	Rainfall in West (treatment
`NC`	Rainfall in North (control)
`SC`	Rainfall in South (control)
`NWC`	Rainfall in Northwest (control)

Details

Researchers were interested in whether seeded clouds would produce more rainfall. An experiment was conducted in Tasmania between 1964 and 1971 and rainfall amounts were measured in inches per rainfall period. The researchers measured the amount of rainfall in two target areas: East (TE) and West (TW). They also measured the amount of rainfall in three control locations. Clouds were coded as being either seeded (treatment) or unseeded (control). This is a subset (only Winter months) of the larger CloudSeeding2 dataset. All rainfall amounts are in inches.

Source

Data were accessed from the website www.statsci.org/data/oz/cloudtas.html. This is the web home of the Australasian Data and Story Library (OzDASL).

References

A.J. Miller, D.E. Shaw, L.G. Veitch, and E.J. Smith, (1979) "Analyzing the results of a cloud-seeding experiment in Tasmania" in Communications in Statistics: Theory and Methods, A8 (10), pp. 1017-1047.

`CloudSeeding2` *Cloud Seeding 2*

Description

Rainfall amounts from a cloud seeding experiment

Format

A dataset with 108 observations on the following 8 variables.

`Period`	ID for time period
`Seeded`	Treatment coded as `S`=seded or `U`=unseeded
`Season`	Coded as `Autumn`, `Spring`, `Summer`, or `Winter`
`TE`	Rainfall in East (treatment)
`TW`	Rainfall in West (treatment
`NC`	Rainfall in North (control)
`SC`	Rainfall in South (control)
`NWC`	Rainfall in Northwest (control)

Details

Researchers were interested in whether seeded clouds would produce more rainfall. An experiment was conducted in Tasmania between 1964 and 1971 and rainfall amounts were measured in inches per rainfall period. The researchers measured the amount of rainfall in two target areas: East (TE) and West (TW). They also measured the amount of rainfall in three control locations. Clouds were coded as being either seeded (treatment) or unseeded (control). A subset (only Winter months) of these data is stored in CloudSeeding. All rainfall amounts are in inches.

Source

Data were accessed from the website www.statsci.org/data/oz/cloudtas.html. This is the web home of the Australasian Data and Story Library (OzDASL).

References

A.J. Miller, D.E. Shaw, L.G. Veitch, and E.J. Smith, (1979) "Analyzing the results of a cloud-seeding experiment in Tasmania" in Communications in Statistics: Theory and Methods, A8 (10), pp. 1017-1047.

`CO2` *CO2*

Description

Daily carbon dioxide measurements - April through November 2011

Format

A dataset with 237 observations on the following 2 variables.

`CO2`	Carbon dioxide (CO2) level (in parts per million)
`Day`	Number of day in 2001 (April 1 = day 91

Details

Scientists at a research station in Brotjacklriegel, Germany recorded CO2 levels, in parts per million, in the atmosphere for each day from the start of April through November in 2001.

Source

http://gaw.empa.ch/gawsis/reports.asp?StationID=-739519191

`CrackerFiber` *Cracker Fiber in Diets*

Description

Digested calories with different types of fiber in crackers

Format

A dataset with 48 observations on the following 3 variables.

`Subj`	ID for the subject
`Fiber`	Type of fiber: `bran`, `combo`, `control`, or `gum`
`Calories`	Digested calories

Details

Twelve female subjects were fed a controlled diet, with crackers before every meal. There were four different kinds of crackers: control, bran fiber, gum fiber, and a combination of both bran and gum fiber. Over the course of the study, each subject ate all four kinds of crackers, one kind at a time, for a stretch of several days. The order was randomized. The response is the number of digested calories, measured as the difference between calories eaten and calories passed through the system.

Source

Subset of the data at http://lib.stat.cmu.edu/DASL/Datafiles/Fiber.html, originally distributed with with the Data Desk software package.

`Cuckoo` *Cuckoo*

Description

Lengths of cuckoo eggs laid in other birds' nests

Format

A dataset with 120 observations on the following 2 variables.

`Bird`	Type of bird nest: `mdw_pipit` (meadow pipit), `tree_pipit`, `hedge_sparrow`, `robin`, `wagtail`, or `wren`
`Length`	Cuckoo egg length (in mm)

Details

Cuckoos are knows to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. The data give the lengths of cuckoo eggs found in nests of varous other bird species.

Source

Downloaded from DASL at http://lib.stat.cmu.edu/DASL/Datafiles/cuckoodat.html

References

"The Egg of Cuculus Canorus. An Enquiry into the Dimensions of the Cuckoo's Egg and the Relation of the Variations to the Size of the Eggs of the Foster-Parent, with Notes on Coloration", by Oswald H. Latter, Biometrika, Vol. 1, No. 2 (Jan., 1902), pp. 164-176.

Day1Survey *Day1Survey*

Description

Data from a first day class survey

Format

A dataset with 43 observations on the following 13 variables.

Section	Section: 1 or 2
Class	Year in school: Freshman, Sophomore, Junior, or Senior
Sex	F=female or M=male
Distance	Distance (in miles) to get to campus
Height	Height (in inches)
Handedness	Left, Right, or Ambidextrous
Coins	Value of coins student has (in class)
WhiteString	Estimated length of a white string (in inches)
BlackString	Estimated length of a black string (in inches)
Reading	Expected amount of reading during the semester (pages/week)
TV	Hours of TV watched per week
Pulse	Resting pulse rate (beats per minute)
Texting	Number of text messages in past 24 hours

Details

An instructor at a small liberal arts college distributed the a data survey on the first day of class. The data for two different sections of the course are given in this dataset.

Source

Student survey in an introductory statistics class.

Diamonds *Diamonds*

Description

Diamonds

Format

A dataset with 351 observations on the following 6 variables.

Carat	Size of the diamond (in carats)
Color	Coded as D(most white/bright) through J
Clarity	Coded as IF, VVS1, VVS2, VS1, VS2, SI1, SI2, or SI3
Depth	Depth (as a percentage of diameter)

PricePerCt	Price per carat
TotalPrice	Price for the diamond (in dollars)

Details

Data for a sample of diamonds. The clarity of the diamonds ranges from IF (internally flawless) through VVS1 (very,very slightly included), VS1 (very slightly included), to SI3 (slightly included) in the order listed above.

Source

Diamond data obtained from AwesomeGems.com on July 28, 2005.

`Diamonds2` *Diamonds2*

Description

A subset of the Diamonds data

Format

A dataset with 307 observations on the following 6 variables.

Carat	Size of the diamond (in carats)
Color	Coded as D(most white/bright) through G
Clarity	Coded as IF, VVS1, VVS2, VS1, VS2, SI1, SI2, or SI3
Depth	Depth (as a percentage of diameter)
PricePerCt	Price per carat
TotalPrice	Price for the diamond (in dollars)

Details

A subset of the Diamonds data, containing only those with most frequent colors D, E, F, and G. The clarity of the diamonds ranges from IF (internally flawless) through VVS1 (very,very slightly included), VS1 (very slightly included), to SI3 (slightly included) in the order listed above.

Source

Diamond data obtained from AwesomeGems.com on July 28, 2005.

`Election08` *Election08*

Description

State-by-state information from the 2008 U.S. presidential election

Format

A dataset with 51 observations on the following 7 variables.

`State`	Name of the state
`Abr`	Abbreviation for the state
`Income`	Per capita income in the state as of 2007 (in dollars)
`HS`	Percentage of adults with at least a high school education
`BA`	Percentage of adults with at least a college education
`Dem.Rep`	Difference in %Democrat-%Republican (according to 2008 Gallup survey)
`ObamaWin`	`1`= Obama (Democrat) wins state in 2008 or `0`=McCain (Republican wins)

Details

This dataset contains information from all 50 states and the District of Columbia for the 2008 U.S. presidential election.

Source

State income data from: Census Bureau Table 659. Personal INcome Per Capita (in 2007)
High school data from: U.S. Census Bureau, 1990 Census of Population, http://nces.ed.gov/programs/digest/d08/tables/d
College data from: Census Bureau Table 225. Educational Attainment by State (in 2007)
% Democrat and % Republican: http://www.gallup.com/poll/114016/state-states-political-party-affiliation.aspx#1

`Ethanol` *Ethanol*

Description

Ethanol

Format

A dataset with 16 observations on the following 3 variables.

`Sugar`	Type of sugar: `Galactose` or `Glucose`
`O2Conc`	Oxygen concentration
`Ethanol`	Ethanol concentration

Details

Many biochemical reactions are slowed or prevented by the presence of oxygen. For example, there are two simple forms of fermentation, one which converts each molecule of sugar to two molecules of lactic acid, and a second which converts each molecule of sugar to one each of lactic acid, ethanol, and carbon dioxide. This experiment was designed to compare the inhibiting effect of oxygen on the metabolism of two different sugars, glucose and galactose, by Streptococcus bacteria. In this case there were four levels of oxygen that were applied to the two kinds of sugar.

Source

Data are found in Statistics: The Exploration and Analysis of Data by Jay Devore and Roxy Peck (1960). St. Paul, MN: West.

References

The original article is "Effects of oxygen concentration on pyruvate formatelyase in situ and sugar metabolism of Streptocucoccus mutans and Streptococcus samguis," Infection and Immunity, pp. 129-134.

FantasyBaseball *FantasyBaseball*

Description

Draft selection times for a fantasy baseball league

Format

A dataset with 24 observations on the following 9 variables.

Round	Round of the draft (1 to 24)
DJ	Draft time (in seconds) for D.J.
AR	Draft time (in seconds) for A.R.
BK	Draft time (in seconds) for B.K.
JW	Draft time (in seconds) for J.W.
TS	Draft time (in seconds) for T.S.
RL	Draft time (in seconds) for R.L.
DR	Draft time (in seconds) for D.R.
MF	Draft time (in seconds) for M.F.

Details

Time (in seconds) for participants in a draft for a fantasy baseball league to make a selectoion at each round.

Source

Mathamatical Science Baseball League historical records (online).

Fertility *Fertility*

Description

Fertility measurements for a sample of women

Format

A dataset with 333 observations on the following 10 variables.

Age	Age (in years)
LowAFC	Smallest antral follicle count
MeanAFC	Average antral follicle count
FSH	Maximum follicle stimulating hormone level
E2	Fertility level
MaxE2	Maximum fertility level
MaxDailyGn	Maximum daily gonadotropin level
TotalGn	Total gonadotropin level
Oocytes	Number of egg cells
Embryos	Number of embryos

Details

A medical doctor and her team of researchers collected a variety of data on women who were having troubling getting pregnant. A key method for assessing fertility is a count of antral follicles (LowAFC or MeanAFC) that can be performed with noninvasive ultrasound. Researchers are interested in how the other variables are related to these counts.

Source

We thank Dr. Priya Maseelall and her research team for sharing these data.

FGByDistance *FGByDistance*

Description

Field goal results in the NFL (by distance)

Format

A dataset with 51 observations on the following 7 variables.

Row	Case ID
Dist	Distance of the attempt (in yards)
N	Number of kicks attempted from that distance
Makes	Nmber of kicks made from that distance
PropMakes	Proportion of attempts made
Blocked	Number of kicks blocked
PropBlocked	Proportion of kicks blocked

Details

This dataset summarizes all 8520 field goals attempted by place kickers in the National Football League (NFL) during regular season games for the 2000 through the 2008 seasons. Results are counts (attempted, made, and blocked) and proportions (made and blocked) for each distance.

Source

We thank Sean Forman and Doug Drinen of Sports Reference LLC for providing us with the NFL field goal data set.

`Film` *Film*

Description

Film data from Maltin's Movie and Video Guide

Format

A dataset with 100 observations on the following 9 variables.

`Title`	Movie title
`Year`	Year the movie was released
`Time`	Running time (in minutes)
`Cast`	Number of cast members listed in the guide
`Rating`	Maltin rating (range is 1 to 4, in steps of 0.5)
`Description`	Number of lines of text Maltin uses to describe the movie
`Origin`	Country: 0 = USA, 1 = Great Britain, 2 = France, 3 = Italy, 4 = Canada
`Time_code`	`long`=90 minues or longer `short`=under 90 minutes
`Good`	`1`=rating or 3 stars or better `0`=any lower rating

Details

One statistician movie fan decided to use statistics to study the movie ratings in his favorite movie guide, Movie and Video Guide (1996), by Leonard Maltin. Maltin rates movies on a one-star to four-star system, in increments of half-stars, with higher numbers being better. The guide also includes additional information on each film. The statistican used a random number generator to select a simple random sample of 100 movies rated by the Guide.

Source

Data from Leonard Maltin's Movie and Video Guide (1996)

`FinalFourIzzo` *FinalFourIzzo*

Description

NCAA Final Four by seed with indicator for Tom Izzo's teams

Format

A dataset with 1664 observations on the following 4 variables.

`Year`	Year 1985-2010
`Seed`	Seed in NCAA men's basketball tournament: 1 to 16
`Final4`	`1`=made Final Four or `0`=did not make Final Four
`Izzo`	`1`=team coached by Tom Izzo or `0`=not an Izzo team

Details

Each year 64 college teams are selected for the NCAA Division I Men's Basketball tournament, with 16 teams placed in each of four regions. Within each region the teams are seeded from 1 to 16, with the (presumed) best team as the 1 seed and the (presumed) weakest team as the 16 seed; this practice of seeding teams began in 1979 for the NCAA tournament. Only one team from each region (so four teams each year) advances to the Final Four. This dataset if is the same as FinalFourLong, except the data starts in 1985 and we have a extra column that is an indicator for Michigan State teams coached by Tom Izzo.

Source

Final Four teams and their seed can be found at http://www.championshiphistory.com/ncaahoops.php.

`FinalFourLong` *FinalFourLong*

Description

NCAA Final Four by seed - long version

Format

A dataset with 2048 observations on the following 3 variables.

`Year`	Year 1979-2010
`Seed`	Seed in NCAA men's basketball tournament: 1 to 16
`Final4`	`1`=made Final Four or `0`=did not make Final Four

Details

Each year 64 college teams are selected for the NCAA Division I Men's Basketball tournament, with 16 teams placed in each of four regions. Within each region the teams are seeded from 1 to 16, with the (presumed) best team as the 1 seed and the (presumed) weakest team as the 16 seed; this practice of seeding teams began in 1979 for the NCAA tournament. Only one team from each region (so four teams each year) advances to the Final Four. This dataset has a row (case) for each team in the NCAA Divisoin I Men's Basketball tournament from 1979 to 2010 along with its seed and an indicator for whether the team made the Final Four that year.

Source

Final Four teams and their seed can be found at http://www.championshiphistory.com/ncaahoops.php.

`FinalFourShort` *FinalFourShort*

Description

NCAA Final Four by seed - short version

Format

A dataset with 512 observations on the following 4 variables.

`Year`	Year: 1979 to 2010
`Seed`	Seed: 1 to 16
`In`	Number of teams at that seed who made the Final Four that year
`Out`	Number of teams at that seed who did not made the Final Four that year

Details

Each year 64 college teams are selected for the NCAA Division I Men's Basketball tournament, with 16 teams placed in each of four regions. Within each region the teams are seeded from 1 to 16, with the (presumed) best team as the 1 seed and the (presumed) weakest team as the 16 seed; this practice of seeding teams began in 1979 for the NCAA tournament. Only one team from each region (so four teams each year) advances to the Final Four. This dataset is similar to FinalFourLong, except that each row combines the count of the results (make/don't make the Final Four) for each seed, so that In+Out= 4 for each row.

Source

Final Four teams and their seed can be found at http://www.championshiphistory.com/ncaahoops.php.

`Fingers` *Fingers*

Description

Fingers

Format

A dataset with 12 observations on the following 3 variables.

`Subject`	`I`, `II`, `III`, or `IV`
`Drug`	`Caffeine`, `Placebo`, or `Theobromine`
`TapRate`	Finger taps in a fixed time interval

Details

Scientists Scott and Chen, published research that compared the effects of caffeine with those of theobromine (a similar chemical found in chocolate) and with those of a placebo. Their experiment used four human subjects, and took place over several days. Each day each subject swallowed a tablet containing one of caffeine, theobromine, or the placebo. Two hours later they were timed while tapping a finger in a specified manner (that they had practiced earlier, to control for learning effects). The response is the number of taps in a fixed time interval

Source

The data was found in Statistics in Biology, Vol. 1, by C. I. Bliss (1967), New York: McGraw Hill.

References

1The original article is "Comparison of the action of 1-ethyl theobromine and caffeine in animals and man," by C. C. Scott and K. K. Chen, Journal of Pharmacological Experimental Therapy, v. 82 (1944), pp 89-97.

`FirstYearGPA` *FirstYearGPA*

Description

Predicting first-year college GPA

Format

A dataset with 219 observations on the following 10 variables.

`GPA`	First-year college GPA on a 0.0-4.0 scale
`HSGPA`	High school GPA on a 0.0-4.0 scale
`SATV`	Verbal/critical reading SAT score
`SATM`	Math SAT score
`Male`	`1`= male, `0`= female
`HU`	Number of credit hours earned in humanities courses in high school
`SS`	Number of credit hours earned in social science courses in high school
`FirstGen`	`1`= student is the first in her or his family to attend college, `0`=otherwise
`White`	`1`= white students, `0`= others
`CollegeBound`	`1`=attended a high school where >=50% students intended to go on to college, `0`=otherwise

Details

The data in FirstYearGPA contains information from a sample of 219 first year students at a mid-western college that might be used to build a model to predict their first year GPA.

Source

A sample from a larger set of data collected in 1996 by a professor at this college.

FishEggs *FishEggs*

Description

Fish Eggs

Format

A dataset with 35 observations on the following 4 variables.

Age	Age of the fish (in years)
PctDM	Percentage of the total egg material that is solid
Month	Month fish was caught: Sep=September or Nov=November
Sept	Indicator with 1=September or 0=November

Details

Researchers collected samples of female lake trout from Lake Ontario in September and November of 2002 through 2004. A goal of the study was to investigate the fertility of fish that had been stocked in the lake. One measure of the viability of fish eggs is percent dry mass(PctDM) which reflects the energy potential stored in the eggs by recording the percentage of the total egg material that is solid. Values of the PctDM for a sample of 35 lake trout (14 in September and 21 in November) are given in this dataset along with the age (in years) of the fish.

Source

Lantry, OGorman, and Machut (2008) "Maternal Characteristics versus Egg Size and Energy Density," Journal of Great Lakes Research 34(4): 661-674.

FlightResponse *FlightResponse*

Description

Flight Response of Pacific Brant

Format

A dataset with 464 observations on the following 7 variables.

FlockID	Flock ID
Altitude	Altitude of the overflight by the helicopter (in 100m)
Lateral	Lateral distance (in 100m) between the aircraft and flock
Flight	1=more than 10% of flock flies away or 0=otherwise
AltLat	Product of Altitude x Lateral
AltCat	Altitude categories: low=under 3, mid=3 to 6, high=over 6
LatCat	Lateral categories: 1under 10 to 4=over 30

Details

A 1994 study collected data on the effects of air traffic on the behavior of the Pacific Brant (a small migratory goose). The data represent the flight response to helicopter "overflights" to see what the relationship between the proximity of a flight, both lateral and altitudinal, would be to the propensity of the Brant to flee the area. For this experiment, air traffic was restricted to helicopters because previous study had ascertained that helicopters created more radical flight response than other aircraft. The data are in FlightResponse. Each case represents a flock of Brant that has been observed during one overflight in the study. Flocks were determined observationally as contiguous collections of Brants, flock sizes varying from 10 to 30,000 birds.

Source

Data come from the book Statistical Case Studies: A Collaboration Between Academe and Industry, Roxy Peck, Larry D. Haugh, and Arnold Goodman, editors; SIAM and ASA, 1998.

`Fluorescence` *Fluorescence*

Description

Data from an experiment on calcium binding proteins

Format

A dataset with 51 observations on the following 2 variables.

`Calcium`	Log of free calcium concentration
`ProteinProp`	Proportion of protein bound to calcium

Details

Suzanne Rohrback used a novel approach in a series of experiments to examine calcium binding proteins.

Source

Thanks to Suzanne Rohrback for providing these data from her honors experiments at Kenyon College.

`FruitFlies` *FruitFlies*

Description

Sexual activity and lifetimes of fruit flies

Format

A dataset with 125 observations on the following 7 variables.

ID	a numeric vector
Partners	Number of female partners: 0, 1, or 8
Type	0=pregnant, 1=virgin, 9=none
Longevity	Lifespan (in days)
Thorax	Length of thorax (in mm)
Sleep	Percent of day sleeping
Treatment	1 pregnant, 1 virgin, 8 pregnant, 8 virgin, or none

Details

Hanley and Shapiro (1994) report on a study conducted by Partridge and Farquhar (1981) about the sexual behavior of fruit flies. It was already known that increased reproduction leads to shorter life spans for female fruit flies. But the question remained whether an increase in sexual activity would also reduce the life spans of male fruit flies. The researchers designed an experiment to answer this question. They had a total of 125 male fruit flies to use and they randomly assigned each of the 125 to one of the following five groups.

Source

The data are given as part of the data archive on the Journal of Statistics Education website and can be found on the page http://www.amstat.org/publications/jse/jse_data_archive.htm.

References

Hanley and Hapiro, "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention," Journal of Statistics Education v.2, n.1, (1994), http://www.amstat.org/publications/jse/v2n1/datasets.hanley.htm

---

Goldenrod *Goldenrod Galls*

Description

Measurements for a sample of goldenrod galls

Format

A dataset with 1055 observations on the following 9 variables.

Gdiam03	Gall diameter in 2003 (in mm)
Stdiam03	Stem diameter in 2003 (in mm)
Wall03	Wall thickness in 2003 (in mm)
Fate03	b=beetle present e=early death f=living fly larva g=living wasp o=pupal case u=unknown
Gdiam04	Gall diameter in 2004 (in mm)
Stdiam04	Stem diameter in 2004 (in mm)
Wall04	Wall thickness in 2003 (in mm)
Fate04	b=beetle present e=early death f=living fly larva g=living wasp o=pupal case u=unknown
Fly04	Fly in 2004? n or y

Details

Biology students collected measurements on goldenrod galls at the Brown Family Environmental Center at Kenyon College.

Source

Thanks to the Kenyon College Department of Biology for sharing these data.

Grocery *Grocery*

Description

Grocery store sales

Format

A dataset with 36 observations on the following 5 variables.

`Discount`	Amount of discount: `5.00%`, `10.00%` or `15.00%`
`Store`	Store number (1-12)
`Display`	`Featured End of Aisl`, `Featured Middle of A`, or `Not Featured`
`Sales`	Number sold during one week
`Price`	Wholesale price (in dollars)

Details

Grocery stores and product manufacturers are always interested in how well the products on the store shelves sell. An experiment was designed to test whether the amount of discount given on products affected the amount of sales of that product. There were three levels of discount, 5%, 10%, and 15%, and sales were held for a week. The total number of products sold during the week of the sale was recorded. The researchers also recorded the wholesale price of the item put on sale.

Source

These data are not real, though they are simulated to approximate an actual study. The data come from John Grego, Director of the Stat Lab at University of South Carolina.

Gunnels *Gunnels*

Description

Presence/absence of gunnels (eels) at shoreline quadrats

Format

A dataset with 1592 observations on the following 10 variables.

Gunnel	1= gunnel present in the quadrat or 0=gunnel absent
Time	Minutes after midnight
Fromlow	Time in minutes from low tide
Slope	Slope (to nearest 10 degrees) perpendicular to waterline
Rw	Percentage cover in quadrat of rockweed/algae/plants
Amphiso	Density of crustacean food 0=none to 4=high
Subst	Substratum: 1=solid rock, 2=rocky cobbles, 3=mixed pebbles/sand, 4=fine sand, 5=mud, 6=mixed mud/shell detritus, 7=cobbles on solid rock, 8=cobbles on mixed pebbles/sand, 9=cobbles on fine sand, 10=cobbles on mud, 11=cobbles on mixed mud/shell detritus, 12=cobbles on shell detritus, 13=shell detritus
Pool	Standing water deep? 1=yes or 2=no
Water	Standing water in the quadrat? 1=yes or 2=no
Cobble	Rocky cobbles? 1=yes or 2=no

Details

This datset comes from a study on the habitat preferences of a species of eel, called a gunnel. Biologist Jake Shorty sampled quadrats along a coastline and recorded whether or not the species was found in the quadrat.

Source

Thanks to Jake Shorty, Bowdoin biology student, for this dataset.

Hawks *Hawks*

Description

Data for a sample of hawks

Format

A dataset with 908 observations on the following 19 variables.

Month	code8=September to 12=December
Day	Date in the month
Year	Year: 1992-2003
CaptureTime	Time of capture (HH:MM)
ReleaseTime	Time of release (HH:MM)
BandNumber	ID band code
Species	CH=Cooper's, RT=Red-tailed, SS=Sharp-Shinned
Age	A=Adult or I=Imature
Sex	F=Female or M=Male
Wing	Length (in mm) of primary wing feather from tip to wrist it attaches to
Weight	Body weight (in gm)
Culmen	Length (in mm) of the upper bill from the tip to where it bumps into the fleshy part of the bird
Hallux	Length (in mm) of the killing talon
Tail	Measurement (in mm) related to the length of the tail (invented at the MacBride Raptor Center)
StandardTail	Standared measurement of tail length (in mm)

Tarsus	Length of the basic foot bone (in mm)
WingPitFat	Amount of fat in the wing pit
KeelFat	Amount of fat on the breastbone (measured by feel
Crop	Amount of material in the crop, coded from 1=full to 0=empty

Details

Students and faculty at Cornell College in Mount Vernon, Iowa, collected the data over many years at the hawk blind at Lake MacBride near Iowa City, Iowa. The data set that we are analyzing here is a subset of the original data set, using only those species for which there were more than 10 observations. Data were collected on random samples of three different species of hawks: Red-tailed, Sharp-shinned, and Cooper's hawks.

Source

Many thanks to the late Professor Bob Black at Cornell College for sharing these data with us.

HawkTail *HawkTail*

Description

Tail lengths for two hawk species

Format

A dataset with 838 observations on the following 2 variables.

Species	RT=Red-tailed, SS=Sharp-shinned
Tail	Length of tail (in mm)

Details

Tail lengths measured for a sample of 838 hawks observed in Mount Vernon, Iowa. Note: HawkTail2 has similar data in unstacked format for three species of hawks.

Source

Observations by students and faculty at Cornell College.

HawkTail2 *HawkTail2*

Description

Tail lengths for three hawk species

Format

A dataset with observations on the following 3 variables.

`Tail_CH`	Tail length (in mm) for a sample of Cooper's hawks
`Tail_RT`	Tail length (in mm) for a sample of Red-tailed hawks
`Tail_SS`	Tail length (in mm) for a sample of Sharp-shinned hawks

Details

Tail lengths measured for a sample of 908 hawks observed in Mount Vernon, Iowa. Note: HawkTail has similar data in stacked format for just the Red-tailed and Sharp-shinned hawks.

Source

Observations by students and faculty at Cornell College.

`HearingTest` *HearingTest*

Description

HearingTest

Format

A dataset with 96 observations on the following 3 variables.

`Subj`	Subject number (1 - 24)
`List`	List of words: `L1` `L2` `L3` `L4`
`Percent`	Percent (out of 50) of words correctly identified

Details

Audiologists use standard lists of 50 words to test hearing; the words are calibrated, using subjects with normal hearing, to make all 50 words on the list equally hard to hear. The goal of the study described here was to see how four such lists, denoted by L1-L4 in this dataset, compared when played at low volume with a noisy background. The response is the percentage of words identified correctly.

Source

Data downloaded from DASL at http://lib.stat.cmu.edu/DASL/Datafiles/Hearing.html.

References

Loven, F. (1981), "A Study of the Interlist Equivalency of the CID W-22 Word List Presented in Quiet and in Noise." Unpublished MS Thesis, University of Iowa.

HighPeaks *HighPeaks*

Description

Adirondack High Peaks

Format

A dataset with 46 observations on the following 6 variables.

Peak	Name of the mountain
Elevation	Elevation at the highest point (in feet)
Difficulty	Rating of difficulty of the hike: 1 (easy) to 7 (most difficult)
Ascent	Vertical ascent (in feet)
Length	Length of hike (in miles)
Time	Expected trip time (in hours)

Details

Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet (although modern measurements show a couple of the peaks are actually slightly under 4000 feet). A goal for hikers in the region is to become a "46er" by scaling each of these peaks. This dataset give infomation about the hiking trails up each of these peaks.

Source

High Peaks data avaialble at http://www.adirondack.net/tour/hike/highpeaks.cfm. Thanks to Jessica Chapman at St. Lawrence University for recommending this dataset.

Hoops *Hoops*

Description

Hoops

Format

A dataset with 147 observations on the following 22 variables.

Game	An ID number assigned to each game
Opp	Name of the opponent school for the game
Home	Indicator variable where 1 = home game and 0 = away game
OppAtt	Number of field goal attempts by the opposing team
GrAtt	Number of field goal attempts by Grinnell
Gr3Att	Number of three-point field goal attempts by Grinnell
GrFT	Number of free throw attempts by Grinnell
OppFT	Number of free throw attempts by the opponent

GrRB	Total number of Grinnell rebounds
GrOR	Number of Grinnell offensive rebounds
OppDR	Number of defensive rebounds the opposing team had
OppPoint	Points scored in the game by the opponent
GrPoint	Points scored in the game by Grinnell
GrAss	Number of assists Grinnell had in the game
OppTO	Number of turnovers the opposing team gave up
GrTO	Number of turnovers Grinnell gave up
GrBlocks	Number of blocks Grinnell had in the game
GrSteal	Number of steals Grinnell had in the game
X40Point	Indicator variable that is 1 if some Grinnell player scored 40 or more points
X30Point	Indicator variable that is 1 if some Grinnell player scored 30 or more points
WinLoss	1=Grinnell win or 0=Grinnell loss
PtDiff	Point differential for the game (Grinnell score minus Opponent's score)

Details

Since 1991, David Arseneault, men's basketball coach of Grinnell College, has developed a unique, fast-paced style of basketball that he calls "the system." This dataset comes from the 147 games the Grinnell team played within its athletics conference between the 1997-98 season through the 2005-06 season.

Source

These data were collected by Grinnell College students Eric Ohrn and Ben Johannsen.

HorsePrices *HorsePrices*

Description

HorsePrices

Format

A dataset with 50 observations on the following 5 variables.

HorseID	ID code for each horse
Price	Price (in dollars)
Age	Age of the horse (in years)
Height	Height fo the horse (in hands)
Sex	f=female m=male

Details

Undergraduate students at Cal Poly collected data on prices of 50 horses advertised for sale on the internet. Predictor variables of price include the age and height of the horse (in hands), as well as its sex.

Source

Cal Poly students using a horse sale website.

`Houses` *Houses*

Description

Houses

Format

A dataset with 20 observations on the following 3 variables.

`Price`	Selling price (in dollars)
`Size`	Size of the house (in square feet)
`Lot`	Area of the house's lot (in square feet)

Details

This dataset contains selling prices for 20 houses that were sold in 2008 in a small midwestern town. The file also contains data on the size of each house (in square feet) and the size of the lot (in square feet) that the house is on.

Source

Data collected from zillow.com in June 2008.

`ICU` *ICU*

Description

Patients at an Intensive Care Unit (ICU)

Format

A dataset with 200 observations on the following 9 variables.

`ID`	Patient ID code
`Survive`	`1`=patient survived to discharge or `0`=patient died
`Age`	Age (in years)
`AgeGroup`	`1`= young (under 50), `2`= middle (50-69), `3` = old (70+)
`Sex`	`1`=female or `0`=male
`Infection`	`1`=infection suspected or `0`=no infection
`SysBP`	Systolic blood pressure (in mm of Hg)
`Pulse`	Hear rate4 (beats per minute)
`Emergency`	`1`=emergency admission or `0`=elective admission

Details

This dataset contains information for a sample of 200 patients who were part of a larger study conducted in a hospital's Intensive Care Unit (ICU). Since an ICU often deals with serious, life-threatening cases, a key variable to study is patient survival, which is coded in the Survive variable as 1 if the patient lived to be discharged and 0 if the patient died.

Source

Data downladed from The Data and Story Library (DASL), http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html.

InfantMortality *InfantMortality*

Description

Infant mortality rates in the United States (1920-2000)

Format

A dataset with 9 observations on the following 2 variables.

`Mortality`	Deaths witin one year of birth (per 1000 births
`Year`	1920-2000 (by decades)

Details

Infant mortality (deaths within one year of birth per 1,000 births) in the US from 1920 - 2000 (by decade).

Source

CDC National Vital Statistics Reports at http://www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57 14.pdf

---

InsuranceVote *InsuranceVote*

Description

Congressional votes

Format

A dataset with 435 observations on the following 9 variables.

`Party`	Party affilication: `D`=Democrat or `R`=Republican
`Dist.`	Congressional district (State-Number)
`InsVote`	Vote on the health insurance bill: `1`=yes or `0`=no
`Rep`	Indicator for Republicans
`Dem`	Indicator for Democrats

Private	Percentage of non-senior citizens in district with private health insurance
Public	Percentage of non-senior citizens in district with public health insurance
Uninsured	Percentage of non-senior citizens in district with no health insurance
Obama	District winner in 2008 presidential election: 1=Obama 0=McCain

Details

On 7 November 2009 the U.S. House of Representatives voted, by the narrow margin of 220-215, for a bill to enact health insurance reform. Most Democrats voted yes while almost all Republicans voted no. This dataset contains data for each of the 435 representatives.

Source

Insurance data are from the American Community Survey (http://www.census.gov/acs/www/data_documentation/data_m
Roll call of congressional votes on this bill can be found at http://clerk.house.gov/evs/2009/roll887.xml.

Jurors *Jurors*

Description

Jurors

Format

A dataset with 52 observations on the following 4 variables.

Period	Sequential 2-week periods ove the course of a year
PctReport	Percentage of selected jurors who report
Year	1998 or 2000
I2000	Indicator for data from the year 2000

Details

Tom Shields, jury commissioner for the Franklin County Municipal Court in Columbus, Ohio, is responsible for making sure that the judges have enough potential jurors to conduct jury trials. Jury duty for this court is two weeks long, so Tom must bring together a new group of potential jurors twenty-six times a year. Random sampling methods are used to obtain a sample of registered voters in Franklin County every two weeks, and these individuals are sent a summons to appear for jury duty. One of the most difficult aspects of Tom's job is to get those registered voters who receive a summons to actually appear at the courthouse for jury duty. This cataset contains the 1998 and 2000 data for the percentages of individuals who reported for jury duty after receiving a summons. The reporting dates vary slightly from year to year, so they are coded sequentially from 1, the first group to report in January, to 26, the last group to report in December. A variety of methods were used after 1998 to try to increase participation rates.

Source

Franklin County Municipal Court

Kids198 *Kids198*

Description

Measurements of Children

Format

A dataset with 198 observations on the following 5 variables.

`Height`	Height (in inches)
`Weight`	Weight (in pounds)
`Age`	Age (in months)
`Sex`	`0`=male or `1`=female
`Race`	`0`=white or `1`=other

Details

This dataset comes from a 1977 anthropometric study of body measurements for children. Subjects in this sample are between the ages of 8 and 18 years old, selected at random from the much larger dataset of the oringinal study.

Source

A sample of 198 cases from the NIST's AnthroKids dataset at http://ovrt.nist.gov/projects/anthrokids/

LeafHoppers *LeafHoppers*

Description

Lefetimes for potato leafhoppers on various sugar diets

Format

A dataset with 8 observations on the following 2 variables.

`Diet`	`Control`, `Fructose`, `Glucose`, or `Sucrose`
`Days`	Number of days until half the leafhoppers in a dish died

Details

The goal of this study was to compare the effects of four diets on the lifespan of small insects called potato leafhoppers. One of the four was a control diet: just distilled water with no nutritive value. Each of the other three diets had a particular sugar added to the distilled water, one of glucose, sucrose, or fructose. Leafhoppers were sorted into groups of eight and each group was put into one of eight lab dishes. Each of the four diets was added to two dishes, chosen using chance.

Source

"Survival and behavioral responses of the potato leafhopper, Empoasca Fabae (Harris), on synthetic media," MS thesis by Douglas Dahlman (1963), Iowa State University. The data can be found in Analyzing Experimental Data by Regression by David M. Allen and Foster B. Cady, Belmont, CA: Lifetime Learning (Wadsworth).

Leukemia *Leukemia*

Description

Treatment results for leukemia patients

Format

A dataset with 51 observations on the following 9 variables.

`Age`	Age at diagnosis (in years)
`Smear`	Differential percentage of blasts
`Infil`	Percentage of absolute marrow leukemia infiltrate
`Index`	Percentage labeling index of the bone marrow leukemia cells
`Blasts`	Absolute number of blasts, in thousands
`Temp`	Highest temperature of the patient prior to treatment, in degrees Farenheit
`Resp`	`1`=responded to treatment or `0`=failed to respond
`Time`	Survival time from diagnosis (in months)
`Status`	`0`=dead or `1`=alive

Details

A study involved 51 untreated adult patients with acute myeloblastic leukemia who were given a course of treatment, after which they were assessed as to their response.

Source

Data come from Statistical Analysis Using S-Plus (Brian S. Everitt; first edition 1994, Chapman & Hall).

LongJumpOlympics *LongJumpOlympics*

Description

Winning distances in men's Olympic long jump competitions (1920-2008)

Format

A dataset with 26 observations on the following 2 variables.

`Year`	Year of the Olympics (1900-2008)
`Gold`	Winning men's long jump distance (in meters)

Details

Gold medal winning distances for the men's long jump at the Olympics from 1900 to 2008.

Source

Historcial Olympic long ump results at http://trackandfield.about.com/od/longjump/qt/olymlongjumpmen.htm

---

`LostLetter` *LostLetter*

Description

Which "lost" letters will be returned by the public?

Format

A dataset with 140 observations on the following 8 variables.

`Location`	Where letter was "lost": `DesMoines`, `GrinnellCampus`, or `GrinnellTown`
`Address`	Address on teh letter: `Confederacy` or `Peaceworks`
`Returned`	`1`=letter was returned or `0`=letter was not returned
`DesMoines`	Indicator for letters left in Des Moines
`GrinnellTown`	Indicator for letters left in the town of Grinnell
`GrinellCampus`	Indicator for letters left on the Grinnell campus
`Peaceworks`	Indicator for letters addressed to Iowa Peaceworks
`Confederacy`	Indicator for letters addressed to Friends of the Confederacy

Details

In 1999 Grinnell College students Laurelin Muir and Adam Gratch conducted an experiment for an introductory statistics class. They intentionally "lost" 140 letters in either the city of Des Moines, the town of Grinnell, or on the Grinnell College campus. Half of each sample were addressed to Friends of the Confederacy and the other half to Iowa Peaceworks. The students kept track of which letters were eventually returned.

Source

Student project at Grinnell College

Marathon *Marathon*

Description

Training records for a marathon runner

Format

A dataset with 1128 observations on the following 9 variables.

Date	Training date
Miles	Miles for training run
Time	Training time (in minutes:seconds:hundredths)
Pace	Running pace (in minutes:seconds:hundredths per mile)
ShoeBrand	Addidas, Asics, Brooks, Izumi, Mizuno, or New Balance
TimeMin	Training time (in minutes)
PaceMin	Running pace (in minutes per mile)
Short	1= 5 miles or less or 0=more than 5 miles
After2004	1= for runs after 2004 or 0=for earlier runs

Details

Information from training records of a marathoner over a five-year period from 2002-2006.

Source

Data from training records of one of the Stat2 authors.

Markets *Markets*

Description

Daily changes in two stock market indices

Format

A dataset with 56 observations on the following 5 variables.

DJIAch	Change in Dow Jones Industrial Average
Date	Date: 06-Aug-09 to 02-Nov-09
Nik225ch	Change in Nikkei 225 stock average
Up	Indicator for positive Nikkei change
lagNik	Previous day's Nikkei change

Details

This dataset contains data on daily changes from two stock markets over 56 days from 06-Aug-09 to 02-Nov-09. The Dow Jones Industrial Average is based in New York and the Nikkei 225 is a stock index in Japan.

Source

Dow Jones Industrial Average: http://markets.cbsnews.com/cbsnews/quote/historical?Month=11&Symbol=310%3A998
Historical Nikkei 225 index: http://markets.cbsnews.com/cbsnews/quote/historical?Month=11&Symbol=992%3A19000

MathEnrollment *Math Enrollments*

Description

Semester enrollments in mathematics courses

Format

A dataset with 11 observations on the following 3 variables.

Ayear	Academic year (for the fall)
Fall	Fall semester total enrollments
Spring	Spring semster total enrollments

Details

Total enrollments in mathematics courses at a small liberal arts college were obtained for each semester from Fall 2001 to Spring 2012.

Source

The data were obtained from http://Registrar.Kenyon.edu on June 1, 2012.

MathPlacement *Math Placement*

Description

Results from a Math Placement exam at a liberal arts college

Format

A dataset with 2696 observations on the following 16 variables.

Student	Identification number for each student
Gender	0=Female, 1=Male
PSATM	PSAT score in MAth
SATM	SAT score in Math

ACTM	ACT Score in Math
Rank	Adjusted rank in HS class
Size	Number of students in HS class
GPAadj	Adjusted GPA
PlcmtScore	Score on math placement exam
Recommends	Recommended course: R0 R01 R1 R12 R2 R3 R4 R6 R8
Course	Actual course taken
Grade	Course grade
RecTaken	1=recommended course, 0=otherwise
TooHigh	1=took course above recommended, 0=otherwise
TooLow	1=took course below recommended, 0=otherwise
CourseSuccess	1=B or better grade, 0=grade below B

Details

Scores and course results for students taking a math placement exam at a college.

Source

Personal correspondence

MedGPA *MedGPA*

Description

Medical school admission status and information on GPA and standardized test scores

Format

A dataset with 55 observations on the following 11 variables.

Accept	Status: A=accepted to medical school or D=denied admission
Acceptance	Indicator for Accept: 1=accepted or 0=denied
Sex	F=female or M=male
BCPM	Bio/Chem/Physics/Math grade point average
GPA	College grade point average
VR	Verbal reasoning (subscore)
PS	Physical sciences (subscore)
WS	Writing sample (subcore)
BS	Biological sciences (subscore)
MCAT	Score on the MCAT exam (sum of CR+PS+WS+BS)
Apps	Number of medical schools applied to

Details

This dataset has information gathered on 55 medical school applicants from a liberal arts college in the Midwest.

Source

Data collected at a midwestern liberal arts college.

MentalHealth *Mental Health Admissions*

Description

Admissions to a mental health emergency room and full moons

Format

A dataset with 36 observations on the following 3 variables.

`Month`	Month of the year
`Moon`	Relatinship to full moon: `After`, `Before`, or `During`
`Admission`	Number of emergency room admissions

Details

Some researchers in the early 1970s set out to study whether there is a "full-moon" effect on emergency room admissions at a mental health hospital. They separated the data over 12 months into rates before the full moon (mean number of patients seen 4-13 days before the full moon), during the full moon (the number of patients seen on the full moon day), and after the full moon (mean number of patients seen 4-13 days after the full moon).

Source

Introduction to Mathematical Statistics and its Applications by Richard J. Larsen and Morris L. Marx. Prentice Hall:Englewood Cliffs, NJ, 1986.

References

The original discussion of the study is in Blackman, S., and Catalina, D. (1973). "The moon and the emergency room." Perceptual and Motor Skills 37, 624-626.

MetabolicRate *Metabolic Rate of Caterpillars*

Description

Body size and metabolic rate of Manduca Sexta caterpillars

Format

A dataset with 305 observations on the following 7 variables.

`Computer`	ID number of the computer used to measure metabolic rate

BodySize	Size of the caterpillar (in grams)
LogBodySize	Log (base 10) of BodySize
Instar	Number from 1 (smallest) to 5 (largest) indicating stage of the caterpillar's life
CO2ppm	Carbon dioxide concentration (in ppm)
Mrate	Metabolic rate
LogMrate	Log (base 10) of metablic rate

Details

Marisa Stearns collected and analyzed body size and metabolic rates for Manduca Sexta caterpillars.

Source

We thank Professor Itagaki and his research students for sharing these data.

MetroHealth83 *MetroHealth83*

Description

Health services data for 83 metropolitan areas

Format

A dataset with 83 observations on the following 16 variables.

City	Name of the metropolitan area
NumMDs	Number of physicians
RateMDs	Number of physicians per 100,000 people
NumHospitals	Nmber of community hospitals
NumBeds	Number of hospital beds
RateBeds	Number of hospital beds per 100,000 people
NumMedicare	Number of Medicare recipients in 2003
PctChangeMedicare	Percent change in Medicare recipients (2000 to 2003)
MedicareRate	Number of Medicare recipients per 100,000 people
SSBNum	Number of Social Security recipients in 2004
SSBRate	Number of Social Security recipients per 100,000 people
SSBChange	Percent change in Social Security recipients (2000 to 2004)
NumRetired	Number of retired workers
SSINum	Number of Supplemental Security Income recipients in 2004
SSIRate	Number of Supplemental Security Income recipients per 100,000 people
SqrtMDs	Square root of numver of physicians

Details

The U.S. Census Bureau regularly collects infomration for many metropolitan areas in the United States, including data on number of phsycians and number (and size) of hospitals. This dataset has such information for 83 different metropolian areas.

Source

U.S. Census Bureau: 2006 State and Metropolitan Area Data Book (Table B-6)
http://www.census.gov/prod/2006pubs/smadb/smadb-06.pdf

`Milgram` *Milgram*

Description

Attitudes towards ethics of a famous Milgram experiment

Format

A dataset with 37 observations on the following 2 variables.

`Results`	Treatment group: `Actual`, `Complied`, or `Refused`
`Score`	Ethical score from 1 (not at all ethical) to 9 (completely ethical)

Details

One of the most famous and most disturbing psychological studies of the twentieth century took place in the laboratory of Stanley Milgram at Yale University. Milgram's subjects were asked to monitor the answers of a "learner" and to push a button to deliver shocks whenever the learner gave a wrong answer. The more wrong answers, the more powerful the shock. Even Milgram himself was surprised by the results: Every one of his subjects ended up delivering what they thought was a dangerous 300-volt shock to a slow "learner" as punishment for repeated wrong answers.
Even though the "shocks" were not real and the "learner" was in on the secret, the results triggered a hot debate about ethics and experiments with human subjects. To study attitudes on this issue, Harvard graduate student Maryann de Mateo conducted a randomized comparative experiment. Her subjects were 37 high school teachers who did not know about the Milgram study. Using chance, Maryann assigned each teacher to one of three treatment groups: Group 1: Actual results. Each subject in this group read a description of Milgram's study, including the actual results that every subject delivered the highest possible "shock."
Group 2: Many complied. Each subject read the same description given to the subjects in Group 1, except that the actual results were replaced by fake results, that many but not all subjects complied.
Group 3. Most refused. For subjects in this group, the fake results said that most subjects refused to comply.
After reading the description, each subject was asked to rate the study according to how ethical they thought it was, from 1 (not at all ethical) to 9 (completely ethical.)

Source

"An experimental study of attitudes toward deception" by Mary Ann DiMatteo. Unpublished manuscript, Department of Psychology and Social Relations, Harvard University (1972).

MLB2007Standings *MLB2007Standings*

Description

Data for Major League Baseball teams from the 2007 regular season

Format

A dataset with 30 observations on the following 21 variables.

Team	Name of the team
League	League: AL or NL
Wins	Number of wins for the season (out of 162 games)
Losses	Number of losses for the season
WinPct	Proportion of games won (Wins/162)
BattingAvg	Team batting average
Runs	Number of runs runs scored
Hits	Number of hits
HR	Number of home runs hit
Doubles	Number of doubles hit
Triples	Number of triple hit
RBI	Number of runs batted in
SB	Number of stoen bases
OBP	On base percentage
SLG	Slugging percentage
ERA	Earned run average (earned runs allowed per 9 innings)
HitsAllowed	Number of hits against the team
Walks	Number of walks allowed
StrikeOuts	Number of strikeouts (by the team's pitchers)
Saves	Number of games saved (by the team's pitchers
WHIP	Number of walks and hits per inning pitched

Details

Data for all 30 Major League Baseball (MLB) teams for the 2007 regular season. This includes team batting statistics (BattingAvg through SLG) and team pitching statistics (ERA through WHIP)

Source

Data downloaded from baseball-reference.com:
http://www.baseball-reference.com/leagues/MLB/2007-standings.shtml
http://www.baseball-reference.com/leagues/MLB/2007.shtml

MothEggs *Moth Eggs*

Description

Body size and eggs produced for a species of moths

Format

A dataset with 39 observations on the following 2 variables.

BodyMass	Log of body size measured in grams
Eggs	Number of eggs present

Details

Researchers were interested an association between body size and the number of eggs produced by a species of moths.

Source

We thank Professor Itagaki and his students for sharing this data from experiments on Manduca Sexta.

NCbirths *NCbirths*

Description

Data from births in North Carolina in 2001

Format

A dataset with 1450 observations on the following 15 variables.

ID	Patient ID code
Plural	1=single birth, 2=twins, 3=triplets
Sex	Sex of the baby 1=male 2=female
MomAge	Mother's age (in years)
Weeks	Completed weeks of gestation
Marital	Marital status: 1=married or 2=not married
RaceMom	Mother's race: 1=white, 2=black, 3=American Indian, 4=Chinese 5=Japanese, 6=Hawaiian, 7=Filipino, or 8=Other Asian or Pacific Islander
HispMom	Hispanic origin of mother: C=Cuban, M=Mexican, N=not Hispanic O=Other Hispanic, P=Puerto Rico, S=Central/South America
Gained	Weight gained during pregnancy (in pounds)
Smoke	Smoker mom? 1=yes or 0=no
BirthWeightOz	Birth weight in ounces
BirthWeightGm	Birth weight in grams
Low	Indicator for low birth weight, 1=2500 grams or less
Premie	Indicator for premature birth, 1=36 weeks or sooner
MomRace	Mother's race: black, hispanic, other, or white

Details

This dataset contains data on a sample of 1450 birth records that statistician John Holcomb selected from the North Carolina State Center for Health and Environmental Statistics.

Source

Thanks to John Holcomb at Cleveland State University for sharing these data.

NFL2007Standings *NFL2007Standings*

Description

Standings for National Football League teams in 2007

Format

A dataset with 32 observations on the following 10 variables.

`Team`	Team name
`Conference`	Conference: `AFC` or `NFC`
`Division`	Division within conference: `ACE`, `ACN`, `ACS`, `ACW`, `NCE`, `NCN`, `NCS`, `NCW`
`Wins`	Number of wins (out of 16 games)
`Losses`	Number of losses
`WinPct`	Proportion of games won (Wins/16)
`PointsFor`	Total points scored by the team
`PointsAgainst`	Total points scored against the team
`NetPts`	PointsFor minus PointsAgainst
`TDs`	Number of touchdowns scored by the team

Details

Data for all 32 National Football League (NFL) teams for the 2007 regular season.

Source

1Data downloaded from www.nfl.com

Nursing *Nursing*

Description

Characteristics of nursing homes in New Mexico.

Format

A dataset with 52 observations on the following 7 variables.

Beds	Number of beds in the nursing home
InPatientDays	Annual medical in-patient days (in hundreds)
AllPatientDays	Annual total patient days (in hundreds)
PatientRevenue	Annual patinet care revenue (in hundreds of dollars)
NurseSalaries	Annual nursing salaries (in hundreds of dollars)
FacilitiesExpend	Annual facilities expenditure (in hundreds of dollars)
Rural	1=rural or 0=non-rural

Details

The data were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico in 1988.

Source

Downloaded from DASL at http://lib.stat.cmu.edu/DASL/Datafiles/Nursingdat.html

References

Howard L. Smith, Niell F. Piland, and Nancy Fisher, "A Comparison of Financial Performance, Organizational Characteristics, and Management Strategy Among Rural and Urban Nursing Facilities," Journal of Rural Health, Winter 1992, pp 27-40.

Olives *Olives*

Description

Measurements of the pesticide fenthion in olive oil over time

Format

A dataset with 18 observations on the following 7 variables.

SampleNumber	Code (1-6) for sample of olive oil
Group	Code for group: 1 or 2
Day	Time (in days) when sample was measured: 0, 281, or 365
Fenthion	Amount of fenthion (pesticide)
FenthionSulphoxide	Amount of fenthion sulfide
FenthionSulphone	Amount of fenthion sulphone
Time	Code (0, 3, or 4) for the number of days

Details

Fenthion is a pesticide used against the olive fruit fly in olive groves. It is toxic to humans so it is important that there be no residue left on the fruit or in olive oil that will be consumed. One theory was that if there is residue of the pesticide left in the olive oil, it would dissipate over time. Chemists set out to test that theory by taking a random sample of small amounts of olive oil with fenthion residue and measuring the amount of fenthion in the oil at three different times over the year - day 0, day 281 and day 365.

Source

Data provided by Rosemary Roberts and discussed in "Persistence of fenthion residues in olive oil" by Chaido Lentza-Rizos, Elizabeth J. Avramides, and Rosemary A. Roberts in Pest Management Science, Vol. 40, Issue 1, Jan. 1994, pp. 63-69.

`Orings` *Orings*

Description

Number of damaged O-rings on space shuttle launches and launch temperature

Format

A dataset with 24 observations on the following 2 variables.

`Temp`	Code for temperature (in degrees F): `Above65` `Below65`
`Failures`	Number of O-ring failures

Details

The space shuttle Challenger explodes shortly after liftoff in 1987. The subsequent investigation focused on the failure of O-ring seals, which allowed liquid hydrogen and oxygen to mix and explode. These failures might be related to temperature at the launch site which was near freezing (32 degrees F) on that day. This dataset shows the number of O-ring failures on previous shuttle launches, along with an indicator for whether the temperature was above or below 65 degrees F.

Source

Data can be found in "Risk analysis of the space shuttle: Pre-challenger prediction of failure" by Siddhartha R. Dalal, Edward B. Fowlke, and Bruce Hoadley in Journal of the American Statistical Association, Vol. 84, No. 408 (Dec. 1989), pp 945-957

`Overdrawn` *Overdrawn*

Description

Overdrawn

Format

A dataset with 450 observations on the following 4 variables.

`Age`	Age of the student (in years)
`Sex`	`0`=male or `1`=female
`DaysDrink`	Number of days drinking alcohol (in past 30 days)
`Overdrawn`	Has student overdrawn a checking account? `0`=no or `1`=yes

Details

Researchers conducted a survey of 450 undergraduates in large introductory courses at either Mississippi State University or the University of Mississippi. There were close to 150 questions on the survey, but only four of these variables are included in this dataset. (You can consult the paper to learn how the variables beyond these 4 affect the analysis.) The primary interest for the researchers was factors relating to whether or not a student has ever overdrawn a checking account.

Source

Worthy S.L., Jonkman J.N., Blinn-Pike L. (2010), "Sensation-Seeking, Risk-Taking, and Problematic Financial Behaviors of College Students," Journal of Family and Economic Issues, 31: 161-170

`PalmBeach` *PalmBeach*

Description

Votes for Geroge Bush and Pat Buchanan in Florida counties for the 2000 U.S. presidential election

Format

A dataset with 67 observations on the following 3 variables.

`County`	Name of the Florida county
`Buchanan`	Number of votes for Par Buchanan
`Bush`	number of votes for George Bush

Details

The race for the presidency of the United States in the fall of 2000 was very close, with the electoral votes from Florida determining the outcome. In the disputed final tally in Florida, George W. Bush won by just 537 votes over Al Gore, out of almost 6 million votes cast. About 2.3 were awarded to other candidates. One of those other candidates was Pat Buchanan, who did much better in Palm Beach County than he did anywhere else. Palm Beach County used a unique "butterfly ballot" that had candidate names on either side of the page with "chads" to be punched in the middle. This non-standard ballot seemed to confuse some voters, who punched votes for Buchanan that may have been intended for a different candidate. This dataset shows the number of votes for Bush and Buchanan in each Florida county.

Source

Florida county data for the 2000 presidential election can be found at http://election.dos.state.fl.us/elections/resultsarchive

---

`Pedometer` *Pedometer*

Description

Daily walking amounts recorded on a personal pedometer from September-December 2011

Format

A dataset with 68 observations on the following 8 variables.

`Steps`	Total number of steps for the day
`Moderate`	Number of steps at a moderate walking speed
`Min`	Number of minutes walking at a moderate speed
`kcal`	Number of calories burned walking at a moderate speed
`Mile`	Total number of miles walked
`Rain`	Type of weather (`rain` or `shine`)
`Day`	Day of the week (`U`=Sunday, `M`=Monday, `T`=Tuesday, `W`=Wednesday, `R`=Thursday, `F`=Friday, `S`=Saturday
`DayType`	Coded as `Weekday` or `Weekend`

Details

A statistics professor regularly keeps a pedometer in his pocket. It records not only the number of steps taken each day, but also the number of steps taken at a moderate pace, the number of minutes walked at a moderate pace, and the number of miles total that he walked. He also added to the data set the day of the week, whether it was rainy, sunny, or cold (on sunny days he often biked, but on rainy or cold days he did not), and whether it was a weekday or weekend.

Source

One of the Stat2 authors

`Perch` *Perch*

Description

Size of perch caught in a Finnish lake

Format

A dataset with 56 observations on the following 4 variables.

`Obs`	Observation number
`Weight`	Weight (in grams)
`Length`	Length (in centimeters)
`Width`	Width (in centimeters

Details

This dataset comes from a sample of fish (perch) caught at Lake Laengelmavesi in Finland.

Source

JSE Data Archive, http://www.amstat.org/publications/jse/jse_data_archive.htm, submitted by Juha Puranen.

PigFeed *PigFeed*

Description

Effects of additives to pig feed on weight gain

Format

A dataset with 12 observations on the following 3 variables.

WgtGain	Daily wight gain (hundredths of a pound over 1.00)
Antibiotic	Antibiotic in the feed? No or Yes
B12	Vitamin B12 in the feed? No or Yes

Details

A scientist in Iowa was interested in additives to standard pig chow that might increase the rate at which the pigs gained weight. Two factors of interest were vitamin B12 and antibiotics. To perform the experiment, the scientist randomly assigned 12 pigs, three to each of the diet combinations (Antibiotic only, B12 only, both, and neither).

Source

Data are found in Statistical Methods by George W. Snedecor and William G. Cochran (1967). Ames, IA: The Iowa State University Press.

References

Original source is Iowa Agricultural Experiment Station (1952). Animal Husbandry Swine Nutrition Experiment No. 577.

Pines *Pines*

Description

Data from pine seedlings planted in 1990

Format

A dataset with 1000 observations on the following 15 variables.

Row	Row number in pine plantation
Col	Column number in pine plantation
Hgt90	Tree height at time of planting (cm)
Hgt96	Tree height in September 1996 (cm)
Diam96	Tree trunk diameter in September 1996 (cm)
Grow96	Leader growth during 1996 (cm)

Hgt97	Tree height in September 1997 (cm)
Diam97	Tree trunk diameter in September 1997 (cm)
Spread97	Widest lateral spread in September 1997 (cm)
Needles97	Needle length in September 1997 (mm)
Deer95	Type of deer damage in September 1995: 0 = none, 1 = browsed
Deer97	Type of deer damage in September 1997: 0 = none, 1 = browsed
Cover95	Thorny cover in September 1995: 0 = none; 1 = some; 2 = moderate; 3 = lots
Fert	Indicator for fertilizer: 0 = no, 1 = yes
Spacing	Distance (in feet) between trees (10 or 15)

Details

This dataset contains information data from an experiment conducted by the Department of Biology at Kenyon College at a site near the campus in Gambier, Ohio. In April 1990, student and faculty volunteers planted 1000 white pine (Pinus strobes) seedlings at the Brown Family Environmental Center. These seedlings were planted in two grids, distinguished by 10- and 15-foot spacings between the seedlings. Several variables were measured and recorded for each seedling over time (in 1990, 1996, and 1997).

Source

Thanks to the Kenyon College Department of Biology for sharing these data.

Political *Political*

Description

Survey of political activity for Grinnell College students

Format

A dataset with 59 observations on the following 9 variables.

Year	Class year (1 to 4)
Sex	0=male or 1=female
Vote	Voting status: 0=not eligible, 1=eligible/not registered, 2=registered/didn't vote, 4=voted
Paper	Read news (per week): 0=never, code1=less than once, 2=once, 3=2 or 3 times, 4=daily
Edit	Read editorial page? 0=no or 1=yes
TV	Watch TV news: 0=never, code1=less than once, 2=once, 3=2 or 3 times, 4=daily
Ethics	Politics should be ruled by: 1=ethical considerations to 5=practical power
Inform	How informed are you about politics? 1=uninformed to 5=very well informed
Participate	Missing if Vote=0, 0 if Vote=1 or 2, 1 if Vote=3

Details

Students Jennifer Wolfson and Meredith Goulet conducted a survey in the spring of 1992 of Grinnell College students to ascertain patterns of political behavior. They took a simple random sample of 60 students who were U.S. citizens and conducted phone interviews. Using several "call backs"

they obtained 59 responses.

Source

Student survey at Grinnell College

`Pollster08` *Pollster08*

Description

Polls for 2008 U.S. presidential election

Format

A dataset with 102 observations on the following 11 variables.

`PollTaker`	Polling organization
`PollDates`	Dates the poll data were collected
`MidDate`	Midpoint of the polling period
`Days`	Number of days after August 28th (end of Democratic convention)
`n`	Sample size for the poll
`Pop`	`A`=all, `LV`=likely voters, `RV`=registered voters
`McCain`	Percent supporting John McCain
`Obama`	Percent supporting Barak Obama
`Margin`	Obama percent minus McCain percent
`Charlie`	Indicator for polls after Charlie Gibson interview with VP candidate Sarah Palin (9/11)
`Meltdown`	Indicator for polls after Lehman Brothers bankruptcy (9/15)

Details

The file Pollster08 contains data from 102 polls that were taken during the 2008 U.S. Presidential campaign. These data include all presidential polls reported on the internet site pollster.com that were taken between August 29th, when John Mc- Cain announced that Sarah Palin would be his running mate as the Republican nominee for vice president, and the end of September.

Source

Downloaded from pollster.com

`Popcorn` *Popcorn*

Description

Unpopped kernels in bags of microwave popcorn

Format

A dataset with 12 observations on the following 3 variables.

Unpopped	Number of unpopped kernels (adjusted for size difference
Brand	Orville or Seaway
Trial	Trial number

Details

Two students, Lara and Lisa, conducted an experiment to compare Orville Redenbacher's Light Butter Flavor vs. Seaway microwave popcorn. They made 12 batches of popcorn, 6 of each type, cooking each batch for four minutes. They noted that the microwave oven seemed to get warmer as they went along so they kept track of six trials and randomly chose which brand would go first for each trial. For a response variable they counted the number of unpopped kernels and then adjusted the count for Seaway for having more ounces per bag of popcorn (3.5 vs 3.0).

Source

Student project

PorscheJaguar *PorscheJaguar*

Description

Compare prices for Porsche and Jaguar cars offered for sale at an internet site

Format

A dataset with 60 observations on the following 5 variables.

Car	Car model: Jaguar or Porsche
Price	Price (in $1,000's)
Age	Age of the car (in years)
Mileage	Previous miles driven (in 1,000's)
Porsche	Indicator for Porsche (1) or Jaguar (0)

Details

Two students collected samples of Porsche and Jaguar cars that were offered for sale at an internet site. In addition to asking price, they recorded the model year (converting to age) and mileage of each advertised car.

Source

Student project data collected from autotrader.com in Spring 2007.

PorschePrice *PorschePrice*

Description

Prices for Porsche cars offered for sale at an internet site.

Format

A dataset with 30 observations on the following 3 variables.

`Price`	Asking price for the car (in $1,000's)
`Age`	Age of the car (in years)
`Mileage`	Previous miles driven (in 1,000's)

Details

A student was interested in pries for used Porsche sports cars being sold on the internet. He selected a random sample of 30 Porsches from the ones being advertised at autotrader.com. For each car he recorded the asking price, mileage, and model year (which he converted to age).

Source

Data collected for a student project from autotrader.com in February 2007.

`Pulse` *Pulse*

Description

Pulse rates before and after exercise for a sample of statistics students

Format

A dataset with 232 observations on the following 7 variables.

`Active`	Pulse rate (beats per minute) after exercise
`Rest`	Resting pulse rate (beats per minute)
`Smoke`	`1`=smoker or `0`=nonsmoker
`Gender`	`1`=female or `0`=male
`Exercise`	Typical hours of exercise (per week)
`Hgt`	Height (in inches)
`Wgt`	Weight (in pounds)

Details

Students in a Stat2 class recorded resting pulse rates (in class), did three "laps" walking up/down a nearby set of stairs, and then measured their pulse rate after the exercise. They provided additional information about height, weight, exercise, and smoking habits via a survey.

Source

Data compiled over several semesters from students taking a Stat2 course.

Putts1 *Putts1*

Description

Putting results for a golfing statistician

Format

A dataset with 587 observations on the following 2 variables.

`Length`	Length if the putt (in feet
`Made`	`1`=made the putt or `0`=missed the putt

Details

A statistician golfer kept careful records of every putt he attempted when playing golf, recording the length of the putt and whether or not he was successful in making the putt. This dataset has one case for each of the 587 attempted putts. A different form of the same data (Putts2) accumulates counts of makes and misses for each putt length.

Source

Personal observations by one of the Stat2 authors

Putts2 *Putts2*

Description

Putting results for a golfing statistician (by length of the putts)

Format

A dataset with 5 observations on the following 4 variables.

`Length`	Length of the attempted putt (in feet)
`Made`	Number of putts made at this length
`Missed`	Number of putts missed at this length
`Trials`	Total number of putts attempted at this length

Details

A statistician golfer kept careful records of every putt he attempted when playing golf, recording the length of the putt and whether or not he was successful in making the putt. For each different length, this dataset records the number putts made, missed, and the total number of attempts from that length. A similar dataset, Putts1, has one case for each of the 587 attempted putts, showing the length and outcome.

Source

Personal observations by one of the Stat2 authors

`ReligionGDP` *ReligionGDP*

Description

Data on religiosity of countries from the Pew Global Attitudes Project

Format

A dataset with 44 observations on the following 9 variables.

`Country`	Name of country
`Religiosity`	A measure of degree of religiosity for residents of the country
`GDP`	Per capita Gross Domestic Product in the country
`Africa`	Indicator for countries in Africa
`EastEurope`	Indicator for countries in Eastern Europe
`MiddleEast`	Indicator for countries in the Middle East
`Asia`	Indicator for countries in Asia
`WestEurope`	Indicator for countries in Western Europe
`Americas`	Indicator for countries in North/South America

Details

The Pew Research Center's Global Attitudes Project surveyed people around the world and asked (among many other questions) whether they agreed that "belief in God is necessary for morality," whether religion is very important in their lives, and whether they pray at least once per day. The variable Religiosity is the sum of the percentage of positive responses on these three items, measured in each of 44 countries. The dataset also includes the per capita GDP for each country and indicator variables that record the part of the world the country is in.

Source

Data from the 2007 Spring Survey conducted through the Pew Global Attitudes Project at http://www.pewglobal.org.

`Retirement` *Retirement*

Description

Contributions to a supplemental retirement account (1997-2012)

Format

A dataset with 16 observations on the following 2 variables.

`Year`	1997-2012
`SRA`	Annual contribution to the Supplemental Retirement Account

Details

A faculty member opened a supplemental retirement account (SRA) in 1997 to investment money for retirement. This dataset shows the annual contributions to that account. Annual contributions were adjusted downward during sabbatical years in order to maintain a steady family income.

Source

Individual records kept by the faculty member.

`RiverElements` *RiverElements*

Description

Concentrations of elements in river water samples from upstate NY

Format

A dataset with 12 observations on the following 27 variables.

`River`	One of four rivers: `Grasse`, `Oswegatchie`, `Raquette`, or `St. Regis`
`Site`	Location: 1=UpStream, 2=MidStream, 3=Downstream
`Al`	Aluminum
`Ba`	Barium
`Br`	Bromine
`Ca`	Calcium
`Ce`	Cerium
`Cu`	Copper
`Dy`	Dysprosium
`Er`	Erbim
`Fe`	Iron
`Gd`	Gadolinium
`Ho`	Holmum
`K`	Potassium
`La`	Lathanum
`Li`	Lithium
`Mg`	Magnesium
`Mn`	Manganese
`Nd`	Neodymium
`Pr`	Proseyodymium
`Rb`	Rubidium
`Si`	Silicon
`Sr`	Strontium

Y	Yttrium
Yb	Ytterbium
Zn	Zinc
Zr	Zirconium

Details

Some geologists were interested in the water chemistry of rivers in upstate New York. They took water samples at three different locations in four rivers (Grasse, Oswegatchie, Raquette, and St. Regis). The sampling sites were chosen to investigate how the composition of the water changes as it flows from the source to the mouth of each river. The sampling sites were labeled as upstream, midstream, and downstream. This dataset contains the concentations (parts per million) of a variety of elements in thoise water samples. The dataset RiverIron contains the information for iorn (FE) alone, along with the log of the concentration.

Source

Thanks to Dr. Jeff Chiarenzelli of the St. Lawrence University Geology Department for the data.

References

Chiarenzelli, Lock, Cady, Bregani and Whitney, "Variation in river multi-element chemistry related to bedrock buffering: an example from the Adirondack region of northern New York, USA", Environmental Earth Sciences, Volume 67, Number 1 (2012), 189-204

RiverIron *River Iron*

Description

Amounts of iron in water samples of four rivers

Format

A dataset with 12 observations on the following 4 variables.

River	One of four rivers: Grasse, Oswegatchie, Raquette, or St. Regis
Site	Location of the site: DownStream, MidStream or Upstream
Iron	Iron concentration in the water sample (parts per million)
LogIron	Log (base 10) of iron concentration

Details

Some geologists were interested in the water chemistry of rivers in upstate New York. They took water samples at three different locations in four rivers (Grasse, Oswegatchie, Raquette, and St. Regis). The sampling sites were chosen to investigate how the composition of the water changes as it flows from the source to the mouth of each river. The sampling sites were labeled as upstream, midstream, and downstream. This dataset contains the concentations of iron in the samples. The dataset RiverElements has similar concentration data for many other elements.

Source

Thanks to Dr. Jeff Chiarenzelli of the St. Lawrence University Geology Department for the data.

References

Chiarenzelli, Lock, Cady, Bregani and Whitney, "Variation in river multi-element chemistry related to bedrock buffering: an example from the Adirondack region of northern New York, USA", Environmental Earth Sciences, Volume 67, Number 1 (2012), 189-204

`SampleFG` *SampleFG*

Description

A sample of 30 field goal attempts in the National Football League

Format

A dataset with 30 observations on the following 13 variables.

`ID`	ID number
`PlayerID`	Code for player
`LastName`	Last name
`FirstName`	First name
`Year`	Year
`Team`	Abbreviation for team name
`Date`	Code for date: mmdddyy
`FGAttempts`	Field goals attempted by the kicker that game
`FGMade`	Field goals made by the kicker that game
`Attempt`	Which attempt during the game?
`Result`	`1`=made the field goal or `0`=missed
`Yards`	Number of yards for the field goal attempt
`Block`	`1`=attempt blocked or `0`=not blocked

Details

This is a subset of just 30 field goal attempts selected at random from the larger sample of attempts made by NFL kickers that is summarized in FGByDistance.

Source

We thank Sean Forman and Doug Drinen of Sports Reference LLC for providing us with the NFL field goal data set.

`SandwichAnts` *Sandwich Ants*

Description

Ant counts on samples of different sandwiches

Format

A dataset with 48 observations on the following 5 variables.

`Trial`	Trial number
`Bread`	Type of bread: `Multigrain`, `Rye`, `White`, or `Wholemeal`
`Filling`	Type of filling: `HamPickles`, `PeanutButter`, or `Vegemite`
`Butter`	Butter on the sandwich? `no` or `yes`
`Ants`	Number of ants on the sandwich

Details

As young students, Dominic Kelly and his friends enjoyed watching ants gather on pieces of sandwiches. Later, as a university student, Dominic decided to study this with a more formal experiment. He chose three types of sandwich fillings (vegemite, peanut butter, and ham & pickles), four types of bread (multigrain, rye, white, and wholemeal), and put butter on some of the sandwiches.
To conduct the experiment he randomly chose a sandwich, broke off a piece, and left it on the ground near an ant hill. After several minutes he placed a jar over the sandwich bit and counteed the number of ants. He repeated the process, allowing time for ants to return to the hill after each trial, until he had two samples for each combination of the three factors.

Source

Margaret Mackisack, "Favourite Experiments: An Addendum to What is the Use of Experiments Conducted by Statistics Students?", Journal of Statistics Education (1994)
http://www.amstat.org/publications/jse/v2n1/mackisack.supp.html

---

SATGPA *SAT scores and GPA*

Description

A sample of SAT scores and grade point averages for statistics students

Format

A dataset with 24 observations on the following 3 variables.

`MathSAT`	Score (out of 800) on the mathematics portion of the SAT exam
`VerbalSAT`	Score (out of 800) on the verbal portion of the SAT exam
`GPA`	Grade point average (0.0-4.0 scale)

Details

In recent years many colleges have re-examined the traditional role the scores on the Scholastic Aptitude Tests (SAT's) play in making decisions on which students to admit. Do SAT scores really

help predict success in college? To investigate this question a group of 24 introductory statistics students supplied the data in this dataset showing their score on the Verbal portion of the SAT as well as their current grade point average (GPA) on a 0.0-4.0 scale.

Source

Student survey in an introductory statistics course.

`SeaSlugs` *Sea Slugs*

Description

Metamorphose rates for sea slugs exposed to different water samples

Format

A dataset with 36 observations on the following 2 variables.

`Time`	Minutes after tide come in
`Percent`	Proportion of 15 sea slug larvae that metamorphose

Details

Sea slugs, common on the coast of southern California, live on vaucherian seaweed. The larvae from these sea slugs need to locate this type of seaweed to survive. A study was done to try to determine whether chemicals that leach out of the seaweed attract the larvae. Seawater was collected over a patch of this kind of seaweed at 5-minute intervals as the tide was coming in and, presumably, mixing with the chemicals. The idea was that as more seawater came in, the concentration of the chemicals was reduced. Each sample of water was divided into 6 parts. Fifteen larvae were then introduced to this seawater to see what percentage metamorphosed (an indication that the desired chemical was detected).

Source

Data downloaded from http://www.stat.ucla.edu/projects/datasets/seaslug-explanation.html

References

A paper based on these data: Krug, P.J. and R.K. Zimmer. 2000b. Larval settlement: chemical markers for tracing production, transport, and distribution of a waterborne cue. Marine Ecology Progress Series, vol. 207: 283-296.

`Sparrows` *Sparrows*

Description

Weight and wing length for a sample of Savannah sparrows

Format

A dataset with 116 observations on the following 3 variables.

`Treatment`	Nest adjustment: `control`, `enlarged`, or `reduced`
`Weight`	Weight (in grams)
`WingLength`	Wing length (in mm)

Details

Priscilla Erickson from Kenyon College collected data on a stratified random sample of 116 Savannah sparrows at Kent Island. Nests that were reduced, controlled (no change), or enlarged.

Source

We thank Priscilla Erickson and Professor Robert Mauck from the Department of Biology at Kenyon College for allowing us to use these data.

`SpeciesArea` *Species Area*

Description

Land area and number of mammal species for island in Southeast Asia

Format

A dataset with 14 observations on the following 5 variables.

`Name`	Name of the island
`Area`	Area (in sq. km)
`Species`	Number of mammal species
`logArea`	Natural logarithm (base e) of Area
`logSpecies`	Natural logarithm (base e) of Species

Details

This dataset shows the number of mammal species and the area for 13 islands in Southeast Asia. Biologists have speculated that the number of species is related to the size of an island and would like to be able to predict the number of species given the size of an island.

Source

Heaney, Lawrence R. (1984) Mammalian species richness on islands on the Sunda Shelf, Southeast Asia, Oecologia.

Speed *Speed*

Description

Highway fatality rates 1987-2007

Format

A dataset with 21 observations on the following 3 variables.

`Year`	Year (1987-2007
`FatalityRate`	Number of fatalites on interstate highways (per 100 million vehicle-miles)
`StateControl`	`0`=1987-1994 or `1`=1995-2007

Details

In 1987 the federal government allowed the speed limit on interstate highways to be 65 mph in most areas. In 1995 federal restrictions were eliminated, so that states assumed control of setting speed limits on interstate highways. This data set compares fatality rates for years before and after the states assumed control for highway speed limits.

Source

Data from the National Highway Saftey Administration website at http://www-fars.nhtsa.dot.gov/Main/index.aspx

Swahili *Swahili*

Description

Attitudes towards the Swahili language among Kenyan school children

Format

A dataset with 480 observations on the following 4 variables.

`Province`	`NAIROBI` or `PWANI`
`Sex`	`female` or `male`
`Attitude.Score`	Score (out a possible 200 points) on a survey of attitude towards the Swahili language
`School`	Code for the school: `A` through `L`

Details

Hamisi Babusa, a Kenyan scholar, administered a survey to 480 students from Pwani and Nairobi provinces about their attitudes towards the Swahili language. In addition, the students took an exam on Swahili. From each province, the students were from 6 schools (3 girls schools and 3 boys schools) with 40 students sampled at each school, so half of the students from each province were

males and the other half females. The survey instrument contained 40 statements about attitudes towards Swahili and students rated their level of agreement to each. Of these questions, 30 were positive questions and the remaining 10 were negative questions. On an individual question the most positive response would be assigned a value of 5 while the most negative response would be assigned a value of 1. By summing (adding) the responses to each question, we can find an overall Attitude Score for each student. The highest possible score would be 200 (an individual who gave the most positive possible response to every question). The lowest possible score would be 40 (an individual who gave the most negative response to every question).

Source

Thanks to Dr. Babusi of Kenyatta University for sharing these data.

TextPrices *Text Prices*

Description

Prices and number of pages for a sample of college textbooks

Format

A dataset with 30 observations on the following 2 variables.

`Pages`	Number of pages in the textbook
`Price`	Price of the textbook (in dollars)

Details

Two undergraduate students at Cal Poly - San Luis Obispo took a random sample of 30 textbooks from the campus bookstore in the fall of 2006. They recorded the price and number of pages in each book, in order to investigate the question of whether number of pages can be used to predict price.

Source

Student project

ThreeCars *Three Cars*

Description

Compare prices for Porsche, Jaguar, and BMW cars offered for sale at an internet site

Format

A dataset with 90 observations on the following 8 variables.

CarType	BMW, Jaguar, or Porsche
Price	Asking price (in $1,000's)
Age	Age of the car (in years)
Mileage	previous miles driven (in 1,000's)
Car	0=Porsche, 1=Jaguar, 2=BMW
Porsche	Indicator with 1= Porsche and 0=otherwise
Jaguar	Indicator with 1= Jaguar and 0=otherwise
BMW	Indicator with 1= BMW and 0=otherwise

Details

Two students collected samples of Porsche, Jaguar, and BMW cars that were offered for sale at an internet site. In addition to asking price, they recorded the model year (converting to age) and mileage of each advertised car. The PorschePrice dataset has only the Porsche data and the Porsche-Jaguar dataset has the data for those two models.

Source

Student project data collected from autotrader.com in Spring 2007.

TipJoke *Tip Joke*

Description

Effect of a waiter leaving a joke or an advertisement on getting a tip

Format

A dataset with 211 observations on the following 5 variables.

Card	Type of card used: Ad, Joke, or None
Tip	1=customer left a tip or 0=no tip
Ad	Indicator for Ad card
Joke	Indicator for Joke card
None	Indicator for no card

Details

Can telling a joke affect whether or not a waiter in a coffee bar receives a tip from a customer? A study investigated this question at a coffee bar at a famous resort on the west coast of France. The waiter randomly assigned coffee-ordering customers to one of three groups: When receiving the bill one group also received a card telling a joke, another group received a card containing an advertisement for a local restaurant, and a third group received no card at all. He recorded whether or not each customer left a tip.

Source

Gueguen, Nicholas (2002), "The Effects of a Joke on Tipping When it is Delivered at the Same Time as the Bill," Journal of Applied Social Psychology, 32, 1955-1963.

`Titanic` *Titanic*

Description

List and outcomes for passengers on the Titanic

Format

A dataset with 1313 observations on the following 6 variables.

`Name`	Passenger name
`PClass`	Passenger class: `*`=missing, `1st`, `2nd`, or `3rd`
`Age`	Age (in years)
`Sex`	`female` or `male`
`Survived`	`1`=survived or `0`=died
`SexCode`	`1`=female or `0`=male

Details

The Titanic was a British luxury ocean liner that sank famously in the icy North Atlantic on its maiden voyage in April of 1912. Of the approximately 2200 passengers on board, 1500 died. The high death rate was blamed largely on the inadequate supply of lifeboats, a result of the manufacturer's claim that the ship was "unsinkable." A partial data set of the passenger list was compiled by Philip Hinde in his Encyclopedia Titanica and is given in this dataset.

Source

Philip Hinde's Encyclopedia Titanica, http://www.encyclopedia-titanica.org/. Data may also be downloaded from the Australasian Data and Story Library (OzDASL) at http://www.statsci.org/data/general/titanic.html.

`TMS` *TMS*

Description

Effects of transcranial magnetic stimulation (TMS) on migraine headaches

Format

A dataset with 2 observations on the following 4 variables.

`Group`	Treatment group: `Placebo` or `TMS`

Yes	Count of number of patients that were pain-free
No	Count of number of patients that had pain
Trials	Number of patients in the group

Details

A study investigated whether a handheld device that sends a magnetic pulse into a person's head might be an effective treatment for migraine headaches. Researchers recruited 200 subjects who suffered from migraines and randomly assigned them to receive either the TMS (transcranial magnetic stimulation) treatment or a sham (placebo) treatment from a device that did not deliver any stimulation. Subjects were instructed to apply the device at the onset of migraine symptoms and then assess how they felt two hours later. This dataset is a two-way table of the results.

Source

Based on results in R. B. Lipton, et. al. (2010) "Single-pulse Transcranial Magnetic Stimulation for Acute Treatment of Migraine with Aura: A Randomised, Double-blind, Parallel-group, Sham-controlled Trial," 9(4):373-380.

TomlinsonRush *LaDainian Tomlinson Rushing Yards*

Description

Rushing yards for each game LaDainian Tomlinson played in the 2006 National Football League (NFL regular) season.

Format

A dataset with 16 observations on the following 4 variables.

Game	Week number in the 2006 season
Opponent	Name of opposing team
Attempts	Number of rushing attempts
Yards	Total yards gamed rushing for the game

Details

For each of the sixteen games the San Diego Chargers played in the 2006 NFL regular season we have the number of times LaDainian Tomlinson ran the ball and the total yards he gained.

Source

Data downloaded from http://www.pro-football-reference.com/players/T/TomlLa00/gamelog/2006/

TwinsLungs *TwinsLungs*

Description

Comparing lung function for twins between rural and urban environments

Format

A dataset with 14 observations on the following 3 variables.

`Pair`	Code for the twin pair: `A` - `G`
`Environ`	Living environment: `Rural` or `Urban`
`Percent`	Percentage of radioactivity remaining in lungs

Details

This dataset is from a study to compare the effect of living environment (rural or urban) on human lung function, where the researchers were able to locate seven pairs of twins with one twin in each pair living in the country, the other in a city. To measure lung function, twins inhaled an aerosol of radioactive Teflon particles. By measuring the level of radioactivity immediately and then again after an hour, the scientists could measure the rate of "tracheobronchial clearance." The percentage of radioactivity remaining in the lungs after an hour told how quickly subjects' lungs cleared the inhaled particles.

Source

"Urban factor and tracheobronchial clearance" by Per Camner and Klas Philipson in Archives of Environmental Health, V. 27 (1973), page 82. Data can be found in Introduction to Mathematical Statistics and its Applications, 2nd Edition by Richard J. Larson and Morris L. Marx. Englewood Cliffs, NJ: Prentice Hall, p. 548.

USstamps *USstamps*

Description

Price of US stamp for first class mail 1885-2012

Format

A dataset with 25 observations on the following 2 variables.

`Year`	Years when stamp price changed
`Price`	Cost of a US first class stamp (in cents)

Details

The data record the year and price for each change in price for a US first class (1 ounce, domestic letter) stamp since 1885.

Source

http://about.usps.com/who-we-are/postal-history/domestic-letter-rates-1863-2011.htm

---

Volts *Volts*

Description

Voltage drop over time as a capacitor discharges

Format

A dataset with 50 observations on the following 2 variables.

`Voltage`	Voltage (in volts)
`Time`	Time after charging (in seconds)

Details

A capacitor was charged with a nine-volt battery and then a voltmeter recorded the voltage as the capacitor was discharged. Measurements were taken every .02 seconds.

Source

Measurements recorded by one of the authors.

WalkingBabies *WalkingBabies*

Description

An experiment to see if special exercises help babies learn to walk sooner

Format

A dataset with 24 observations on the following 2 variables.

`Group`	Treatments: `exercise control`, `final report`, `special exercises`, or `weekly report`
`Age`	Age (in months) when first walking

Details

Scientists wondered if they could get babies to walk sooner by prescribing a set of special exercises. Their experimental design included four groups of babies and the following treatments:
Special exercises: Parents were shown the special exercises and encouraged to use them with their children. They were phoned weekly to check on their child's progress.
Exercise control: These parents were not shown the special exercises, but they were told to make sure their babies spent at least 15 minutes a day exercising.
Weekly report: Parents in this group were not given instructions about exercise. Like the parents in the treatment group, however, they received a phone call each week to check on progress.
Final report: These parents were not given weekly phone calls or instructions about exercises. They reported at the end of the study.

Source

Zelazo, Phillip R., Nancy Ann Zelazo, and Sarah Kolb (1972), "Walking in the Newborn," Science, v. 176, pp. 314-315.

`WeightLossIncentive` *WeightLossIncentive*

Description

An experiment to see if financial incentives improve weight loss

Format

A dataset with 38 observations on the following 3 variables.

`WeightLoss`	Weight loss (in pounds) after four months
`Group`	Treatment group: `Control` or `Incentive`
`Month7Loss`	Weight loss (in pounds) after seven months

Details

Researchers investigated whether financial incentives would help people lose weight more successfully. Some participants in the study were randomly assigned to a treatment group that was offered financial incentives for achieving weight loss goals, while others were assigned to a control group that did not use financial incentives. All participants were monitored over a four month period and the net weight change (Before - After in pounds) at the end of this period was recorded for each individual. Then the individuals were left alone for three months with a followup weight check at the seven-month mark to see whether weight losses persisted after the original four months of treatment.
The 4-month data alone (with missing values omitted) is stored in WeightLossIncentive4.
The 7-month data alone (with missing values omitted) is stored in WeightLossIncentive7.

Source

"Financial incentive-based approaches for weight loss," Journal of the American Medical Association by Volpp, John, Troxel, et. al., Vol. 200, no. 22, pp 2631-2637, (Dec. 2008)

WeightLossIncentive4 *WeightLossIncentive4*

Description

Weight loss after four months with/without a financial incentive

Format

A dataset with 36 observations on the following 2 variables.

`WeightLoss`	weight loss (in pounds) after 4 months
`Group`	Treatment group: `Control` or `Incentive`

Details

Researchers investigated whether financial incentives would help people lose weight more successfully. Some participants in the study were randomly assigned to a treatment group that was offered financial incentives for achieving weight loss goals, while others were assigned to a control group that did not use financial incentives. All participants were monitored over a four month period and the net weight change (Before - After in pounds) at the end of this period was recorded for each individual. Then the individuals were left alone for three months with a followup weight check at the seven-month mark to see whether weight losses persisted after the original four months of treatment. This dataset has only the non-missing 4-month data. The 7-month data are in WeightLossIncentive7 and both measurements (including missing values) are in WeightLossIncentive.

Source

"Financial incentive-based approaches for weight loss," Journal of the American Medical Association by Volpp, John, Troxel, et. al., Vol. 200, no. 22, pp 2631-2637, (Dec. 2008)

WeightLossIncentive7 *WeightLossIncentive7*

Description

Weight loss after seven months with/without a financial incentive

Format

A dataset with 33 observations on the following 2 variables.

`Group`	Treatment group: `Control` or `Incentive`
`Month7Loss`	Weight loss (in pounds) after seven months

Details

Researchers investigated whether financial incentives would help people lose weight more successfully. Some participants in the study were randomly assigned to a treatment group that was offered financial incentives for achieving weight loss goals, while others were assigned to a control group that did not use financial incentives. All participants were monitored over a four month period and the net weight change (Before - After in pounds) at the end of this period was recorded for each individual. Then the individuals were left alone for three months with a followup weight check at the seven-month mark to see whether weight losses persisted after the original four months of treatment. This dataset has only the non-missing 7-month data. The 4-month data are in WeightLossIncentive4 and both measurements (including missing values) are in WeightLossIncentive.

Source

"Financial incentive-based approaches for weight loss," Journal of the American Medical Association by Volpp, John, Troxel, et. al., Vol. 200, no. 22, pp 2631-2637, (Dec. 2008)

`WordMemory` *WordMemory*

Description

Percentage of different types of words recalled

Format

A dataset with 40 observations on the following 4 variables.

`Subject`	Code to identify each subject: `A` to `J`
`Abstract`	Words were abstract? `No` or `Yes`
`Frequent`	Words were common? `No` or `Yes`
`Percent`	Percentage of words recalled (out of 25)

Details

One hundred words were presented to each subject in a randomized order. The goal of the experiment was to see whether some kinds of words were easier to remember than others. In particular, are common words like potato, love, diet, and magazine easier to remember than less common words like manatee, hangnail, fillip, and apostasy? Are concrete words like coffee, dog, kale, and tamborine easier than abstract words like beauty, sympathy, fauna, and guile? There were 25 words each of four kinds, obtained by crossing the two factors of interest, Abstraction (concrete or abstract) and Frequency (common or rare).

Source

Data from a student laboratory project, Department of Psychology and Education, Mount Holyoke College.

YouthRisk2007 *YouthRisk2007*

Description

Risky behavior (riding with a drnk driver) in youths

Format

A dataset with 13387 observations on the following 6 variables.

ride.alc.driver	1=rode with a drinking driver in past 30 days or 0=did not
female	1=female or 0=male
grade	Year in high school: 9, 10, 11, or 12
age4	Age (in years)
smoke	Ever smoked? 1=yes or 0=no
DriverLicense	Have a driver's license? 1=yes or 0=no

Details

This dataset is derived from the 2007 Youth Risk Behavior Surveillance System (YRBSS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC) to monitor the prevalence of health-risk youth behaviors. This datset focuses on whether or not youths have recently (in past 30 days) ridden with a drunk driver.

Source

The article "Which Young People Accept a Lift From a Drunk or Drugged Driver?" in Accident Analysis and Prevention (July 2009. pp. 703-9) provides more details.

References

A more recent version of the full dataset is available at http://www.cdc.gov/brfss/technical_infodata/surveydata.htm.

YouthRisk2009 *YouthRisk2009*

Description

Survey of students in grades 9-12 concerning health-risk behiviors

Format

A dataset with 500 observations on the following 6 variables.

Sleep	Average hours sleep on school night, coded with 10 (
]tab 8 hours, 7 hours, 6 hours, 5 hours, or 4 or less hours	
Sleep7	Seven or more hours of sleep? 0=No, 1=Yes
SmokeLife	Ever smoked? No or Yes

SmokeDaily	Regular smoker? No or Yes
MarijuaEver	Ever smoked marijuana? 0=No or 1=Yes
Age	Age (in years)

Details

Data from the Centers for Disease Control's Youth Risk Behavior Surveillance System (YRBSS).

Source

http://www.cdc.gov/HealthyYouth/yrbs/index.htm

Index