Package ‘Stat2Data’

October 12, 2022

Type  Package
Title  Datasets for Stat2
Version  2.0.0
Date  2018-12-29
Author  Ann Cannon, George Cobb, Bradley Hartlaub, Julie Legler, Robin Lock, Thomas Moore, Allan Rossman, Jeffrey Witmer
Maintainer  Robin Lock <rlock@stlawu.edu>
License  GPL-3
LazyLoad  yes
Depends  R (>= 2.10)
Suggests  stats,graphics
URL  https://github.com/statmanrobin/Stat2Data
BugReports  https://github.com/statmanrobin/Stat2Data
Encoding  UTF-8
RoxygenNote  6.1.0
NeedsCompilation  no
Repository  CRAN
Date/Publication  2019-01-04 11:00:07 UTC

R topics documented:

  Stat2Data-package ............................................. 6
  AccordPrice ................................................... 7
  AHCAvote2017 ..................................................... 7
  Airlines .......................................................... 8
  Alfalfa ........................................................... 8
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlitoConfirmation</td>
<td>9</td>
</tr>
<tr>
<td>Amyloid</td>
<td>10</td>
</tr>
<tr>
<td>AppleStock</td>
<td>10</td>
</tr>
<tr>
<td>ArcheryData</td>
<td>11</td>
</tr>
<tr>
<td>AthleteGrad</td>
<td>12</td>
</tr>
<tr>
<td>AudioVisual</td>
<td>12</td>
</tr>
<tr>
<td>AutoPollution</td>
<td>13</td>
</tr>
<tr>
<td>Backpack</td>
<td>14</td>
</tr>
<tr>
<td>BaseballTimes</td>
<td>14</td>
</tr>
<tr>
<td>BaseballTimes2017</td>
<td>15</td>
</tr>
<tr>
<td>BeeStings</td>
<td>16</td>
</tr>
<tr>
<td>BirdCalcium</td>
<td>16</td>
</tr>
<tr>
<td>BirdNest</td>
<td>17</td>
</tr>
<tr>
<td>Blood1</td>
<td>18</td>
</tr>
<tr>
<td>BlueJays</td>
<td>18</td>
</tr>
<tr>
<td>BrainpH</td>
<td>19</td>
</tr>
<tr>
<td>BreesPass</td>
<td>20</td>
</tr>
<tr>
<td>BritishUnions</td>
<td>20</td>
</tr>
<tr>
<td>ButterfliesBc</td>
<td>21</td>
</tr>
<tr>
<td>CAFE</td>
<td>21</td>
</tr>
<tr>
<td>CalciumBP</td>
<td>22</td>
</tr>
<tr>
<td>CanadianDrugs</td>
<td>23</td>
</tr>
<tr>
<td>CancerSurvival</td>
<td>24</td>
</tr>
<tr>
<td>Caterpillars</td>
<td>24</td>
</tr>
<tr>
<td>CavsShooting</td>
<td>25</td>
</tr>
<tr>
<td>Cereal</td>
<td>26</td>
</tr>
<tr>
<td>ChemoTHC</td>
<td>26</td>
</tr>
<tr>
<td>ChildSpeaks</td>
<td>27</td>
</tr>
<tr>
<td>ClintonSanders</td>
<td>28</td>
</tr>
<tr>
<td>Clothing</td>
<td>28</td>
</tr>
<tr>
<td>CloudSeeding</td>
<td>29</td>
</tr>
<tr>
<td>CloudSeeding2</td>
<td>30</td>
</tr>
<tr>
<td>CO2</td>
<td>31</td>
</tr>
<tr>
<td>CO2Germany</td>
<td>32</td>
</tr>
<tr>
<td>CO2Hawaii</td>
<td>32</td>
</tr>
<tr>
<td>CO2SouthPole</td>
<td>33</td>
</tr>
<tr>
<td>Contraceptives</td>
<td>33</td>
</tr>
<tr>
<td>cooksplot</td>
<td>34</td>
</tr>
<tr>
<td>CountyHealth</td>
<td>35</td>
</tr>
<tr>
<td>CrabShip</td>
<td>36</td>
</tr>
<tr>
<td>CrackerFiber</td>
<td>36</td>
</tr>
<tr>
<td>CreditRisk</td>
<td>37</td>
</tr>
<tr>
<td>Cuckoo</td>
<td>37</td>
</tr>
<tr>
<td>Day1Survey</td>
<td>38</td>
</tr>
<tr>
<td>DiabeticDogs</td>
<td>39</td>
</tr>
<tr>
<td>Diamonds</td>
<td>40</td>
</tr>
<tr>
<td>Diamonds2</td>
<td>40</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>41</td>
</tr>
</tbody>
</table>
R topics documented:

Election08 ................................................................. 41
Election16 ................................................................. 42
ElephantsFB ................................................................. 43
ElephantsMF ................................................................. 44
emplogitplot1 .............................................................. 44
emplogitplot2 .............................................................. 46
Ethanol ................................................................. 48
Eyes ................................................................. 48
Faces ................................................................. 49
FaithfulFaces ................................................................. 50
FantasyBaseball ................................................................. 50
FatRats ................................................................. 51
Fertility ................................................................. 52
FGByDistance ................................................................. 52
Film ................................................................. 53
FinalFourIzzo ................................................................. 54
FinalFourIzzo17 ................................................................. 54
FinalFourLong ................................................................. 55
FinalFourLong17 ................................................................. 56
FinalFourShort ................................................................. 57
FinalFourShort17 ................................................................. 57
Fingers ................................................................. 58
FirstYearGPA ................................................................. 59
FishEggs ................................................................. 59
Fitch ................................................................. 60
FlightResponse ................................................................. 61
FloridaDP ................................................................. 61
Fluorescence ................................................................. 62
FranticFingers ................................................................. 63
FruitFlies ................................................................. 63
FruitFlies2 ................................................................. 64
FunnelDrop ................................................................. 65
GlowWorms ................................................................. 66
Goldenrod ................................................................. 66
GrinnellHouses ................................................................. 67
Grocery ................................................................. 68
Gunnels ................................................................. 69
Handwriting ................................................................. 69
Hawks ................................................................. 70
HawkTail ................................................................. 71
HawkTail2 ................................................................. 72
HearingTest ................................................................. 72
HeatingOil ................................................................. 73
HighPeaks ................................................................. 74
Hoops ................................................................. 74
HorsePrices ................................................................. 75
Houses ................................................................. 76
HousesNY ................................................................. 76
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU</td>
<td>77</td>
</tr>
<tr>
<td>InfantMortality2010</td>
<td>78</td>
</tr>
<tr>
<td>Inflation</td>
<td>78</td>
</tr>
<tr>
<td>Insurance Vote</td>
<td>79</td>
</tr>
<tr>
<td>IQGuessing</td>
<td>80</td>
</tr>
<tr>
<td>Jurors</td>
<td>80</td>
</tr>
<tr>
<td>Kershaw</td>
<td>81</td>
</tr>
<tr>
<td>KeyWestWater</td>
<td>82</td>
</tr>
<tr>
<td>Kids198</td>
<td>83</td>
</tr>
<tr>
<td>Leafhoppers</td>
<td>83</td>
</tr>
<tr>
<td>LeafWidth</td>
<td>84</td>
</tr>
<tr>
<td>Leukemia</td>
<td>85</td>
</tr>
<tr>
<td>LeveeFailures</td>
<td>85</td>
</tr>
<tr>
<td>LewyBody2Groups</td>
<td>86</td>
</tr>
<tr>
<td>LewyDLBad</td>
<td>87</td>
</tr>
<tr>
<td>LongJumpOlympics</td>
<td>87</td>
</tr>
<tr>
<td>LongJumpOlympics2016</td>
<td>88</td>
</tr>
<tr>
<td>LosingSleep</td>
<td>88</td>
</tr>
<tr>
<td>LostLetter</td>
<td>89</td>
</tr>
<tr>
<td>Marathon</td>
<td>90</td>
</tr>
<tr>
<td>Markets</td>
<td>90</td>
</tr>
<tr>
<td>MathEnrollment</td>
<td>91</td>
</tr>
<tr>
<td>MathPlacement</td>
<td>91</td>
</tr>
<tr>
<td>MedGPA</td>
<td>92</td>
</tr>
<tr>
<td>Meniscus</td>
<td>93</td>
</tr>
<tr>
<td>MentalHealth</td>
<td>94</td>
</tr>
<tr>
<td>MetabolicRate</td>
<td>94</td>
</tr>
<tr>
<td>MetroCommutes</td>
<td>95</td>
</tr>
<tr>
<td>MetroHealth83</td>
<td>96</td>
</tr>
<tr>
<td>Migraines</td>
<td>96</td>
</tr>
<tr>
<td>Milgram</td>
<td>97</td>
</tr>
<tr>
<td>MLB2007Standings</td>
<td>98</td>
</tr>
<tr>
<td>MLBStandings2016</td>
<td>99</td>
</tr>
<tr>
<td>MothEggs</td>
<td>100</td>
</tr>
<tr>
<td>MouseBrain</td>
<td>101</td>
</tr>
<tr>
<td>MusicTime</td>
<td>102</td>
</tr>
<tr>
<td>NCBirths</td>
<td>102</td>
</tr>
<tr>
<td>NFL2007Standings</td>
<td>103</td>
</tr>
<tr>
<td>NFLStandings2016</td>
<td>104</td>
</tr>
<tr>
<td>Nursing</td>
<td>105</td>
</tr>
<tr>
<td>OilDeapsorption</td>
<td>105</td>
</tr>
<tr>
<td>Olives</td>
<td>106</td>
</tr>
<tr>
<td>Orings</td>
<td>107</td>
</tr>
<tr>
<td>Overdrawn</td>
<td>108</td>
</tr>
<tr>
<td>Oysters</td>
<td>108</td>
</tr>
<tr>
<td>PalmBeach</td>
<td>109</td>
</tr>
<tr>
<td>PeaceBridge2003</td>
<td>110</td>
</tr>
<tr>
<td>PeaceBridge2012</td>
<td>110</td>
</tr>
<tr>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------</td>
</tr>
<tr>
<td>Pedometer</td>
<td>111</td>
</tr>
<tr>
<td>Perch</td>
<td>111</td>
</tr>
<tr>
<td>PigFeed</td>
<td>112</td>
</tr>
<tr>
<td>Pines</td>
<td>113</td>
</tr>
<tr>
<td>PKU</td>
<td>114</td>
</tr>
<tr>
<td>Political</td>
<td>114</td>
</tr>
<tr>
<td>Pollster08</td>
<td>115</td>
</tr>
<tr>
<td>Popcorn</td>
<td>116</td>
</tr>
<tr>
<td>PorscheJaguar</td>
<td>116</td>
</tr>
<tr>
<td>PorschePrice</td>
<td>117</td>
</tr>
<tr>
<td>Pulse</td>
<td>118</td>
</tr>
<tr>
<td>Putts1</td>
<td>118</td>
</tr>
<tr>
<td>Putts2</td>
<td>119</td>
</tr>
<tr>
<td>Putts3</td>
<td>119</td>
</tr>
<tr>
<td>RacialAnimus</td>
<td>120</td>
</tr>
<tr>
<td>RadioactiveTwins</td>
<td>121</td>
</tr>
<tr>
<td>RailsTrails</td>
<td>122</td>
</tr>
<tr>
<td>Rectangles</td>
<td>123</td>
</tr>
<tr>
<td>ReligionGDP</td>
<td>124</td>
</tr>
<tr>
<td>RepeatedPulse</td>
<td>124</td>
</tr>
<tr>
<td>ResidualOil</td>
<td>125</td>
</tr>
<tr>
<td>Retirement</td>
<td>126</td>
</tr>
<tr>
<td>Ricci</td>
<td>126</td>
</tr>
<tr>
<td>RiverElements</td>
<td>127</td>
</tr>
<tr>
<td>RiverIron</td>
<td>128</td>
</tr>
<tr>
<td>SampleFG</td>
<td>129</td>
</tr>
<tr>
<td>SandwichAnts</td>
<td>130</td>
</tr>
<tr>
<td>SATGPA</td>
<td>131</td>
</tr>
<tr>
<td>Sealce</td>
<td>131</td>
</tr>
<tr>
<td>SeaSlugs</td>
<td>132</td>
</tr>
<tr>
<td>SleepingShrews</td>
<td>133</td>
</tr>
<tr>
<td>sluacf</td>
<td>134</td>
</tr>
<tr>
<td>Sparrows</td>
<td>135</td>
</tr>
<tr>
<td>SpeciesArea</td>
<td>135</td>
</tr>
<tr>
<td>Speed</td>
<td>136</td>
</tr>
<tr>
<td>SugarEthanol</td>
<td>136</td>
</tr>
<tr>
<td>SuicideChina</td>
<td>137</td>
</tr>
<tr>
<td>Swahili</td>
<td>138</td>
</tr>
<tr>
<td>Tadpoles</td>
<td>139</td>
</tr>
<tr>
<td>TechStocks</td>
<td>139</td>
</tr>
<tr>
<td>TeenPregnancy</td>
<td>140</td>
</tr>
<tr>
<td>TextPrices</td>
<td>140</td>
</tr>
<tr>
<td>ThomasConfirmation</td>
<td>141</td>
</tr>
<tr>
<td>ThreeCars</td>
<td>142</td>
</tr>
<tr>
<td>ThreeCars2017</td>
<td>142</td>
</tr>
<tr>
<td>TipJoke</td>
<td>143</td>
</tr>
<tr>
<td>Titanic</td>
<td>144</td>
</tr>
<tr>
<td>TMS</td>
<td>144</td>
</tr>
</tbody>
</table>
This package included datasets for both the first and second editions of the text.

Author(s)

Ann Cannon, George Cobb, Bradley Hartlaub, Julie Legler, Robin Lock, Thomas Moore, Allan Rossman, Jeffrey Witmer
AccordPrice

Maintainer: Robin Lock <rlock@stlawu.edu>

---

**AccordPrice**

*Prices of Used Honda Accords (in 2017)*

**Description**

Age, price, and mileage of used Honda Accords in 2017

**Format**

A data frame with 30 observations on the following 3 variables.

- **Age**: Age of used Honda Accord car
- **Price**: Price (in $1,000’s)
- **Mileage**: Mileage (in 1,000’s of miles)

**Details**

Information on used Honda Accords obtained from cars.com.

**Source**

Cars.com, February 2017 using zip code 44107, Lakewood, Ohio

---

**AHCAvote2017**

*Congressional Votes on American Health Care Act (in 2017)*

**Description**

Congressional votes on the American Health Care Act in 2017

**Format**

A data frame with 430 observations on the following 11 variables.

- **STATE**: State name
- **Dist**: Congressional district
- **Party**: Party affiliation (D=Democrat, R=Republican)
  - Dem 1=Democrat, 0=Republican
  - Rep 1=Republican, 0=Democrat
- **uni2013**: Percentage of citizens without health care in 2013
- **uni2015**: Percentage of citizens without health care in 2015
- **uniChange**: uni2015 - uni2013
- **Member**: Name of representative
- **AHCAvote**: 1=yes, 0=no
  - Trump 1=Trump won district, 0=Clinton won district
Details
On May 4, 2017, the U.S. House of Representatives voted, by the narrow margin of 217-213, to pass the American Health Care Act. Most Republicans voted Yes, while all Democrats voted No.

Source
https://fivethirtyeight.com/features/obamacare-has-increased-insurance-coverage-everywhere/
https://docs.google.com/spreadsheets/d/1VfkHtzBTP5gf4jAu8tcVQgsBJ1IDvXEHjuMqY1OgYbA/edit#gid=0

---

**Airlines**

**Ontime Records for Two Airlines at Two Airports**

**Description**
Ontime arrivals for American and Delta airlines at LaGuardia and O'Hare airports

**Format**
A data frame with 10333 observations on the following 5 variables.

- **airline**: American or Delta
- **airport**: LGA=LaGuardia ORD=O'Hare
- **OnTime**: no or yes
- **IndORD**: Is the airport ORD? (1=yes or 0=no)
- **IndDelta**: Is the airline Delta? (1=yes or 0=no)

**Details**
Ontime/late data for individual flights to LaGuardia and O'Hare airports by American and Delta airlines.

**Source**
Data collected on 9/20/16 from http://www.transtats.bts.gov/ot_delay/OT_DelayCause1.asp?pn=1

---

**Alfalfa**

**Alfalfa Growth**

**Description**
Growth of alfalfa sprouts in acidic conditions

**Format**
A dataset with 15 observations on the following 3 variables.
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 (windows) 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


Examples

data(Alfalfa)
Source

These numbers are taken from Kastellec, J.P., Lax, J.R., and Phillips, J. (2010), "Public Opinion and Senate Confirmation of Supreme Court Nominees," Journal of Politics, 72(3): 767-84. In this paper the authors used opinion polls and an advanced statistical method known as multilevel regression and poststratification to determine the StateOpinion levels.

<table>
<thead>
<tr>
<th>Amyloid</th>
<th>Amyloid-beta and Cognitive Impairment</th>
</tr>
</thead>
</table>

Description

Amyloid-beta and cognitive impairment for a sample of Catholic priests

Format

A data frame with 57 observations on the following 2 variables.

Group mAD=Alzheimer's, MCI=mild impairment, NCI=no impairment

Abeta Amount of Abeta from the posterior cingulate cortex (pmol/g tissue)

Details

Amyloid-beta (Abeta) is a protein fragment that has been linked to Alzheimer’s disease. Autopsies from a sample of Catholic priests included measurements of Abeta (pmol/g tissue from the posterior cingulate cortex) from three groups: subjects who had exhibited no cognitive impairment before death, subjects who had exhibited mild cognitive impairment, and subjects who had mild to moderate Alzheimer’s disease.

Source

Violetta N. Pivtoraiko, Eric E. Abrahamson, Sue E. Leurgans, Steven T. DeKosky, Elliott J. Mufson, Milos D. Ikonomovic (2015) Cortical pyroglutamate amyloid-beta levels and cognitive decline in Alzheimer’s disease. Neurobiology of Aging (36) 12-19. Data are read from Figure 1, panel d.

<table>
<thead>
<tr>
<th>AppleStock</th>
<th>Daily Price and Volume of Apple Stock</th>
</tr>
</thead>
</table>

Description

Daily prices and trading volume of Apple stock from July 21st to August 21st in 2016
ArcheryData

Format
A data frame with 66 observations on the following 4 variables.

Date  Date as mm/dd/yyyy
Price  Closing price of Apple stock
Change  Change in price from previous day
Volume  Number of shares traded (in millions)

Details
Closing price of Apple stock (AAPL) for each trading day in a three month period from 7/21/2016 to 10/21/2016 as well as the change in stock price and number of shares traded.

Source
Data downloaded from Nasdaq historical prices at http://www.nasdaq.com/symbol/aapl/historical

Scores in an Archery Class

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
**AudioVisual**

---

**AthleteGrad**

*Athletic Participation, Race, and Graduation*

**Description**

Six-year graduation data for 214,555 students in 2004

**Format**

A data frame with 214555 observations on the following 3 variables.

- Student Athlete or NonAthlete
- Race Black or White
- Grad 1=graduated within 6 years, otherwise 0

**Details**

Six-year graduation data from 2004 for male non-athletes and for male athletes, where "Athlete" means football or basketball player. These data show Simpson’s Paradox.

**Source**

Victor Matheson, College of the Holy Cross, collected the summary statistics.

Data are derived from the summary tables in:


---

**AudioVisual**

*Reaction Times to Audio and Visual Stimuli*

**Description**

Data from an experiment on reaction times to audio or visual stimuli by Oberlin College students.

**Format**

A data frame with 72 observations on the following 4 variables.

- Subject SubjectIDs coded s1 to s36
- ResponseTime Time to respond to a stimulus (in ms)
- Stimulus Type of stimulus (auditory or visual)
- Group Musician or NonMusician
Details

Subjects in a reaction time study were asked to press a button as fast as possible after being exposed to either an auditory stimulus (a burst of white noise) or a visual stimulus (a circle flashing on a computer screen). Average reaction times (ms) were recorded for between 10 and 20 trials for each type of stimulus for each subject. Data also identifies which subjects are musicians.

Source

Arjuna Pettit, Jr. and Jeremy Potterfield at Oberlin College

Description

Measurements of noise levels for different filters to reduce pollution levels of automobiles.

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 filter or 2=new filter
- **Side**: Side of vehicle: code 1=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](http://lib.stat.cmu.edu/DASL/Stories/airpollutionfilters.html).

References

**Description**

Backpack weights for a sample of college students

**Format**

A data frame 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**


---

**Description**

Game times and boxscore information for baseball games

**Format**

A data frame with 15 observations on the following 7 variables.

- **Game**: Code for opposing teams
- **League**: AL= American League or NL=National League
## Details

Data were collected for 15 Major League Baseball (MLB) games played on August 26, 2008. This dataset was used in first edition, but updated to BaseballTimes2017 for the second edition.

### Source

Data from boxscores at www.baseball-reference.com

---

### Description

Times for one day’s major league baseball games

### Format

A data frame with 14 observations on the following 7 variables.

<table>
<thead>
<tr>
<th>Game</th>
<th>League</th>
<th>Runs</th>
<th>Margin</th>
<th>Pitchers</th>
<th>Attendance</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLB teams that played</td>
<td>AL=American League, IL=Interleague, or NL=National League</td>
<td>Runs scored by the two teams combined</td>
<td>Margin of victory (Winner-Loser score)</td>
<td>Total number of pitchers used (both teams)</td>
<td>Number of spectators at the game</td>
<td>Total time for the game (in minutes)</td>
</tr>
</tbody>
</table>

### Details

Data from all MLB games played on August 11, 2017. There were no extra-innings game nor any rain delays.

### Source

https://www.baseball-reference.com/boxes/?month=8&day=11&year=2017
**BeeStings**  
*Do Bee Stings Depend on Previous Stings?*

**Description**
Data from an experiment to see if the number of bee stings depends on previous stings.

**Format**
A data frame with 18 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>Occasion</th>
<th>Trial: I to IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>Fresh or Stung</td>
</tr>
<tr>
<td>Stingers</td>
<td>Number of stingers</td>
</tr>
</tbody>
</table>

**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 trials.


**Source**

---

**BirdCalcium**  
*Effect of a Hormone on Bird Calcium Levels*

**Description**
An experiment on the effects of a hormone on blood calcium levels in robins.

**Format**
A data frame with 20 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Bird</th>
<th>ID number for each bird (1 to 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>female or male</td>
</tr>
<tr>
<td>Hormone</td>
<td>Treated with hormone (no or yes)</td>
</tr>
<tr>
<td>Group</td>
<td>Combined Sex and Hormone (F No, F Yes, M No, or M Yes)</td>
</tr>
<tr>
<td>Ca</td>
<td>Blood calcium level (mg per 100 ml)</td>
</tr>
</tbody>
</table>
Details

An experiment looked at the effects of treatment with a hormone for increasing the concentration of calcium in birds. Twenty birds (robins) were used in the study, ten male and ten female, equally divided between the hormone and no hormone treatments.

Source

Bliss, Chester (1970), Statistics in Biology, McGraw-Hill

---

BirdNest

**Nest Characteristics for Different Bird Species**

**Description**

Nest and species characteristics for North American passerines

**Format**

A data frame with 84 observations on the following 12 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Latin species name</td>
</tr>
<tr>
<td>Common</td>
<td>Common species name</td>
</tr>
<tr>
<td>Page</td>
<td>Page in a bird manual describing the species</td>
</tr>
<tr>
<td>Length</td>
<td>Mean body length for the species (in cm)</td>
</tr>
<tr>
<td>Nesttype</td>
<td>Type of nest</td>
</tr>
<tr>
<td>Location</td>
<td>Location of nest</td>
</tr>
<tr>
<td>No.eggs</td>
<td>Number of eggs</td>
</tr>
<tr>
<td>Color</td>
<td>Egg color (0=plain/solid or 1=speckled/spotted)</td>
</tr>
<tr>
<td>Incubate</td>
<td>Mean length of time (in days) the species incubates eggs in the nest</td>
</tr>
<tr>
<td>Nestling</td>
<td>Mean length of time (in days) the species cares for babies in the nest until fledged</td>
</tr>
<tr>
<td>Totcare</td>
<td>Total care time = Incubate+Nestling</td>
</tr>
<tr>
<td>Closed</td>
<td>1=closed nest (pendant, spherical, cavity, crevice, burrow), 0=open nest (saucer, cup)</td>
</tr>
</tbody>
</table>

**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 84 separate species of North American passerines.

**Source**

Project by Amy Moore at Grinnell College

**References**

Blood1

**Blood Pressure, Weight, and Smoking Status**

**Description**

Systolic blood pressure, weight and smoking status for a sample of 500 adults

**Format**

A data frame with 500 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SystolicBP</td>
<td>Systolic blood pressure (mm of Hg)</td>
</tr>
<tr>
<td>Smoke</td>
<td>Y=smoker or N=non-smoker</td>
</tr>
<tr>
<td>Overwt</td>
<td>1=normal, 2=overweight, or 3=obese</td>
</tr>
</tbody>
</table>

**Details**

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

**Source**


BlueJays

**Blue Jay Measurements**

**Description**

Body measurements for a sample of blue jays

**Format**

A data frame with 123 observations on the following 9 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BirdID</td>
<td>ID tag for bird</td>
</tr>
<tr>
<td>KnownSex</td>
<td>Sex coded as F or M</td>
</tr>
<tr>
<td>BillDepth</td>
<td>Thickness of the bill measured at the nostril (in mm)</td>
</tr>
<tr>
<td>BillWidth</td>
<td>Width of the bill (in mm)</td>
</tr>
<tr>
<td>BillLength</td>
<td>Length of the bill (in mm)</td>
</tr>
<tr>
<td>Head</td>
<td>Distance from tip of bill to back of head (in mm)</td>
</tr>
<tr>
<td>Mass</td>
<td>Body mass (in grams)</td>
</tr>
<tr>
<td>Skull</td>
<td>Distance from base of bill to back of skull (in mm)</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex coded as 0=female or 1=male</td>
</tr>
</tbody>
</table>
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

BrainpH

Brain pH Measurements

---

Description

Brain tissue pH at time of death

Format

A data frame with 54 observations on the following 5 variables.

- pH  Brain tissue pH
- Sex  F or M
- Ethnicity  AfricanAmerican, Asian, Caucasian, or PacificIslander
- Age  Age at death
- DeathType  Cause of death (Cardiac, Other, or Suicide)

Details

These are data from a PNAS article (supplemental file) on pH in brain tissue samples for controls and for people who had Major Depressive Disorder. We extracted just the controls (roughly 3/4 of whom died of cardiac arrest).

Source


Data extracted from Supporting Information, Table S4: Li et al. www.pnas.org/cgi/content/short/1305814110
**BreesPass**

*Drew Brees Passing Statistics (2016)*

**Description**

Passing statistics for football quarterback Drew Brees in 2016

**Format**

A data frame with 16 observations on the following 5 variables.

- **Game**  Game number (1 is the first game of the regular season)
- **Opponent**  Opponent abbreviation
- **Completed**  Number of completed passes
- **Attempts**  Pass attempts
- **Yards**  Passing yards

**Details**

Drew Brees was the quarterback for the NFL’s New Orleans Saints football team in 2016. This dataset shows some of his passing statistics for each of the 16 regular season games.

**Source**


---

**BritishUnions**

*Attitudes Towards British Trade Unions*

**Description**

Poll attitudes towards British trade unions

**Format**

A data frame 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

B. chariclea Measurements

Description
Measurements for a sample of butterflies in Greenland

Format
A data frame with 32 observations on the following 4 variables.
- Temp: Average temperature for preceding summer (Celsius)
- Wing: Average wing length (mm)
- Sex: Female or Male
- Species: all are Bc, Boloria chariclea

Details
Scientists measured wing length of a species of butterfly, Boloria chariclea (Bc), in Greenland each year from 1996 through 2013. They also recorded summer temperatures.

Source
Digitized data from plots in Bowden, J. et al., "High-Arctic butterflies become smaller with rising temperatures", published in Biology Letters 11: 20150574

CAFE
US Senate Votes on Corporate Average Fuel Economy Bill

Description
Senate votes for Corporate Average Fuel Economy (CAFE) bill

Format
A data frame with 100 observations on the following 7 variables.
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 from SUNY Oswego for the data.

CalciumBP

Do Calcium Supplements Lower Blood Pressure?

Description

An experiment on calcium supplements and blood pressure in 21 men

Format

A data frame with 21 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium or Placebo</td>
<td>Beginning-ending blood pressure</td>
</tr>
</tbody>
</table>

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

---

CanadianDrugs | Canadian Drugs Senate Vote

Description

US Senate vote on Klobuchar amendment to lower drug prices

Format

A data frame with 94 observations on the following 6 variables.

- Senator: Name of the Senator
- Contributions: Amount of money received from the pharmaceutical industry over 6 years
- Party: D=Democrat or R=Republican
- State: Abbreviation for Senator’s state
- RollCall: Nay or Yea
- Vote: Against or With what drug makers wanted

Details

January 2017 vote in the U.S. Senate related to repeal part of ObamaCare. The "Klobuchar amendment" to a bill was introduced with the purpose of lowering drug prices by allowing prescription drugs to be imported from Canada.

The data exclude two senators who did not vote on the amendment and four senators who were new to Congress and thus had received no money from the drug industry. The remaining 94 senators represent 49 states (every state except California) and each of these senators had received at least $3,000.

Source

Data obtained from:

- http://maplight.org/us-congress/interest/H4300/view/all
CancerSurvival  

Survival Times for Different Cancers

Description
Cancer survival with ascorbate supplement

Format
A dataset with 64 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Survival</th>
<th>Survival time (in days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organ</td>
<td>Breast, Bronchus, Colon, Ovary, or Stomach</td>
</tr>
</tbody>
</table>

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

Caterpillars  

Measurements of Manduca Sexta Caterpillars

Description
Measurements on a sample of Manduca Sexta caterpillars

Format
A data frame 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.

---

**CavsShooting**  
**Cleveland Cavalier’s Shooting (2016-2017)**

**Description**

Shooting percentages for two Cav players

**Format**

A data frame with 1940 observations on the following 3 variables.

- **Player**: Frye or Irving
- **ShotType**: Two or Three
- **Hit**: 1=made or 0=missed

**Details**

Shooting success on 2-point shots and 3-point shots for the 2016-17 NBA season for two Cleveland Cavalier basketball players, Kyrie Irving and Channing Frye. Each case is a shot attempt. These data show Simpson’s Paradox.
ChemoTHC

Source


Cereal Nutrition Content of Breakfast Cereals

Description

Nutrition content for a sample of 36 different brands of breakfast cereals

Format

A data frame with 36 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Cereal</th>
<th>Brandname of cereal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calories</td>
<td>Calories per serving</td>
</tr>
<tr>
<td>Sugar</td>
<td>Grams of sugar per serving</td>
</tr>
<tr>
<td>Fiber</td>
<td>Grams of fiber per serving</td>
</tr>
</tbody>
</table>

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 THC for Antinausea Treatment in Chemotherapy

Description

Comparison of two treatments for nausea in chemotherapy

Format

A data frame with 2 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prochlorperazine or THC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective</td>
<td>Count of effective cases</td>
</tr>
</tbody>
</table>
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 was 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 2 x 2 table of counts.

Source

Description
Age at first speaking and aptitude test scores

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

<table>
<thead>
<tr>
<th>Child</th>
<th>ID for each child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age at first speaking (in months)</td>
</tr>
<tr>
<td>Gesell</td>
<td>Gesell Aptitude Test Score</td>
</tr>
</tbody>
</table>

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)
ClintonSanders  Clinton/Sanders Primary Results (2016)

Description
2016 US Democratic Presidential primary results

Format
A data frame with 31 observations on the following 5 variables.

- State: ID for primary state
- Delegates: Percentage of delegates won by Clinton
- PaperTrail: Was a paper trail available for votes cast? (No Paper Trail or Paper Trail)
- PopularVote: Percentage of votes won by Clinton
- AfAmPercent: Percentage of African-Americans in the state

Details
In 2016 Hillary Clinton won the Democratic nomination for U.S. President over Bernie Sanders. A paper was circulated that claimed to show evidence of election fraud based, among other things, on Clinton doing better in states that don’t have a paper trail for votes cast in a primary election than she did in states that have a paper trail. Data is for the 31 states that held Democratic primaries in 2016.

Source
https://docs.google.com/spreadsheets/d/1cszGOhbmHDTHH5ntaGPmeX55RgMMaoBhqO1Wx-9TRk/edit#gid=0
http://kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22%22%7D

Clothing  Sales for a Clothing Retailer

Description
Data on 60 customers at a clothing retailer
CloudSeeding

Format

A data frame with 60 observations on the following 8 variables.

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

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 with David Cameron who completed a more extensive consulting project for the retailer.

CloudSeeding

Cloud Seeding Experiment (Winter Only)

Description

Rainfall amounts from a cloud seeding experiment (winter only)

Format

A data frame with 28 observations on the following 7 variables.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeded</td>
<td>Treatment coded as S=seeded or U=unseeded</td>
</tr>
<tr>
<td>Season</td>
<td>All in Winter</td>
</tr>
<tr>
<td>TE</td>
<td>Rainfall in East (treatment)</td>
</tr>
<tr>
<td>TW</td>
<td>Rainfall in West (treatment)</td>
</tr>
<tr>
<td>NC</td>
<td>Rainfall in North (control)</td>
</tr>
<tr>
<td>SC</td>
<td>Rainfall in South (control)</td>
</tr>
<tr>
<td>NWC</td>
<td>Rainfall in Northwest (control)</td>
</tr>
</tbody>
</table>
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


CloudSeeding2

Cloud Seeding Experiment (Four Seasons)

Description

Rainfall amounts from a cloud seeding experiment

Format

A data frame with 108 observations on the following 8 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>ID for time period</td>
</tr>
<tr>
<td>Seeded</td>
<td>Treatment coded as S=seeded or U=unseeded</td>
</tr>
<tr>
<td>Season</td>
<td>Coded as Autumn, Spring, Summer, or Winter</td>
</tr>
<tr>
<td>TE</td>
<td>Rainfall in East (treatment)</td>
</tr>
<tr>
<td>TW</td>
<td>Rainfall in West (treatment)</td>
</tr>
<tr>
<td>NC</td>
<td>Rainfall in North (control)</td>
</tr>
<tr>
<td>SC</td>
<td>Rainfall in South (control)</td>
</tr>
<tr>
<td>NWC</td>
<td>Rainfall in Northwest (control)</td>
</tr>
</tbody>
</table>

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


---

**Daily CO2 Measurements in Germany**

**Description**

Daily carbon dioxide measurements for April through November 2011

**Format**

A data frame with 237 observations on the following 2 variables.

| CO2 | Carbon dioxide (CO2) level (in parts per million) |
| Day | Number of day in 2011 (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 2011.

This dataset was renamed to CO2Germany for the second edition.

**Source**

http://gaw.empa.ch/gawsis/reports.asp?StationID=-739519191
**Daily CO2 Measurements in Germany**

**Description**

Daily carbon dioxide measurements for April through November 2011

**Format**

A data frame with 237 observations on the following 2 variables.

- **CO2**: Carbon dioxide (CO2) level (in parts per million)
- **Day**: Number of day in 2011 (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 2011.

**Source**

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

---

**CO2 Readings in Hawaii**

**Description**

Monthly carbon dioxide readings at Mauna Loa, Hawaii

**Format**

A data frame with 360 observations on the following 4 variables.

- **Year**: Year (1988 - 2017)
- **Month**: Month (1=Jan. to 12=Dec.)
- **CO2**: Atmospheric carbon dioxide level (ppm)
- **t**: Time interval (t=1 to 360)

**Details**


**Source**

Data downloaded for MOL (Mauna Loa) from the ESRL/GMD data page at https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html
**CO2Readings at the South Pole**

**Description**

Monthly carbon dioxide readings at the South Pole

**Format**

A data frame with 348 observations on the following 4 variables.

- **Year**: Year (1988 - 2016)
- **Month**: Month (1=Jan. to 12=Dec.)
- **CO2**: Atmospheric carbon dioxide level (ppm)
- **t**: Time interval (t=1 to 348)

**Details**


**Source**

Data downloaded for SPO (South Pole) from the ESRL/GMD data page at https://www.esrl.noaa.gov/gmd/.

---

**Contraceptives**

**Drug Interaction with Contraceptives**

**Description**

Drug interaction study with oral contraceptives

**Format**

A data frame with 44 observations on the following 6 variables.

- **ID**: ID number for each of the women
- **StudyPeriod**: 1=first or 2=second
- **Treatment**: Drug or Placebo
- **EE**: Bioavailability of the ethinyl estradiol component of the oral contraceptive (in pg/hr/ml)
- **ComparisonValues**: Comparison values used for a Tukey nonadditivity plot
- **Residuals**: Residuals used for a Tukey nonadditivity plot
Details

Twenty-two female subjects were allocated randomly to one of two treatment sequences in a two period crossover design. The two treatments were a new Drug D or placebo, both given concomitantly with a standard oral contraceptive which was given in both study periods. The oral contraceptive has two components, ethinyl estradiol (EE) and norethindrone (NET). The purpose of the study was to evaluate whether the presence of Drug D affected the bioavailability of each of the oral contraceptive components. Note that our dataset does not include the NET variable.

Source


**cooksplot**

*Plot of standardized residuals vs. leverage with boundaries for unusual cases*

**Description**

This function produces a plot of standardized residuals versus leverage values for a regression model. Horizontal boundaries identify mild or more extreme standardized residuals. Vertical boundaries identify mild and more severe high leverage points. Curved boundaries identify mild and more severe values of Cook’s D.

**Usage**

```r
cooksplot(mod)
```

**Arguments**

- `mod` a regression model from `lm()`

**Details**

The plot shows standardized residuals (vertical) versus leverage values (horizontal) for all cases in a regression model.

Horizontal (blue) boundaries mark standardized residuals beyond +/- 2 (mild) and +/- 3 (more severe).

Vertical (green) boundaries mark leverage points beyond 2(k+1)/4 (mild) and 3(k+1)/n (more severe), where k= number of predictors.

Curved (red) boundaries for mark influential points beyond 0.5 (mild) and 1.0 (more severe) using Cook’s D.

Unusual points are labeled with a case number.

**Value**

A plot showing standardized residuals versus leverage values with boundaries for unusual cases
Examples

```r
data(AccordPrice)
mod1=lm(Price~Age,data=AccordPrice)
cooksplot(mod1)
```

---

**CountyHealth**  
**County Health Resources**

---

**Description**

Medical facilities and doctors in a sample of counties.

**Format**

A data frame with 53 observations on the following 4 variables.

- **County**: County name, state
- **MDs**: Number of medical doctors
- **Hospitals**: Number of community hospitals
- **Beds**: Number of beds in the hospitals

**Details**

Data compiled from information provided by the American Medical Association on the availability of health care in counties in the United States. A random sample of 53 counties was chosen from among counties with at least two community hospitals.

**Source**


Other web sources:  
http://www.ama-assn.org/  
http://www.cms.hhs.gov  
http://www.ssa.gov
CrabShip

Description
Oxygen intake of crabs with different noise sources

Format
A data frame with 34 observations on the following 3 variables.
Mass Oxygen intake of crabs with different noise sources
Oxygen Rate of oxygen consumption (μ moles h⁻¹)
Noise Source of noise (ambient or ship)

Details
Animals that are stressed might increase their oxygen consumption. Biologists measured oxygen consumption of shore crabs that were either exposed to 7.5 minutes of ship noise or 7.5 minutes of ambient harbor noise.

Source

CrackerFiber

Description
Digested calories with different types of fiber in crackers

Format
A data frame 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.

CreditRisk 

<table>
<thead>
<tr>
<th>Overdrawn Checking Account?</th>
</tr>
</thead>
</table>

Description

Variables that might be related to whether students overdraw a checking account.

Format

A data frame 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


Cuckoo 

Measurements of Cuckoo Eggs
Description

Lengths of cuckoo eggs laid in other birds’ nests

Format

A data frame 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 known 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 various other bird species.

Source

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

References


Day1Survey

First Day Survey of Statistics Students

Description

Data from a first day class survey in an introductory statistics course

Format

A data frame 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 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.

Lactic Acid Turnover in Dogs

Description
The rate of lactic acid turnover was measured by two methods for normal and diabetic dogs.

Format
A data frame with 20 observations on the following 4 variables.
Dog     Code for individual dogs (d1 through d10)
Method  Tracer method to measure response (infuse or inject)
Operation Pancreas removed to make the dog diabetic? (no or yes)
Response Rate for biochemical turnover of lactic acid

Details
Five dogs had their pancreas removed to make them diabetic (Operation=yes), the other five were normal (Operation=no). The rate of turnover of lactic acid was measured for each dog by two methods, infusion and injection.

Source
Description
Price and characteristics for a sample of 351 diamonds

Format
A data frame with 351 observations on the following 6 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carat</td>
<td>Size of the diamond (in carats)</td>
</tr>
<tr>
<td>Color</td>
<td>Coded as D (most white/bright) through J</td>
</tr>
<tr>
<td>Clarity</td>
<td>Coded as IF, VVS1, VVS2, VS1, VS2, SI1, SI2, or SI3</td>
</tr>
<tr>
<td>Depth</td>
<td>Depth (as a percentage of diameter)</td>
</tr>
<tr>
<td>PricePerCt</td>
<td>Price per carat</td>
</tr>
<tr>
<td>TotalPrice</td>
<td>Price for the diamond (in dollars)</td>
</tr>
</tbody>
</table>

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.

Description
A subset of 307 cases with the most frequent colors from the Diamonds data

Format
A data frame with 307 observations on the following 6 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carat</td>
<td>Size of the diamond (in carats)</td>
</tr>
<tr>
<td>Color</td>
<td>Coded as D (most white/bright) through G</td>
</tr>
<tr>
<td>Clarity</td>
<td>Coded as IF, VVS1, VVS2, VS1, VS2, SI1, SI2, or SI3</td>
</tr>
<tr>
<td>Depth</td>
<td>Depth (as a percentage of diameter)</td>
</tr>
<tr>
<td>PricePerCt</td>
<td>Price per carat</td>
</tr>
<tr>
<td>TotalPrice</td>
<td>Price for the diamond (in dollars)</td>
</tr>
</tbody>
</table>
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.

---

Dinosaurs  Iridium Levels in Rock Layers to Investigate Dinosaur Extinction

Description

Iridium levels in prehistoric rock layers

Format

A data frame with 28 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Type of rock (Limestone Shale)</td>
</tr>
<tr>
<td>Depth</td>
<td>Depth of the sample (in meters)</td>
</tr>
<tr>
<td>Iridium</td>
<td>Iridium concentration (ppb)</td>
</tr>
</tbody>
</table>

Details

The question of interest is whether a volcanic eruption or asteroid strike had created a dust cloud that led to extinction of most dinosaurs. Rock samples taken in Gubbio, Italy were measured for the concentration of iridium (a rare metal which is more common in asteroids). The deeper the sample, the older the rocks are. A sudden increase in iridium at some point in time would lend support for the asteroid hypothesis.

Source


---

Election08  2008 U.S. Presidential Election

Description

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

Format

A dataframe with 51 observations on the following 7 variables.
Election16

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 and %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)
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

Election16 2016 U.S. Presidential Election

Description
2016 presidential election and state demographic data

Format
A data frame with 50 observations on the following 8 variables.
State State name
Abr Abbreviation for state name
Income Per capita income in the state
HS Percent high school grads
BA Percent college grads
Adv Percent with advanced degrees
Dem.Rep Democratic lean - Republican lean in 2015 Gallup poll
TrumpWin Trump won the state? (1=yes or 0=no)

Details
This dataset contains information from all 50 states and the District of Columbia for the 2016 U.S. presidential election. It is similar to Election08 for the 2008 election.
**Source**

Income data from https://www.census.gov/search-results.html?q=per+capita+income+by+state&search.x=0&search.y=0&search=submit&page=1

2015 data via American Community Survey


---

**ElephantsFB**

*Measurements of Male African Elephants*

---

**Description**

Age and height of male African elephants

**Format**

A data frame with 138 observations on the following 3 variables.

- **Age** Age (in years)
- **Height** Shoulder height (in cm)
- **Firstborn** Firstborn? (1=yes, 0=no)

**Details**

Data on 138 male African elephants that lived through droughts in the first two years of life.

**Source**

Data are from Phyllis Lee, Stirling University, and are related to Lee, P., et al. (2013), "Enduring consequences of early experiences: 40-year effects on survival and success among African elephants (Loxodonta Africana)," Biology Letters, 9: 20130011.
Measurements of African Elephants

Description
Age and height of African elephants

Format
A data frame with 288 observations on the following 3 variables.

- **Age**: Age (in years)
- **Height**: Shoulder height (in cm)
- **Sex**: F=female or M=male

Details
Data on 288 African elephants that lived through droughts in the first two years of life.

Source
Data are from Phyllis Lee, Stirling University, and are related to Lee, P., et al. (2013), "Enduring consequences of early experiences: 40-year effects on survival and success among African elephants (Loxodonta Africana)," Biology Letters, 9: 20130011.

Empirical logit plot for one quantitative variable

Description
This function produces an empirical logit plot for a binary response variable and a single quantitative predictor variable.

Usage
```r
emplogitplot1(formula, data = NULL, ngroups = 3, breaks = NULL, yes = NULL, padj = TRUE, out = FALSE, showplot = TRUE, showline = TRUE, ylab = "Log(Odds)", xlab = NULL, dotcol = "black", linecol = "blue", pch = 16, main = "", ylim = NULL, xlim = NULL, lty = 1, lwd = 1, cex = 1)
```
Arguments

- **formula**: A formula of the form (binary) Response~Predictor
- **data**: A dataframe
- **ngroups**: Number of groups to use (not needed if breaks is used), ngroups="all" uses all unique values
- **breaks**: A vector of endpoints for the bins (not needed if ngroups is used)
- **yes**: Set a value for the response to be counted for proportions (optional)
- **padj**: Should proportions be adjusted to avoid zero and one? (default is TRUE)
- **out**: Should the function return a dataframe with group information? (default is FALSE)
- **showplot**: Show the plot? default is TRUE
- **showline**: Show the regression line? default is TRUE
- **ylab**: Text label for the vertical axis (default is "Log(Odds)")
- **xlab**: Text label for the horizontal axis (default is NULL)
- **dotcol**: Color for the dots (default is "black")
- **linecol**: Color for the line (default is "black")
- **pch**: Plot character for the dots (default is 16)
- **main**: Title for plot
- **ylim**: Limits for the vertical axis
- **xlim**: Limits for the horizontal axis
- **lty**: Line type (default is 1)
- **lwd**: Line width (default is 1)
- **cex**: Multiplier for plot symbols

Details

Values of the quantitative explanatory variable will be grouped into ngroups roughly equal sized groups, unless breaks is used to determine the boundaries of the groups. Using ngroups="all" will make each distinct value of the explanatory variable its own group.

We find an adjusted proportion for the binary response variable within each of the groups with (Number yes +0.5)/(Number of cases+1). This is converted to an adjusted log odds log(adjp/(1-adjp)). The adjustment avoids problems if there are no "successes" or all "successes" in a group. What constitutes a "success" can be specified with yes= and the proportion adjustment can be turned off (if no group proportions are likely to be zero or one) with padj=FALSE.

The function plots the log odds versus the mean of the explanatory variable within each group. A least square line is fit to these points. The plot can be suppressed with showplot=FALSE.

The out=TRUE option will return a dataframe with the boundaries of each group, proportion, adjusted proportion, mean explanatory variable, and (adjusted or unadjusted) log odds.
emplogitplot2

Empirical logit plot for one quantitative variable by categorical groups

Description

This function produces an empirical logit plot for a binary response variable and with a single quantitative predictor variable broken down by a single categorical factor.

Usage

emplogitplot2(formula, data = NULL, ngroups = 3, breaks = NULL, yes = NULL, padj = TRUE, out = FALSE, showplot = TRUE, showline = TRUE, ylab = "Log(Odds)", xlab = NULL, putlegend = "n", levelcol = NULL, pch = NULL, main = "", ylim = NULL, xlim = NULL, lty = NULL, lwd = 1, cex = 1)

Arguments

formula A formula of the form (binary) Response~Quantitative Predictor+Factor
data A dataframe
ngroups Number of groups to use (not needed if breaks is used), ngroups="all" uses all unique values
breaks A vector of endpoints for the bins (not needed if ngroups is used)
yes Set a value for the response to be counted for proportions (optional)
padj Should proportions be adjusted to avoid zero and one? (default is TRUE)
out Should the function return a dataframe with group and factor information? (default is FALSE)
showplot Show the plot? default is TRUE

Value

A dataframe with group information (if out=TRUE)

Examples

data(MedGPA)
emplogitplot1(Acceptance~GPA,data=MedGPA)

GroupTable=emplogitplot1(Acceptance~MCAT,ngroups=5,out=TRUE,data=MedGPA)

emplogitplot1(Acceptance~MCAT,data=MedGPA,breaks=c(0,34.5,39.5,50.5),dotcol="red",linecol="black")
data(Putts1)
emplogitplot1(Made~Length,data=Putts1,ngroups="all")
showline Show the regression lines? default is TRUE
ylab Text label for the vertical axis (default is "Log(Odds)")
xlab Text label for the horizontal axis (default is NULL)
putlegend Position for the legend (default is "n" for no legend)
levelcol Vector of colors for the factor levels
pch Plot character for the dots
main Title for plot
ylim Limits for the vertical axis
xlim Limits for the horizontal axis
lty Line type (default is 1)
lwd Line width (default is 1)
cex Multiplier for plot symbols

Details
Values of the quantitative explanatory variable will be grouped into ngroups roughly equal sized groups, unless breaks is used to determine the boundaries of the groups. Using ngroups="all" will make each distinct value of the explanatory variable its own group.

We find a proportion for the binary response variable within each of the groups created from the quantitative variable crossed with the categorical variable. To avoid problems with proportions of zero and one, we compute an adjusted proportion with \((\text{Number yes} + 0.5)/\text{(Number of cases + 1)}\). This is converted to an adjusted log odds \(\log(\text{adjp}/(1-\text{adjp}))\). What constitutes a "success" can be specified with yes= and the proportion adjustment can be turned off (if no group proportions are likely to be zero or one) with padj=FALSE.

The function plots the log odds versus the mean of the explanatory variable within each group with different colors for each of the categories defined by the categorical variable. A least square line is fit to these points within each categorical group. The plot can be suppressed with showplot=FALSE.

The out=TRUE option will return a dataframe with the boundaries of each group, proportion, adjusted proportion, mean explanatory variable, and (adjusted or unadjusted) log odds.

Value
A dataframe with group information (if out=TRUE)

Examples
```r
data(MedGPA)
emplogitplot2(Acceptance~GPA+Sex,data=MedGPA)

GroupTable2=emplogitplot2(Acceptance~MCAT+Sex,ngroups=5,out=TRUE,data=MedGPA,putlegend="topleft")

emplogitplot2(Acceptance~MCAT+Sex,data=MedGPA,breaks=c(0,34.5,39.5,50.5),
levelcol=c("red","blue"),putlegend="bottomright")```
Effects of Oxygen on Sugar Metabolism

**Description**

Experiment on the effects of oxygen on sugar metabolism by bacteria

**Format**

A data frame 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.

Renamed to SugarEthanol in second edition.

**Source**


The original article is Yamada T., Takahashi-Abbe S., Abbe K. (1985) "Effects of oxygen concentration on pyruvate formatelyase in situ and sugar metabolism of Streptococcus mutans and Streptococcus sanguis," Infection and Immunity, pp. 129-134.

---

Pupil Dilation and Sexual Orientation

**Description**

Data from an experiment relating pupil dilation to sexual orientation.
Faces

Format

A data frame with 106 observations on the following 4 variables.

DilateDiff  Difference in pupil dilation when looking at same-sex and opposite-sex nude photographs
Sex  F=female or M=male
Gay  1=gay or 0=not, based on Kinsey scale score greater than 3
SexMale  0=female or 1=male

Details

DilateDiff is, essentially, the difference in pupil dilation when looking at (a) same-sex nudes and (b) opposite-sex nude photographs. More specifically, multiple measurements of pupil size were taken under each of the two conditions, together with a third condition that involved a neutral stimulus. Within-subject z-scores were then computed, which led to the DilateDiff numbers used here.

Source

G. Rieger and R.C. Savin-Williams (2012), "The Eyes Have It: Sex and Sexual Orientation Differences in Pupil Dilation Patterns," in PLoS ONE. The full study included 325 students. Here we are analyzing a subset of the data that excludes White students.

Facial Attractiveness of Men

Description

Grip strength, attractiveness, and shoulder-hip ratio for men

Format

A data frame with 38 observations on the following 5 variables.

MaxGripStrength  Measurement of strength of hand grip
SHR  Shoulder to hip ratio
Partners  Number of sexual partners (lifetime)
Attractive  Attractiveness rating
AgeFirstSex  Age of first sex

Details

Facial attractiveness of several men was rated by female college students. Maximum grip strength was also measured, along with shoulder to hip ratio, age of first sex, and number of sex partners.

Source

FaithfulFaces  

**Faithfulness from a Photo?**

**Description**

Ratings from a facial photo and actual faithfulness.

**Format**

A data frame with 170 observations on the following 7 variables.

- **SexDimorph**  Rating of sexual dimorphism (masculinity for males, femininity for females)
- **Attract**  Rating of attractiveness
- **Cheater**  Was the face subject unfaithful to a partner? (1=yes or 0=no)
- **Trust**  Rating of trustworthiness
- **Faithful**  Rating of faithfulness
- **FaceSex**  Sex of face (F=female or M=male)
- **RaterSex**  Sex of rater (F=female or M=male)

**Details**

College students were asked to look at a photograph of an opposite-sex adult face and to rate the person, on a scale from 1 (low) to 10 (high), for attractiveness. They were also asked to rate trustworthiness, faithfulness, and sexual dimorphism (i.e., how masculine a male face is and how feminine a female face is). Overall, 68 students (34 males and 34 females) rated 170 faces (88 men and 82 women).

**Source**

This dataset is based on G. Rhodes et al. (2012), "Women can judge sexual unfaithfulness from unfamiliar men’s faces," Biology Letters, November 2012. All of the 68 raters were heterosexual Caucasians, as were the 170 persons who were rated. (We have deleted 3 subjects with missing values and 16 subjects who were over age 35.)

---

FantasyBaseball  

**Selection Times in a Fantasy Baseball Draft**

**Description**

Draft selection times for a fantasy baseball league

**Format**

A data frame 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 selection at each round.

Source
Mathematical Science Baseball League historical records (online).

FatRats

<table>
<thead>
<tr>
<th>Diet and Weight of Rats</th>
</tr>
</thead>
</table>

Description
Experiment on effects of diets on weight gain of rats

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

<table>
<thead>
<tr>
<th>Gain</th>
<th>Weight gain (in grams per week)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>Level of protein (Hi or Lo)</td>
</tr>
<tr>
<td>Source</td>
<td>Source of protein (Beef, Cereal, or Pork)</td>
</tr>
</tbody>
</table>

Details
Data from this experiment compared weight gain for 60 baby rats that were fed different diets. Half of the rats had low-protein diets (Lo) and the rest had high-protein (Hi). The source of protein was either beef, cereal, or pork.

Source
C. P. Wilsie, Iowa State College Agricultural Station (1944) via Snedecor and Cochran
Fertility Data for Women Having Trouble Getting Pregnant

**Description**

Fertility measurements for a sample of women who have difficulty getting pregnant.

**Format**

A data frame 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 trouble 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.

---

Results of NFL Field Goal Attempts

**Description**

Field goal results in the National Football League (NFL) by distance.

**Format**

A data frame 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 Number 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 Data from Leonard Maltin’s Guide

Description

Film data from Maltin’s Movie and Video Guide

Format

A data frame with 100 observations on the following 9 variables.

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

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 statistician 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)

<table>
<thead>
<tr>
<th>FinalFourIzzo</th>
<th>NCAA Final Four by Seed and Tom Izzo (through 2010)</th>
</tr>
</thead>
</table>

Description

NCAA Final Four by seed with indicator for Tom Izzo’s teams from 1985 - 2010.

Format

A dataset with 1664 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Year (1985 - 2010)</td>
</tr>
<tr>
<td>Seed</td>
<td>Seed in NCAA men’s basketball tournament: 1 to 16</td>
</tr>
<tr>
<td>Final4</td>
<td>1=made Final Four or 0=did not make Final Four</td>
</tr>
<tr>
<td>Izzo</td>
<td>1=team coached by Tom Izzo or 0=not an Izzo team</td>
</tr>
</tbody>
</table>

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 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.

Updated to FinalFourIzzo17 in second edition.

Source

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

<table>
<thead>
<tr>
<th>FinalFourIzzo17</th>
<th>NCAA Final Four by Seed and Tom Izzo (through 2017)</th>
</tr>
</thead>
</table>

Description

NCAA Final Four by seed with indicator for Tom Izzo’s teams for 1985 - 2017
Format

A data frame with 2112 observations on the following 4 variables.

- Year  Year 1985 - 2017
- 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 is an extension of FinalFourIzzo (that ended in 2017) and the same as FinalFourLong2017, except the data starts in 1985 and we have an 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

Description

NCAA Final Four by seed with individual cases for each team each year

Format

A data frame 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 Division 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.

Updated to FinalFourLong17 in second edition.

Source

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

---

**FinalFourLong17**

**NCAA Final Four by Seed (Long Version through 2017)**

**Description**

NCAA Final Four by seed with individual cases for each team each year

**Format**

A data frame with 2496 observations on the following 4 variables.

- **Year**  Year (1979 - 2017)
- **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 Division I Men's Basketball tournament from 1979 to 2017 along with its seed and an indicator for whether the team made the Final Four that year. This dataset is an extension of FinalFourLong (that went through 2010).

**Source**

Final Four teams and their seed can be found at http://www.championshiphistory.com/ncaahoops.php
FinalFourShort17

CAA Final Four by Seed (Short Version through 2010)

Description

NCAA Final Four participation summarized each year by seed

Format

A data frame with 512 observations on the following 4 variables.

- **Year**: Year (1979 - 2010)
- **Seed**: Seed in NCAA men’s basketball tournament: 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 make 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.

Updated to FinalFourShort17 in second edition.

Source

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

---

FinalFourShort17

NCAA Final Four by Seed (Short Version through 2017)

Description

NCAA Final Four participation summarized each year by seed

Format

A data frame with 624 observations on the following 4 variables.

- **Year**: Year 1979 to 2017
Seed  Seed in NCAA men’s basketball tournament: 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 make 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 FinalFourLong2017, 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. This dataset is an extension of FinalFourShort (that went through 2010).

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

<table>
<thead>
<tr>
<th>Fingers</th>
<th>Finger Tap Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description
Finger tap rates after drug administration

Format
A data frame with 12 observations on the following 4 variables.

Subject  Subject code (I, II, III, or IV)
Drug     Drug administered (Ca=caffeine, Pl=placebo, or Th=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


FishEggs

**FishEggs**

**FirstYearGPA**

**First Year GPA for College Students**

**Description**

Predicting first-year college GPA

**Format**

A data frame with 219 observations on the following 10 variables.

- **GPA** First-year college GPA on a 0.0 to 4.0 scale
- **HSGPA** High school GPA on a 0.0 to 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**

Fertility measurement for eggs from a sample of 35 lake trout
Format

A data frame with 35 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age of the fish (in years)</td>
</tr>
<tr>
<td>PctDM</td>
<td>Percentage of the total egg material that is solid</td>
</tr>
<tr>
<td>Month</td>
<td>Month fish was caught: Sep=September or Nov=November</td>
</tr>
<tr>
<td>Sept</td>
<td>Indicator with 1=September or 0=November</td>
</tr>
</tbody>
</table>

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


---

Fitch

Body Measurements of Mammal Species

Description

Body measurements for a sample of 28 mammal species from a Fitch paper on acoustic allometry

Format

A data frame with 28 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>species of mammal</td>
</tr>
<tr>
<td>Order</td>
<td>Order (Carnivora or Primates)</td>
</tr>
<tr>
<td>Wt</td>
<td>Body weight (in kg)</td>
</tr>
<tr>
<td>Skull</td>
<td>Skull length (in cm)</td>
</tr>
<tr>
<td>Palate</td>
<td>Palate length (in cm)</td>
</tr>
</tbody>
</table>

Details

Data on mammal species from a Zoology paper about acoustic allometry by W. Tecumseh Fitch.

Source

**FlightResponse**  
*Response of Migratory Geese to Helicopter Overflights*

### Description

Flight response of Pacific Brant to overflights of helicopters

### Format

A dataset with 464 observations on the following 7 variables.

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

### 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

Fluorescence

Format
A data frame with 326 observations on the following 4 variables.

Penalty    Was death penalty given? (No or Yes)
Defendant  Race of the defendant (Black or White)
White.Victim Was the victim white? (1=yes or 0=no)
Black.Victim Was the victim black? (1=yes or 0=no)

Details
Mike Radelet’s data on imposition of the death penalty for murderers in Florida broken down by race of the victim and defendant.

Source

Measuring Calcium Binding to Proteins

Description
Data from an experiment on calcium binding to proteins

Format
A data frame 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.
**FranticFingers**

### Finger Tap Rates

**Description**

Finger tap rates after drug administration

**Format**

A data frame with 12 observations on the following 4 variables.

- **ID** Case ID
- **Rate** Finger taps in a fixed time interval
- **Subj** Subject code (A, B, C, or D)
- **Drug** Drug administered (Ca=caffeine, Pl=placebo, or Th=theobromine)

**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**


---

**FruitFlies**

### Fruit Fly Sexual Activity and Longevity

**Description**

Sexual activity and lifetimes of fruit flies

---
**Format**

A data frame with 125 observations on the following 7 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>Partners</td>
<td>Number of female partners: 0, 1, or 8</td>
</tr>
<tr>
<td>Type</td>
<td>0=pregnant, 1=virgin, 9=none</td>
</tr>
<tr>
<td>Longevity</td>
<td>Lifespan (in days)</td>
</tr>
<tr>
<td>Thorax</td>
<td>Length of thorax (in mm)</td>
</tr>
<tr>
<td>Sleep</td>
<td>Percent of day sleeping</td>
</tr>
<tr>
<td>Treatment</td>
<td>1 pregnant, 1 virgin, 8 pregnant, 8 virgin, or none</td>
</tr>
</tbody>
</table>

**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**

http://www.amstat.org/publications/jse/v2n1/datasets.hanley.html

---

**FruitFlies2**

**Fruit Fly Sexual Activity and Male Competition**

**Description**

Results from an experiment on male fruit flies with different levels of sexual activity and competition from other males.

**Format**

A data frame with 201 observations on the following 7 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mated</td>
<td>Was the fly allowed mating opportunities? (n or y)</td>
</tr>
<tr>
<td>Alone</td>
<td>Did the fly live alone? (y=yes or n= no, lived near another male)</td>
</tr>
</tbody>
</table>
Mating  How many mating opportunities was the fly given?
Total  Total duration of mating time over all opportunities (in seconds)
Size   Size of the thorax (in mm)
Lifespan  Lifespan (in hours, starting at the 12th day)
Activity  Number of times a movement detector was tripped starting in the 12th day

Details

Researchers randomly assigned virgin male fruit flies to one of two treatments: live alone or live in an environment where they can sense one other male fly. Flies were randomly allocated to either have mating opportunities with female flies or to not have such opportunities. Those flies that were given mating opportunities were given 3, 4, or 5 opportunities to mate (Mating measures this number). Researchers also measured size, lifespan and activity levels of the fruit flies.

Source

The file we are using is the link called survival at
http://rsbl.royalsocietypublishing.org/content/suppl/2013/02/25/rsbl.2012.1188.DC1.html
The article talking about the data is at
http://rsbl.royalsocietypublishing.org/content/9/2/20121188.full

<table>
<thead>
<tr>
<th>FunnelDrop</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Funnel Drop Times</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Description**

Experiment with a ball swirling thorough a funnel

**Format**

A data frame with 120 observations on the following 3 variables.

Funnel  Height of the funnel (inches)
Tube    Height of the drop tube (inches)
Time    Time (in seconds) for the ball to drop/swirl though the funnel

**Details**

Data from a class experiment to see where a steel ball was rolled through a plastic tube into a long plastic funnel. The angle of the funnel and the angle of the tube with respect to the flat table could be adjusted by changing the height of either (Funnel measured from the table, Tube measured from the top of the funnel). The ball rolls down the tube, then swirls around the funnel until dropping out at the bottom. Total trip time was measured with a stopwatch. Heights were adjusted after every two drops in a randomized order.
Source

The funnel dropping experiment was originally described in Gunter, B. (1993) "Through a Funnel Slowly with Ball Bearing and Insight to Teach Experimental Design," The American Statistician, Vol. 47. These data come from a class experiment based on the setup in that article.

---

GlowWorms

**Female Glow-worms**

---

Description

Brightness and fecundity of female glow-worms

Format

A data frame with 26 observations on the following 2 variables.

- **Lantern**  Length of glow lantern (in mm)
- **Eggs**  Number of eggs laid

Details

Data on 26 female glow-worms captured in Finland. Female glow-worms attract males by glowing with part of their abdomen (lantern). Researchers believe the brightness of glow might be related to mating success.

Source


http://dx.doi.org/10.1098/rsbl.2015.0599

---

Goldenrod

**Goldenrod Galls**

---

Description

Measurements for a sample of goldenrod galls
**Format**

A data frame 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**: Fate = 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 2004 (in mm)
- **Fate04**: Fate = 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.

---

**GrinnellHouses**  
*House Sales in Grinnell, Iowa*

**Description**

Data on houses sold between 2005 and 2015 in Grinnell, Iowa

**Format**

A data frame with 929 observations on the following 15 variables.

- **Date**: Coded value for date of sale (Jan 1, 2005=16436)
- **Address**: Street address of the house
- **Bedrooms**: Number of bedrooms
- **Baths**: Number of bathrooms
- **SquareFeet**: The square footage of the home’s living space
- **LotSize**: Lot size (in acres)
- **YearBuilt**: Year the house was built; many pre-1900 homes are listed as 1900
- **YearSold**: The year the house was sold, for this case
- **MonthSold**: The month the house was sold (1=Jan, 2=Feb, to 12=Dec)
- **DaySold**: Day of the month the house was sold (1 to 31)
Details

A local Grinnell realtor, Matt Karjalahti, put these data together to see what patterns might be found, perhaps with an improvement in how one sells houses or buys them. He asked Grinnell College economists, Lee Logan and Eric Ohrn, to help with the analysis and we obtained the data from them.

Source

Thanks to Grinnell realtor Matt Karjalahti who originally collected the data and Grinnell College economists Lee Logan and Eric Ohrn who gave us the data.

<table>
<thead>
<tr>
<th>Grocery</th>
<th>Grocery Sales and Discounts</th>
</tr>
</thead>
</table>

Description

Grocery store sales with different discounts

Format

A data frame with 36 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Discount</th>
<th>Amount of discount: 5.00%, 10.00%, or 15.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store</td>
<td>Store number (1-12)</td>
</tr>
<tr>
<td>Display</td>
<td>Featured End of Aisle, Featured Middle of Aisle, or Not Featured</td>
</tr>
<tr>
<td>Sales</td>
<td>Number sold during one week</td>
</tr>
<tr>
<td>Price</td>
<td>Wholesale price (in dollars)</td>
</tr>
</tbody>
</table>

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**

*Are Gunnels Present at Shoreline?*

**Description**

Presence/absence of gunnels (eels) at shoreline quadrats

**Format**

A data frame with 1592 observations on the following 10 variables.

<table>
<thead>
<tr>
<th>Gunnel</th>
<th>1= gunnel present in the quadrat or 0=gunnel absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Minutes after midnight</td>
</tr>
<tr>
<td>Fromlow</td>
<td>Time in minutes from low tide</td>
</tr>
<tr>
<td>Slope</td>
<td>Slope (to nearest 10 degrees) perpendicular to waterline</td>
</tr>
<tr>
<td>Rw</td>
<td>Percentage cover in quadrat of rockweed/algae/plants</td>
</tr>
<tr>
<td>Amphiso</td>
<td>Density of crustacean food: 0=none to 4=high</td>
</tr>
<tr>
<td>Subst</td>
<td>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</td>
</tr>
<tr>
<td>Pool</td>
<td>Standing water deep? 1=yes or 2=no</td>
</tr>
<tr>
<td>Water</td>
<td>Standing water in the quadrat? 1=yes or 2=no</td>
</tr>
<tr>
<td>Cobble</td>
<td>Rocky cobbles? 1=yes or 2=no</td>
</tr>
</tbody>
</table>

**Details**

This dataset 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.

---

**Handwriting**

*Guess Author’s Sex from Handwriting?*

**Description**

Survey data to see if subjects can guess author’s sex from handwriting specimens
Format

A data frame with 204 observations on the following 8 variables.

- Individual Survey Respondent Number
- Gender Gender of Respondent (0 = male, 1 = female)
- Survey1 Percent correct on Survey 1
- Survey2 Percent correct on Survey 2
- FemaleID Percent correct in identifying female specimens on Survey 1
- MaleID Percent correct in identifying male specimens on Survey 1
- Both Percent correctly identified on Survey 1 AND Survey 2
- DIFF Survey1 - Survey2

Details

Bradley and colleagues at Clarke University gave two identical surveys to a sample of 203 students (each student did the survey twice). Each survey contains 25 writing specimens and students were asked to identify whether the author is male or female. Of the 25 specimens, 12 are written by a female, 13 by a male.

An example of the survey form can be found at https://docs.google.com/forms/d/1sO6vIozsORbqaCTsA7Ta0qZL7_6_MCEPJ7tYeKYYvI/viewform

Source


Hawks

Measurements on Three Hawk Species

Description

Data for a samples of hawks from three different species

Format

A data frame with 908 observations on the following 19 variables.

- Month 8=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
Details

Students and faculty at Cornell College in Mount Vernon, Iowa, collected 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

Tail Lengths of Hawks

Description

Tail lengths for two hawk species

Format

A data frame with 838 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Species</th>
<th>RT=Red-tailed, SS=Sharp-shinned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tail</td>
<td>Length of tail (in mm)</td>
</tr>
</tbody>
</table>

Details

Tail lengths measured for a sample of 838 hawks observed in Mount Vernon, Iowa. Note: HawkTail2 has these data in unstacked format and they are a subset of the data in Hawks which has a third species (Cooper’s hawk).
HawkTail2

**Tail Lengths of Hawks (Unstacked)**

**Description**

Tail lengths for two hawk species

**Format**

A data frame with observations on the following 2 variables.

- **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 hawks observed in Mount Vernon, Iowa. Note: HawkTail has similar data in stacked format. The Hawks dataset has more variables and a third species (Cooper’s hawk).

**Source**

Observations by students and faculty at Cornell College.

HearingTest

**Correctly Identified Words in a Hearing Test**

**Description**

Percentaged of correctly identified words in a hearing test

**Format**

A data frame 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


References


HeatingOil

Description

Monthly US residential consumption of fuel oil (1983-2016)

Format

A data frame with 408 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Year</th>
<th>Year (1983 to 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>Month (1=Jan through 12=Dec)</td>
</tr>
<tr>
<td>t</td>
<td>Time index (1 to 408)</td>
</tr>
<tr>
<td>FuelOil</td>
<td>Residential consumption of fuel oil (in 1,000 barrels/day)</td>
</tr>
</tbody>
</table>

Details

U.S. residential consumption of distillate fuel oil each month from January 1983 through December 2016.

Source

**HighPeaks**  
*Characteristics of Adirondack Hiking Trails*

**Description**

Data on hiking trails for each of the 46 "High Peaks" in the Adirondack mountains

**Format**

A data frame with 46 observations on the following 6 variables.

<table>
<thead>
<tr>
<th>Peak</th>
<th>Name of the mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>Elevation at the highest point (in feet)</td>
</tr>
<tr>
<td>Difficulty</td>
<td>Rating of difficulty of the hike: 1 (easy) to 7 (most difficult)</td>
</tr>
<tr>
<td>Ascent</td>
<td>Vertical ascent (in feet)</td>
</tr>
<tr>
<td>Length</td>
<td>Length of hike (in miles)</td>
</tr>
<tr>
<td>Time</td>
<td>Expected trip time (in hours)</td>
</tr>
</tbody>
</table>

**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 gives information about the hiking trails up each of these peaks.

**Source**

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

---

**Hoops**  
*Grinnell College Basketball Games*

**Description**

Data from games played by the Grinnell College men’s basketball team between 1997 and 2006

**Format**

A data frame with 147 observations on the following 22 variables.

<table>
<thead>
<tr>
<th>Game</th>
<th>An ID number assigned to each game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opp</td>
<td>Name of the opponent school for the game</td>
</tr>
<tr>
<td>Home</td>
<td>Indicator variable where 1 = home game and 0 = away game</td>
</tr>
</tbody>
</table>
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.

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

---

**HorsePrices**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HorseID</td>
<td>ID code for each horse</td>
</tr>
<tr>
<td>Price</td>
<td>Price (in dollars)</td>
</tr>
<tr>
<td>Age</td>
<td>Age of the horse (in years)</td>
</tr>
<tr>
<td>Height</td>
<td>Height of the horse (in hands)</td>
</tr>
<tr>
<td>Sex</td>
<td>f=female m=male</td>
</tr>
</tbody>
</table>

**Prices of Horses**

Price and related characteristics of horses listed for sale on the internet

**Format**

A data frame with 50 observations on the following 5 variables.
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**

*House Prices, Sizes, and Lot Areas*

**Description**

Selling price and characteristics for a sample of 20 houses in a small town

**Format**

A data frame 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.

Updated to HousesNY in second edition.

**Source**

Data collected from zillow.com in June 2008.

---

**HousesNY**

*House Prices in Rural NY*

**Description**

House prices for a sample of houses in Canton NY
ICU

**Format**

A data frame with 53 observations on the following 5 variables.

- **Price**: Estimated price (in $1,000’s)
- **Beds**: Number of bedrooms
- **Baths**: Number of bathrooms
- **Size**: Floor area of the house (in 1,000 square feet)
- **Lot**: Size of the lot (in acres)

**Details**

Data scraped from Zillow.com for a sample of houses near the 13617 area code (Canton, NY a small town in upstate NY). Houses on lots bigger than five acres (often farms) were excluded.

**Source**

Data scraped from the Zillow.com website using tools an app at http://myslu.stlawu.edu/~clee/dataset/zillow/ (April 2017)

---

**ICU**

*Intensive Care Unit Patients*

**Description**

Data for a sample of 200 patients at an Intensive Care Unit (ICU)

**Format**

A data frame 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**: Heart rate (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 downloaded from The Data and Story Library (DASL), http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html.

Infant Mortality Rates

Description

Infant mortality rates in the United States by decade (1920-2010)

Format

A data frame with 10 observations on the following 2 variables.

- Mortality  Deaths within one year of birth (per 1000 births)
- Year  Year (1920-2010 by decades)

Details

Infant mortality (deaths within one year of birth per 1,000 births) in the US from 1920 - 2010 (by decade).

Source


Monthly Consumer Price Index (2009-2016)

Description

Consumer Price Index (CPI) each month for 2009 through 2016

Format

A data frame with 96 observations on the following 5 variables.

- Month  Month: 1=January to 12=December
- Year  Year (2009 to 2016)
- CPI  Consumer Price Index (base=100 in 1984)
- CPIPctDiff  Monthly percent change in CPI
- t  Time index (1 to 96)
Details

Monthly Consumer Price Index for 2009 to 2016 as produced by the Bureau of Labor Statistics (Series Id. CUUR0000SA0). Based on prices for all items in U.S. city average for all consumers (not seasonally) Base period is 1982-1984-100.

Source


<table>
<thead>
<tr>
<th>InsuranceVote</th>
<th>Congressional Votes on a Health Insurance Bill</th>
</tr>
</thead>
</table>

Description

Congressional votes on an ObamaCare health insurance bill in 2009

Format

A dataset with 435 observations on the following 9 variables.

- **Party**: Party affiliation: 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_main/). Roll call of congressional votes on this bill can be found at http://clerk.house.gov/evs/2009/roll887.xml.
**IQGuessing**

**Description**

True IQ and guessed IQ (from a photo) for 40 women.

**Format**

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

- **Age** Age of woman
- **GuessIQ** Guessed IQ
- **TrueIQ** Actual IQ

**Details**

One hundred sixty raters (75 men and 85 women) took part in judging intelligence (on a 1=high to 7=low scale) based on photographs of students. The ratings were converted to z-scores and then put on an IQ scale to compare to actual measured IQ. There were photos of 80 students, 40 men and 40 women. This data set contains data for the 40 women.

**Source**


---

**Jurors**

**Description**

Reporting rates for bi-weekly jury pools in Franklin County Court (Columbus, OH).

**Format**

A data frame with 52 observations on the following 4 variables.

- **Period** Sequential 2-week periods over 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 dataset 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

---

Kershaw Pitch Data

Description

Pitch-by-pitch data for baseball pitcher Clayton Kershaw in the 2013 season

Format

A data frame with 3402 observations on the following 24 variables.

- **BatterNumber**: Number of batters faced so far that game
- **Outcome**: One of 14 possible results for a pitch (e.g., Ball, Ball In Dirt, Called Strike, ..., Swinging Strike (Blocked))
- **Class**: One of three classifications (B=ball, S=strike, or X=in play)
- **Result**: From pitcher’s perspective (Neg=ball or hit, Pos=strike or out)
- **Swing**: Did the batter swing at the pitch? (No or Yes)
- **Time**: Date and time of the pitch (format yyyy-mm-ddThh:mm:ssZ)
- **StartSpeed**: Speed leaving the pitcher’s hand (in mph)
- **EndSpeed**: Speed crossing home plate (in mph)
- **HDev**: Horizontal movement (inches)
- **VDev**: Vertical movement (inches)
- **HPos**: Horizontal position at home plate (inches from center, positive is catcher’s right)
- **VPos**: Vertical position at home plate (inches above the ground)
- **PitchType**: Code for pitch type (CH=changeup, CU=curve, FF=fastball, or SL=slider)
- **Zone**: 1-9 in theoretical strike zone (upper left to lower right), 11-14 are out of strike zone
Nasty  A measure on a 0-100 scale of difficulty of the pitch to hit (100 is most difficult)
Count  Ball strike count (0−0, 0−1, 0−2, 1−1, 1−2, 2−1, 2−2, 3−1, or 3−2)
BallCount  Number of balls before the pitch (0, 1, 2, or 3)
StrikeCount  Number of strikes before the pitch (0, 1, or 2)
Inning  Inning of the game
InningSide  Portion of the inning (bottom= pitcher at home or top=pitcher away)
 Outs  Number of outs when the pitch is thrown
BatterHand  Batter’s stance (L=left or R=right)
ABEvent  Result of the at bat (several possibilities)
Batter  Name of the batter faced

Details

Dataset includes information for 3,402 individual pitches thrown by Los Angeles Dodger baseball pitcher Clayton Kershaw during the 2013 regular season when he won the Cy Young award as the best pitcher in the National League. Many variables are measured using Major League Baseball’s PITCHf/x system that uses camera systems in each ballpark to track characteristics of each pitch thrown.

Source

Data scraped from the MLB GameDay website (http://gd2.mlb.com/components/game/mlb/) using pitchRx

---

**KeyWestWater**  
**Key West Water Temperatures**

**Description**

Hourly water temperatures from Gulf of Mexico near Key West, Florida

**Format**

A data frame with 6572 observations on the following 3 variables.

- **DateTime**  Date and time of reading (format mm/dd/yyyy h:00)
- **WaterTemp**  Water temperature (in degrees Fahrenheit)
- **t**  Time index (1 to 673)

**Details**

Hourly readings of water temperatures from a measuring device in the Gulf of Mexico near Key West, Florida. The hourly temperatures are provided from October 3, 2016 to October 3, 2017 and were obtained from station 8724580. A few missing values have been interpolated to provide a complete series.
Leafhoppers

Source


Data were obtained by Kyle Johnston for his Senior Exercise (a capstone project).

---

Kids198  

**Body Measurements of Children**

**Description**

Body measurements for a sample of 198 children

**Format**

A data frame with 198 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Height (in inches)</td>
</tr>
<tr>
<td>Weight</td>
<td>Weight (in pounds)</td>
</tr>
<tr>
<td>Age</td>
<td>Age (in months)</td>
</tr>
<tr>
<td>Sex</td>
<td>0=male or 1=female</td>
</tr>
<tr>
<td>Race</td>
<td>0=white or 1=other</td>
</tr>
</tbody>
</table>

**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 original study.

**Source**

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

---

Leafhoppers  

**Leafhopper Diet and Longevity**

**Description**

Lifetimes for potato leafhoppers on various sugar diets

**Format**

A data frame with 8 observations on the following 2 variables.
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).

---

### LeafWidth

<table>
<thead>
<tr>
<th>Diet</th>
<th>Control, Fructose, Glucose, or Sucrose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days</td>
<td>Number of days until half the leafhoppers in a dish died</td>
</tr>
</tbody>
</table>

---

### Leaf Measurements

**Description**
Measurements of Dodonaea viscosa leaves

**Format**
A data frame with 252 observations on the following 5 variables.

- **Width**: Average width (in mm)
- **Length**: Average length (in mm)
- **LWRatio**: Length divided by Width
- **Area**: Area (in sq. mm)
- **Year**: Year the leaves were collected

**Details**
Data on samples of leaves from the species Dodonaea viscosa subsp. angustissima (common name hopbush), which have been collected in a certain region of South Australia for many years.

**Source**
Leukemia

Responses to Treatment for Leukemia

Description

Treatment results for leukemia patients

Format

A data frame with 51 observations on the following 9 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age at diagnosis (in years)</td>
</tr>
<tr>
<td>Smear</td>
<td>Differential percentage of blasts</td>
</tr>
<tr>
<td>Infil</td>
<td>Percentage of absolute marrow leukemia infiltrate</td>
</tr>
<tr>
<td>Index</td>
<td>Percentage labeling index of the bone marrow leukemia cells</td>
</tr>
<tr>
<td>Blasts</td>
<td>Absolute number of blasts, in thousands</td>
</tr>
<tr>
<td>Temp</td>
<td>Highest temperature of the patient prior to treatment, in degrees Fahrenheit</td>
</tr>
<tr>
<td>Resp</td>
<td>1=responded to treatment or 0=failed to respond</td>
</tr>
<tr>
<td>Time</td>
<td>Survival time from diagnosis (in months)</td>
</tr>
<tr>
<td>Status</td>
<td>0=dead or 1=alive</td>
</tr>
</tbody>
</table>

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


LeveeFailures

Levee Failures along the Mississippi River

Description

Factors relating to Mississippi River levee failure

Format

A data frame with 82 observations on the following 14 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure</td>
<td>Did the levee fail? (1=yes or 0=no)</td>
</tr>
<tr>
<td>Year</td>
<td>Year</td>
</tr>
</tbody>
</table>
RiverMile Location along the river (mile marker)
Sediments Sediments present? (1=yes or 0=no)
BorrowPit Borrow pit present? (1=yes or 0=no)
Meander Type of meander (1=inside bend, 2=outside bend, 3=chute, 4=straight)
ChannelWidth Width of the river channel (in meters)
FloodwayWidth Width of floodway (in meters, levee to levee, levee to bluff, or bluff to bluff, as appropriate)
ConstrictionFactor Constriction of the floodway over time (1880s to present)
LandCover 1=open water, 2=grassy, 3=agricultural, 4=forest
VegWidth Vegetation buffer width (in meters)
Sinuosity River length divided by valley length for 10 miles up- and down-valley from levee site
Dredging Dredging intensity
Revetement Is there a stone structure (wall) meant to hold up the bank? (1=yes or 0=no)

Details
The goal of this investigation was to test the relative importance of geologic, geomorphic, and other physical factors that have led to levee failures through the past century along much of the Mississippi River.

Source

LewyBody2Groups

Description
Dementia study comparing two groups of patients

Format
A data frame with 39 observations on the following 3 variables.

Type DLB=Dementia with Lewy Bodies or DLB/AD=DLB and Alzheimer’s Disease
APC Annualized Percentage Change from baseline volume of the brain
MMSE Change in functional performance on the Mini Mental State Examination

Details
Brain MRIs were used to study the brains of patients with Dementia with Lewy Bodies, some of whom also were diagnosed with Alzheimer’s Disease.
LongJumpOlympics

Source


LewyDLBad

Lewy Bodies and Dimentia with Alzheimer’s

Description

Dementia Study with Lewy Bodies

Format

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

- **group**: DLB/AD=DLB and Alzheimer’s Disease
- **APC**: Annualized Percentage Change from baseline volume of the brain
- **MMSE**: Change in functional performance on the Mini Mental State Examination

Details

Brain MRIs were used to study the brains of patients with Dementia with Lewy Bodies. These are the cases that were also diagnosed with Alzheimer’s Disease. This is a subset of LewBody2Groups

Source


LongJumpOlympics

Olympic Men’s Long Jump Gold Medal Distance (1900 - 2008)

Description

Winning distances in men’s Olympic long jump competitions (1900 - 2008)

Format

A data frame 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
Historical Olympic long jump results at http://trackandfield.about.com/od/longjump/qt/olymlongjumpmen.htm

---

Description
Gold medal distance for Olympic men’s long jump

Format
A data frame with 28 observations on the following 2 variables.

Year  Olympic Year (1900-2016)
Gold  Gold medal distance (in meters)

Details
Gold medal winning distances for the men’s long jump at the Olympics from 1900 to 2016.

Source
Historical Olympic long jump results at http://trackandfield.about.com/od/longjump/qt/olymlongjumpmen.htm

---

Description
Hours of sleep for teenagers

Format
A data frame with 446 observations on the following 3 variables.

Person  Cased ID number
Age  Age (in years)
Outcome  Average at least 7 hours of sleep? (1=yes or 0=no)
Details

Data from a sample of 446 teens, aged 14 to 18, who answer the question, "On an average school night, how many hours of sleep do you get?" The outcome variable records whether or not each person averages at least 7 hours of sleep.

Source


---

Return Rates for "Lost" Letters

<table>
<thead>
<tr>
<th>Location</th>
<th>Address</th>
<th>Returned</th>
<th>DesMoines</th>
<th>GrinnellTown</th>
<th>GrinnellCampus</th>
<th>Peaceworks</th>
<th>Confederacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DesMoines</td>
<td>Confederacy</td>
<td>Returned</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GrinnellCampus</td>
<td>Peaceworks</td>
<td>Returned</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GrinnellTown</td>
<td>Peaceworks</td>
<td>Returned</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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
Daily Training for a Marathon Runner

Description
Training records for a marathon runner

Format
A dataset with 1128 observations on the following 9 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Training date</td>
</tr>
<tr>
<td>Miles</td>
<td>Miles for training run</td>
</tr>
<tr>
<td>Time</td>
<td>Training time (in minutes:seconds:hundredths)</td>
</tr>
<tr>
<td>Pace</td>
<td>Running pace (in minutes:seconds:hundredths per mile)</td>
</tr>
<tr>
<td>ShoeBrand</td>
<td>Adidas, Asics, Brooks, Izumi, Mizuno, or New Balance</td>
</tr>
<tr>
<td>TimeMin</td>
<td>Training time (in minutes)</td>
</tr>
<tr>
<td>PaceMin</td>
<td>Running pace (in minutes per mile)</td>
</tr>
<tr>
<td>Short</td>
<td>1= 5 miles or less or 0=more than 5 miles</td>
</tr>
<tr>
<td>After2004</td>
<td>1= for runs after 2004 or 0=for earlier runs</td>
</tr>
</tbody>
</table>

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.

Daily Change in Dow Jones and Nikkei Stock Market Indices

Description
Daily changes in two stock market indices

Format
A dataset with 56 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJI Ach</td>
<td>Change in Dow Jones Industrial Average</td>
</tr>
<tr>
<td>Date</td>
<td>Date: 06-Aug-09 to 02-Nov-09</td>
</tr>
<tr>
<td>Nik 225 Ch</td>
<td>Change in Nikkei 225 stock average</td>
</tr>
<tr>
<td>Up</td>
<td>Indicator for positive Nikkei change</td>
</tr>
<tr>
<td>Lag Nik</td>
<td>Previous day’s Nikkei change</td>
</tr>
</tbody>
</table>
MathPlacement

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%3A998313&Year=2009&Range=12&tag=cbsnewsSectionsArea
Historical Nikkei 225 index:
http://markets.cbsnews.com/cbsnews/quote/historical?
Month=11&Symbol=992%3A1900000035&Year=2009&Range=12&tag=cbsnewsSectionsArea

MathEnrollment  Enrollments in Math Courses

Description
Semester enrollments in mathematics courses

Format
A dataset with 11 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AYear</td>
<td>Academic year (for the fall)</td>
</tr>
<tr>
<td>Fall</td>
<td>Fall semester total enrollments</td>
</tr>
<tr>
<td>Spring</td>
<td>Spring semester total enrollments</td>
</tr>
</tbody>
</table>

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 Exam Results

Description
Results from a Math Placement exam at a liberal arts college
**Format**

A dataset with 2696 observations on the following 16 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Identification number for each student</td>
</tr>
<tr>
<td>Gender</td>
<td>0=Female, 1=Male</td>
</tr>
<tr>
<td>PSATM</td>
<td>PSAT score in Math</td>
</tr>
<tr>
<td>SATM</td>
<td>SAT score in Math</td>
</tr>
<tr>
<td>ACTM</td>
<td>ACT Score in Math</td>
</tr>
<tr>
<td>Rank</td>
<td>Adjusted rank in HS class</td>
</tr>
<tr>
<td>Size</td>
<td>Number of students in HS class</td>
</tr>
<tr>
<td>GPAAdj</td>
<td>Adjusted GPA</td>
</tr>
<tr>
<td>PlcmtScore</td>
<td>Score on math placement exam</td>
</tr>
<tr>
<td>Recommends</td>
<td>Recommended course: R0 R01 R1 R12 R2 R3 R4 R6 R8</td>
</tr>
<tr>
<td>Course</td>
<td>Actual course taken</td>
</tr>
<tr>
<td>Grade</td>
<td>Course grade</td>
</tr>
<tr>
<td>RecTaken</td>
<td>1=recommended course, 0=otherwise</td>
</tr>
<tr>
<td>TooHigh</td>
<td>1=took course above recommended, 0=otherwise</td>
</tr>
<tr>
<td>TooLow</td>
<td>1=took course below recommended, 0=otherwise</td>
</tr>
<tr>
<td>CourseSuccess</td>
<td>1=B or better grade, 0=grade below B</td>
</tr>
</tbody>
</table>

**Details**

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

**Source**

Personal correspondence

---

**MedGPA**  
*GPA and Medical School Admission*

**Description**

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

**Format**

A dataset with 55 observations on the following 11 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>Status: A=accepted to medical school or D=dENied admission</td>
</tr>
<tr>
<td>Acceptance</td>
<td>Indicator for Accept: 1=accepted or 0=dENied</td>
</tr>
<tr>
<td>Sex</td>
<td>F=female or M=male</td>
</tr>
<tr>
<td>BCPM</td>
<td>Bio/Chem/Physics/Math grade point average</td>
</tr>
<tr>
<td>GPA</td>
<td>College grade point average</td>
</tr>
<tr>
<td>VR</td>
<td>Verbal reasoning (subscore)</td>
</tr>
<tr>
<td>PS</td>
<td>Physical sciences (subscore)</td>
</tr>
</tbody>
</table>
Meniscus

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.

<table>
<thead>
<tr>
<th>Meniscus</th>
<th>Meniscus Repair Methods</th>
</tr>
</thead>
</table>

Description

Comparing meniscus repair methods on cadaver knees

Format

A data frame with 18 observations on the following 4 variables.

- **Method**  
  Meniscus repair method (1 = Vertical Suture, 2 = Meniscus Arrow, 3 = FasT-Fix)
- **FailureLoad**  
  Load at failure (in Newtons)
- **Displacement**  
  Displacement (in mm)
- **Stiffness**  
  Stiffness (Newtons/mm)

Details

Eighteen, lightly embalmed, cadaveric knee specimens were used in a study to compare three different methods of meniscus repair. The specimens were randomly assigned to one of the three treatments: vertical suture, meniscus arrow, FasT-Fix. They were evaluated on three different response variables: load at failure, stiffness, and displacement.

Source


Dataset downloaded from http://www.stat.ufl.edu/~winner/data/meniscus.txt
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.

<table>
<thead>
<tr>
<th>Month</th>
<th>Month of the year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moon</td>
<td>Relationship to full moon: After, Before, or During</td>
</tr>
<tr>
<td>Admission</td>
<td>Number of emergency room admissions</td>
</tr>
</tbody>
</table>

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


References


MetabolicRate  Metabolic Rate of Caterpillars

Description

Body size and metabolic rate of Manduca Sexta caterpillars
**MetroCommutes**

**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 metabolic 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.

---

**MetroCommutes**

**Commute Times**

**Description**

Commute times for four cities

**Format**

A data frame with 2000 observations on the following 3 variables.

- **City** Boston, Houston, Minneapolis, or Washington
- **Distance** Distance of commute (in miles)
- **Time** Time of commute (in minutes)

**Details**

The data are distances (miles) and times (minutes) of daily commute (one-way) for random samples of 500 commuters in each of four cities (Boston, Houston, Minneapolis, Washington) in 2007. The random samples were taken from the Metropolitan Public Use File of the 2007 American Housing Survey.

**Source**

**MetroHealth83**  
*Health Services in Metropolitan Areas*

**Description**

Health services data for 83 metropolitan areas

**Format**

A dataset with 83 observations on the following 16 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Name of the metropolitan area</td>
</tr>
<tr>
<td>NumMDs</td>
<td>Number of physicians</td>
</tr>
<tr>
<td>RateMDs</td>
<td>Number of physicians per 100,000 people</td>
</tr>
<tr>
<td>NumHospitals</td>
<td>Number of community hospitals</td>
</tr>
<tr>
<td>NumBeds</td>
<td>Number of hospital beds</td>
</tr>
<tr>
<td>RateBeds</td>
<td>Number of hospital beds per 100,000 people</td>
</tr>
<tr>
<td>NumMedicare</td>
<td>Number of Medicare recipients in 2003</td>
</tr>
<tr>
<td>PctChangeMedicare</td>
<td>Percent change in Medicare recipients (2000 to 2003)</td>
</tr>
<tr>
<td>MedicareRate</td>
<td>Number of Medicare recipients per 100,000 people</td>
</tr>
<tr>
<td>SSBNum</td>
<td>Number of Social Security recipients in 2004</td>
</tr>
<tr>
<td>SSBRate</td>
<td>Number of Social Security recipients per 100,000 people</td>
</tr>
<tr>
<td>NumRetired</td>
<td>Number of retired workers</td>
</tr>
<tr>
<td>SSINum</td>
<td>Number of Supplemental Security Income recipients in 2004</td>
</tr>
<tr>
<td>SSIRate</td>
<td>Number of Supplemental Security Income recipients per 100,000 people</td>
</tr>
<tr>
<td>SqrtMDs</td>
<td>Square root of number of physicians</td>
</tr>
</tbody>
</table>

**Details**

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

This dataset is in the first edition, but replaced by CountyHealth in the second edition.

**Source**

U.S. Census Bureau: 2006 State and Metropolitan Area Data Book (Table B-6)  
Description
Effects of transcranial magnetic stimulation (TMS) on migraine headaches

Format
A data frame with 2 observations on the following 4 variables.

Group  Treatment group (Placebo or TMS)
Yes  Count of number of patients that were pain-free in each group
No  Count of number of patients that had pain in each group
Trials  Number of patients in each 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.

This dataset was called TMS in the first edition.

Source

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  Standings and Team Statistics from the 2007 Baseball Season

Description

Data for Major League Baseball teams from the 2007 regular season

Format

A dataset with 30 observations on the following 21 variables.

<table>
<thead>
<tr>
<th>Team</th>
<th>Name of the team</th>
</tr>
</thead>
<tbody>
<tr>
<td>League</td>
<td>League: AL or NL</td>
</tr>
<tr>
<td>Wins</td>
<td>Number of wins for the season (out of 162 games)</td>
</tr>
</tbody>
</table>
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). Updated to MLBStandings2016 in second edition.

Source

Data downloaded from baseball-reference.com:

Description

Major League Baseball (MLB) standings and team statistics for the 2016 season

Format

A data frame with 30 observations on the following 21 variables.

- Team  Team name
- League  AL=Ameri. or NL=National
- Wins  Number of wins for the season (out of 162 games)
Details

Data for all 30 Major League Baseball (MLB) teams for the 2016 regular season. This includes team batting statistics (BattingAvg through SLG) and team pitching statistics (ERA through WHIP).

Source

Data downloaded from baseball-reference.com:

Description

Body size and eggs produced for a species of moths

Format

A dataset with 39 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BodyMass</td>
<td>Log of body size measured in grams</td>
</tr>
<tr>
<td>Eggs</td>
<td>Number of eggs present</td>
</tr>
</tbody>
</table>
Details

Researchers were interested in 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.

---

**MouseBrain**  
**Effects of Serotonin in Mice**

**Description**

Effects of altering serotonin levels on social interactions of mice

**Format**

A data frame with 48 observations on the following 3 variables.

Contacts  Number of social contacts the mouse had during the experiment  
Sex  F=female or M=male  
Genotype  Minus, Mixed, or Plus (see description below)

**Details**

Serotonin is a chemical that influences mood balance in humans. But how does it affect mice? Scientists genetically altered mice by "knocking out" the expression of a gene, tryptophan hydroxylase 2 (Tph2), that regulates serotonin production. With careful breeding, the scientists produced three types of mice that we label as “Minus” for Tph2-/-, “Plus” for Tph2+++, “Mixed” for Tph2+/-.. The variable Genotype records Minus/Plus/Mixed. The variable Contacts is the number of social contacts that a mouse had with other mice during an experiment and the variable Sex is “M” for males and “F” for females.

**Source**

http://dx.doi.org/10.1098/rsbl.2015.0057

Once you go to the above link, to get the data, click on the "Figures and Data" tab. Then click on the "Juvenile SocInter Behavior Data" link to download a hairy data file that needs to be cleaned a great deal to get our data.
MusicTime  

Estimating Time with Different Music Playing

Description

Estimates of 45 seconds with different music playing

Format

A data frame with 60 observations on the following 6 variables.

- **MusicBg**: Music playing in the background (no or yes)
- **Subject**: Code for each subject (subj1 through subj20)
- **Sex**: Subject's sex (f=female or m=male)
- **TimeGuess**: Subject's time estimating 45 seconds (in seconds)
- **Music**: Type of music (calm, control, or upbeat)
- **Accuracy**: Absolute value of TimeGuess minus 45

Details

Participants were asked to judge when 45 seconds had passed in silence (control), while listening to an upbeat song (Metropolis, by David Guetta and Nicky Romero), and while listening to a calm song (Bach's Das Wohltemperierte Klavier, Prelude in C Major). The order in which the three conditions were experienced was randomized for each participant. Time until subject guessed 45 seconds had elapsed (TimeGuess) and the magnitude of the difference from 45 (Accuracy) were recorded.

Source

Data collected by Ksenia Vlasov at Oberlin College.

NCbirths  

North Carolina Birth Records

Description

Data from births in North Carolina in 2001
NFL2007Standings

Format

A dataset with 1450 observations on the following 15 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Patient ID code</td>
</tr>
<tr>
<td>Plural</td>
<td>1=single birth, 2=twins, 3=triplets</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex of the baby: 1=male, 2=female</td>
</tr>
<tr>
<td>MomAge</td>
<td>Mother’s age (in years)</td>
</tr>
<tr>
<td>Weeks</td>
<td>Completed weeks of gestation</td>
</tr>
<tr>
<td>Marital</td>
<td>Marital status: 1=married or 2=not married</td>
</tr>
<tr>
<td>RaceMom</td>
<td>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</td>
</tr>
<tr>
<td>HispMom</td>
<td>Hispanic origin of mother: C=Cuban, M=Mexican, N=not Hispanic 0=Other Hispanic, P=Puerto Rico, S=Central/South America</td>
</tr>
<tr>
<td>Gained</td>
<td>Weight gained during pregnancy (in pounds)</td>
</tr>
<tr>
<td>Smoke</td>
<td>Smoker mom?: 1=yes or 0=no</td>
</tr>
<tr>
<td>BirthWeightOz</td>
<td>Birth weight in ounces</td>
</tr>
<tr>
<td>BirthWeightGm</td>
<td>Birth weight in grams</td>
</tr>
<tr>
<td>Low</td>
<td>Indicator for low birth weight, 1=2500 grams or less</td>
</tr>
<tr>
<td>Premie</td>
<td>Indicator for premature birth, 1=36 weeks or sooner</td>
</tr>
<tr>
<td>MomRace</td>
<td>Mother’s race: black, hispanic, other, or white</td>
</tr>
</tbody>
</table>

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

NFL Standings for 2007 Regular Season

Description

Standings for National Football League teams in 2007

Format

A dataset with 32 observations on the following 10 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>Team name</td>
</tr>
<tr>
<td>Conference</td>
<td>Conference: AFC or NFC</td>
</tr>
<tr>
<td>Division</td>
<td>Division within conference: ACE, ACN, ACS, ACW, NCE, NCN, NCS, NCW</td>
</tr>
<tr>
<td>Wins</td>
<td>Number of wins (out of 16 games)</td>
</tr>
<tr>
<td>Losses</td>
<td>Number of losses</td>
</tr>
</tbody>
</table>
**Details**


**Source**

Data downloaded from www.nfl.com

---

### NFLStandings2016

**NHL Standings for 2016 Regular Season**

---

**Description**

Standings and team statistics for National Football League (NFL) teams in the 2016 season

**Format**

A data frame with 32 observations on the following 11 variables.

- **Team** Team name
- **Wins** Wins in the 2016 regular season (out of 16 games)
- **Losses** Losses in the 2016 regular season
- **Ties** Ties in the 2016 regular season (ties are very rare in the NFL)
- **WinPct** Winning percentage = (Wins+0.5*Ties)/16 games
- **PointsFor** Points scored
- **PointsAgainst** Points allowed
- **NetPts** Points scored minus Points allowed
- **YardsFor** Offensive yards gained by the team
- **YardsAgainst** Offensive yards against the team
- **TDs** Touchdowns scored

**Details**

Standings for the 2016 regular season of the National Football League (NFL) along with points and scored and allowed for each team in its 16 games.
OilDeapsorption

Source
Data downloaded from:
http://www.pro-football-reference.com/years/2016/

<table>
<thead>
<tr>
<th>Nursing</th>
<th>Nursing Homes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beds</td>
<td>Number of beds in the nursing home</td>
</tr>
<tr>
<td>InPatientDays</td>
<td>Annual medical in-patient days (in hundreds)</td>
</tr>
<tr>
<td>AllPatientDays</td>
<td>Annual total patient days (in hundreds)</td>
</tr>
<tr>
<td>PatientRevenue</td>
<td>Annual patient care revenue (in hundreds of dollars)</td>
</tr>
<tr>
<td>NurseSalaries</td>
<td>Annual nursing salaries (in hundreds of dollars)</td>
</tr>
<tr>
<td>FacilitiesExpend</td>
<td>Annual facilities expenditure (in hundreds of dollars)</td>
</tr>
<tr>
<td>Rural</td>
<td>1=rural or 0=non-rural</td>
</tr>
</tbody>
</table>

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

OilDeapsorbtion

Effect of Ultrasound on Oil Deapsorption

Description
Experiment to measure the effect of ultrasound on deapsorbing oil from sand
### Olives

**Description**

Measurements of the pesticide fenthion in olive oil over time

**Format**

A dataset with 18 observations on the following 7 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SampleNumber</td>
<td>Code (1-6) for sample of olive oil</td>
</tr>
<tr>
<td>Group</td>
<td>Code for group: 1 or 2</td>
</tr>
<tr>
<td>Day</td>
<td>Time (in days) when sample was measured: 0, 281, or 365</td>
</tr>
<tr>
<td>Fenthion</td>
<td>Amount of fenthion (pesticide)</td>
</tr>
<tr>
<td>FenthionSulphoxide</td>
<td>Amount of fenthion sulfide</td>
</tr>
<tr>
<td>FenthionSulphone</td>
<td>Amount of fenthion sulphone</td>
</tr>
<tr>
<td>Time</td>
<td>Code (0, 3, or 4) for the number of days</td>
</tr>
</tbody>
</table>

---

### Fenthion in Olive Oil

**Description**

Measurements of the pesticide fenthion in olive oil over time

**Format**

A dataset with 18 observations on the following 7 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SampleNumber</td>
<td>Code (1-6) for sample of olive oil</td>
</tr>
<tr>
<td>Group</td>
<td>Code for group: 1 or 2</td>
</tr>
<tr>
<td>Day</td>
<td>Time (in days) when sample was measured: 0, 281, or 365</td>
</tr>
<tr>
<td>Fenthion</td>
<td>Amount of fenthion (pesticide)</td>
</tr>
<tr>
<td>FenthionSulphoxide</td>
<td>Amount of fenthion sulfide</td>
</tr>
<tr>
<td>FenthionSulphone</td>
<td>Amount of fenthion sulphone</td>
</tr>
<tr>
<td>Time</td>
<td>Code (0, 3, or 4) for the number of days</td>
</tr>
</tbody>
</table>
Orings

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


Space Shuttle O-Rings

Description

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

Format

A dataset with 24 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Temp</th>
<th>Code for temperature (in degrees F): Above65 Below65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failures</td>
<td>Number of O-ring failures</td>
</tr>
</tbody>
</table>

Details

The space shuttle Challenger exploded 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**

*Description*

Survey of college students to look at factors related to having overdrawn a checking account.

*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.

Renamed as CreditRisk in second edition.

*Source*


**Oysters**

*Description*

Comparing methods for measuring the size of oysters

*Format*

A data frame with 30 observations on the following 5 variables.

- **ID**: ID number of each oyster
- **Weight**: Weight (in grams)
- **Volume**: Volume (in cubic centimeters)
- **ThreeD**: Measurement from a 3D system (pixels)
- **TwoD**: Measurement from a 2D cross-section (pixels)
Details
In 2001 engineers at an R&D lab Agri-Tech, Inc, in Woodstock, Virginia, designed a 3-D system that they hoped would improve on the existing 2-D system for measuring the size of oysters. The 3-D system used computer scanning to estimate an oyster volume, whereas the old 2-D system estimated a cross-sectional area. Data shows the result of both systems, as well as the actual weight and volume of each oyster used in calibration.

Source

PalmBeach

Palm Beach Butterfly Ballot

Description
Votes for George 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.

<table>
<thead>
<tr>
<th>County</th>
<th>Buchanan</th>
<th>Bush</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of the Florida county</td>
<td>Number of votes for Pat Buchanan</td>
<td>Number of votes for George Bush</td>
</tr>
</tbody>
</table>

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% of the votes cast in Florida 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.
PeaceBridge2012

Source

Florida county data for the 2000 presidential election can be found at http://election.dos.state.fl.us/elections/resultsarchive/Index.asp?ElectionDate=11/7/00

---


Description

Monthly traffic (in 1,000’s) across the Peace Bridge between Canada and the U.S.

Format

A data frame with 156 observations on the following 4 variables.

- Year  Year (2003 to 2015)
- Month  Month (1 to 12)
- Traffic  Vehicles (in 1,000’s)
- t  Time frame (1 to 156)

Details

Monthly traffic (in thousands of vehicles) across the Peace Bridge between the U.S. and Canada near Niagara Falls between January 2003 and December 2015. Note PeaceBridge2012 has only the last four years of this series.

Source


---


Description

Monthly traffic (in 1,000’s) across the Peace Bridge between Canada and the U.S.

Format

A data frame with 48 observations on the following 4 variables.

- Year  Year (2012 to 2015)
- Month  Month (1 to 12)
- Traffic  Vehicles (in 1,000’s)
- t  Time frame (1 to 48)
Details


Source


---

**Pedometer**  
*Pedometer Walking Data*

**Description**

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

**Format**

A dataset with 68 observations on the following 8 variables.

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

**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 Sizes*
Description

Size of perch caught in a Finnish lake

Format

A dataset with 56 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Obs</th>
<th>Observation number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>Weight (in grams)</td>
</tr>
<tr>
<td>Length</td>
<td>Length (in centimeters)</td>
</tr>
<tr>
<td>Width</td>
<td>Width (in centimeters)</td>
</tr>
</tbody>
</table>

Details

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

Source


---

Description

Effects of additives to pig feed on weight gain

Format

A dataset with 12 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>WgtGain</th>
<th>Daily weight gain (hundredths of a pound over 1.00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibiotic</td>
<td>Antibiotic in the feed? No or Yes</td>
</tr>
<tr>
<td>B12</td>
<td>Vitamin B12 in the feed? No or Yes</td>
</tr>
</tbody>
</table>

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.

Description

Data from pine seedlings planted in 1990

Format

A dataset with 1000 observations on the following 15 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>Row number in pine plantation</td>
</tr>
<tr>
<td>Col</td>
<td>Column number in pine plantation</td>
</tr>
<tr>
<td>Hgt90</td>
<td>Tree height at time of planting (cm)</td>
</tr>
<tr>
<td>Hgt96</td>
<td>Tree height in September 1996 (cm)</td>
</tr>
<tr>
<td>Diam96</td>
<td>Tree trunk diameter in September 1996 (cm)</td>
</tr>
<tr>
<td>Grow96</td>
<td>Leader growth during 1996 (cm)</td>
</tr>
<tr>
<td>Hgt97</td>
<td>Tree height in September 1997 (cm)</td>
</tr>
<tr>
<td>Diam97</td>
<td>Tree trunk diameter in September 1997 (cm)</td>
</tr>
<tr>
<td>Spread97</td>
<td>Widest lateral spread in September 1997 (cm)</td>
</tr>
<tr>
<td>Needles97</td>
<td>Needle length in September 1997 (mm)</td>
</tr>
<tr>
<td>Deer95</td>
<td>Type of deer damage in September 1995: 0 = none, 1 = browsed</td>
</tr>
<tr>
<td>Deer97</td>
<td>Type of deer damage in September 1997: 0 = none, 1 = browsed</td>
</tr>
<tr>
<td>Cover95</td>
<td>Thorny cover in September 1995: 0 = none; 1 = some; 2 = moderate; 3 = lots</td>
</tr>
<tr>
<td>Fert</td>
<td>Indicator for fertilizer: 0 = no, 1 = yes</td>
</tr>
<tr>
<td>Spacing</td>
<td>Distance (in feet) between trees (10 or 15)</td>
</tr>
</tbody>
</table>

Details

This dataset contains 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.

---

**PKU**

*Dopamine levels with PKU in diets*

---

**Description**

Dopamine levels with different amounts of phenylalanine in diets

**Format**

A data frame with 20 observations on the following 4 variables.

- **Subject** Initials to identify each subject
- **Diet** Level of phenylalanine in diet (Low or Normal)
- **DietControl** Ability to follow prescribed diet (Good or Poor)
- **Y** Concentration of dopamine (micrograms per milligram of creatinine)

**Details**

Phenylketonuria (PKU) is an enzyme deficiency that keeps a person from being able to synthesize enough dopamine. The amino acid phenylalanine inhibits the enzyme needed to synthesize dopamine, and so to some extent, a diet low in phenylalanine can moderate the symptoms of PKU. In short, less phenylalanine in the diet should lead to more dopamine in the brain. The dopamine level for each patient was measured after a normal diet and after a week on a low phenylalanine diet.

**Source**


Several of the values were altered slightly in ways that would not change the analysis except to simplify the arithmetic.

---

**Political**

*Political Behavior of College Students*

---

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

---

<table>
<thead>
<tr>
<th>Pollster08</th>
<th>2008 U.S. Presidential Election Polls</th>
</tr>
</thead>
</table>

**Description**

Polls for 2008 U.S. presidential election

**Format**

A dataset with 102 observations on the following 11 variables.

<table>
<thead>
<tr>
<th>PollTaker</th>
<th>Polling organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>PollDates</td>
<td>Dates the poll data were collected</td>
</tr>
<tr>
<td>MidDate</td>
<td>Midpoint of the polling period</td>
</tr>
<tr>
<td>Days</td>
<td>Number of days after August 28th (end of Democratic convention)</td>
</tr>
<tr>
<td>n</td>
<td>Sample size for the poll</td>
</tr>
<tr>
<td>Pop</td>
<td>(A=\text{all, } LV=\text{likely voters, } RV=\text{registered voters})</td>
</tr>
<tr>
<td>McCain</td>
<td>Percent supporting John McCain</td>
</tr>
<tr>
<td>Obama</td>
<td>Percent supporting Barack Obama</td>
</tr>
<tr>
<td>Margin</td>
<td>Obama percent minus McCain percent</td>
</tr>
<tr>
<td>Charlie</td>
<td>Indicator for polls after Charlie Gibson interview with VP candidate Sarah Palin (9/11)</td>
</tr>
<tr>
<td>Meltdown</td>
<td>Indicator for polls after Lehman Brothers bankruptcy (9/15)</td>
</tr>
</tbody>
</table>
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 McCain 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 Popping Success**

#### 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

**Porsche and Jaguar Prices**

#### Description

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

**Format**

A dataset with 60 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Car</th>
<th>Price</th>
<th>Age</th>
<th>Mileage</th>
<th>Porsche</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car model: Jaguar or Porsche</td>
<td>Price (in $1,000's)</td>
<td>Age of the car (in years)</td>
<td>Previous miles driven (in 1,000’s)</td>
<td>Indicator for Porsche (1) or Jaguar (0)</td>
</tr>
</tbody>
</table>

**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.

---

**PorschePrices**

**Description**

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

**Format**

A dataset with 30 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>Price</th>
<th>Age</th>
<th>Mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asking price for the car (in $1,000’s)</td>
<td>Age of the car (in years)</td>
<td>Previous miles driven (in 1,000’s)</td>
</tr>
</tbody>
</table>

**Details**

A student was interested in prices 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). This dataset was replaced by AccordPrice for second edition.

**Source**

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

**Pulse Rates and Exercise**

**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
- **Sex**: 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

**Putting Success by Length (Long Form)**

**Description**

Putting results for a golfing statistician

**Format**

A dataset with 587 observations on the following 2 variables.

- **Length**: Length of 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 puts. 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

Putting Success by Length (Short Form)

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 of 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

Putts3

Hypothetical Putting Data (Short Form)

Description

Hypothetical putting results for a golfing statistician
Format

A data frame 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

This is a hypothetical revision of the table of putting success in Putts2 that helps demonstrate overdispersion.

Source

Modified from personal observations by one of the Stat2 authors.

---

RacialAnimus  Racial Animus and City Demographics

Description

Demographics and a measurement of racial animus in cities based on Google searches

Format

A data frame with 196 observations on the following 7 variables.

MediaMarket  City (State)
Age65Plus   Percentage 65 and older
BachPlus    Percentage with a bachelor’s degree
Black       Percentage of African-Americans
Hispanic    Percentage of Hispanics
ObamaKerry  Percentage of vote won by Obama in 2008 minus Kerry percentage in 2004
Animus      Measurement (0-250) of racial animus

Details

Professor Seth Stephens-Davidowitz studies the level of racial animus across different areas in America by measuring the percent of Google search queries that include racially charged language. A measurement, Animus, is derived from his algorithm and is scaled to be between 0 (low racial animus) and 250 (high racial animus). The dataset includes those values along with demographic information about each media market.
Radioactive Twins

Source


Radioactive Twins  Comparing Twins Ability to Clear Radioactive Particles

Description

Experiment comparing twins (one urban, one rural) ability to clear airborne radioactive particles from their lungs

Format

A data frame with 30 observations on the following 3 variables.

TwinPair  Identifies the twin pairs (1 to 15)

Env  Residential environment (Rural or Urban)

Rate  Clearance rate (percentage radioactive particles remaining after one hour)

Details

To assess lung health, the scientists measured "tracheobronchial clearance rate," that is, in English, "How fast do your lungs get rid of nasty stuff?" Each subject agreed to inhale an aerosol of radioactive Teflon particles. A Geiger counter held to the chest measured the radioactivity just after inhaling, and again one hour later. The clearance rate was the percentage of radioactivity remaining – the lower the better. Subjects were 15 sets of identical twins, each pair with one twin living in an urban environment and the other in a rural environment.

Source

Description

Sample of homes in Northampton, MA to see whether being close to a bike trail enhances the value of the home.

Format

A data frame with 104 observations on the following 30 variables:

- **HouseNum**: Unique house number
- **Acre**: Lot size for the house (in acres)
- **AcreGroup**: Lot size groups (<= 1/4 acre or > 1/4 acre)
- **Adj1998**: Estimated 1998 price (in thousands of 2014 dollars)
- **Adj2007**: Estimated 2007 price (in thousands of 2014 dollars)
- **Adj2011**: Estimated 2011 price (in thousands of 2014 dollars)
- **BedGroup**: Bedroom groups (1-2 beds, 3 beds, or 4+ beds)
- **Bedrooms**: Number of bedrooms
- **BikeScore**: Bike friendliness (0-100 score, higher scores are better)
- **Diff2014**: Difference in price between 2014 estimate and adjusted 1998 estimate (in thousands of dollars)
- **Distance**: Distance (in feet) to the nearest entry point to the rail trail network
- **DistGroup**: Distance groups, compared to 1/2 mile (Closer or Farther Away)
- **GarageSpaces**: Number of garage spaces (0-4)
- **GarageGroup**: Any garage spaces? (no or yes)
- **Latitude**: Latitude (for mapping)
- **Longitude**: Longitude (for mapping)
- **NumFullBaths**: Number of full baths (includes shower or bathtub)
- **NumHalfBaths**: Number of half baths (no shower or bathtub)
- **NumRooms**: Number of rooms
- **PctChange**: Percentage change from adjusted 1998 price to 2014 (value of zero means no change)
- **Price1998**: Zillow 10 year estimate from 2008 (in thousands of dollars)
- **Price2007**: Zillow price estimate from 2007 (in thousands of dollars)
- **Price2011**: Zillow price estimate from 2011 (in thousands of dollars)
- **Price2014**: Zillow price estimate from 2014 (in thousands of dollars)
- **SFGroup**: SquareFeet group (<= 1500 sf or > 1500 sf)
- **SquareFeet**: Square footage of interior finished space (in thousands of sf)
- **StreetName**: Street name
- **StreetNum**: House number on street
- **WalkScore**: Walk friendliness (0-100 score, higher scores are better)
- **Zip**: Location (1060 = Northampton or 1062 = Florence)
Details

This dataset comprises 104 homes in Northampton, MA that were sold in 2007. The authors measured the shortest distance from each home to a railtrail on streets and pathways with Google maps and recorded the Zillow.com estimate of each home’s price in 1998 and 2011. Additional attributes such as square footage, number of bedrooms and number of bathrooms are available from a realty database from 2007. We divide the houses into two groups based on distance to the trail (DistGroup).

Source

From July 2015 JSE Datasets and Stories: "Rail Trails and Property Values: Is There an Association?", Ella Hartenian, Smith College and Nicholas J. Horton, Amherst College.


---

<table>
<thead>
<tr>
<th>Rectangles</th>
<th>Measurements of Rectangles</th>
</tr>
</thead>
</table>

Description

Measurements for a hypothetical set of nine rectangles.

Format

A data frame with 9 observations on the following 5 variables.

- **Case** ID number for each rectangle
- **Width** Width (1, 4, or 10)
- **Length** Length (1, 4, or 10)
- **Area** Area
- **logArea** Log (base 10) of area

Details

Areas for rectangles of width 1, 4, or 10 and length of 1, 4, or 10.

Source

Areas computed for a hypothetical set of rectangles.
ReligionGDP

**Religion and GDP for Countries**

---

**Description**

Data on religiosity of countries from the Pew Global Attitudes Project

**Format**

A dataset with 44 observations on the following 9 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>Name of country</td>
</tr>
<tr>
<td>Religiosity</td>
<td>A measure of degree of religiosity for residents of the country</td>
</tr>
<tr>
<td>GDP</td>
<td>Per capita Gross Domestic Product in the country</td>
</tr>
<tr>
<td>Africa</td>
<td>Indicator for countries in Africa</td>
</tr>
<tr>
<td>EastEurope</td>
<td>Indicator for countries in Eastern Europe</td>
</tr>
<tr>
<td>MiddleEast</td>
<td>Indicator for countries in the Middle East</td>
</tr>
<tr>
<td>Asia</td>
<td>Indicator for countries in Asia</td>
</tr>
<tr>
<td>WestEurope</td>
<td>Indicator for countries in Western Europe</td>
</tr>
<tr>
<td>Americas</td>
<td>Indicator for countries in North/South America</td>
</tr>
</tbody>
</table>

**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.

---

RepeatedPulse

**Pulse Rates at Various Times of Day**

---

**Description**

A student measured her pulse several times a day over 26 days.
### ResidualOil

**Format**

A data frame with 104 observations on the following 3 variables.

- **Pulse**: Pulse rate (beats per minute)
- **Time**: Time of day (evening, morning, noon, one)
- **Day**: Day 1 to Day 26

**Details**

A student measured her pulse in the morning, at noon, at 1:00, and in the evening for each of 26 days.

**Source**

Data supplied by a student at Oberlin College.

### US Residual Oil Production (Quarterly 1983-2016)

**Description**

Quarterly production of residual oil in the U.S. from 1983 to 2016

**Format**

A data frame with 136 observations on the following 7 variables.

- **Year**: Year (1983 to 2016)
- **Qtr**: Month (1=Jan-Mar, 2=Apr-June, 3=July-Sep, 4=Oct-Dec)
- **t**: Time index (1 to 136)
- **Oil**: Residual fuel oil distribution (in million gallons/day)
- **LogOil**: Natural logarithm of Oil

**Details**

The U.S. Energy Information Administration tracks the production and distribution of various types of petroleum products. The category for this dataset is called residual oil, which are heavier oils (often called No. 5. and No. 6) that remain after lighter oils (such as No. 4 home heating oil) are distilled away in the refining process. It is used in steam-powered ships, power plants, and other industrial applications.

**Source**

U.S. Energy Information Administration website - Refiner sales volumes for residual fuel oil and No. 4 heating oil at https://www.eia.gov/petroleum/data.php#consumption. Specific webpage is https://www.eia.gov/dnav/pet/pet_cons_refres_d_nus_VTR_mgalpd_m.htm.
Ricci  
Retirement  
Yearly Contributions to a Supplemental Retirement Account

Description
Contributions to a supplemental retirement account (1997-2012)

Format
A dataset with 16 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Year</th>
<th>1997-2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRA</td>
<td>Annual contribution to the Supplemental Retirement Account</td>
</tr>
</tbody>
</table>

Details
A faculty member opened a supplemental retirement account (SRA) in 1997 to invest 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.

Ricci  
Firefighter Promotion Exam Scores

Description
Data on firefighter promotion exams as part of the Ricci v. DeStafano court case

Format
A data frame with 118 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Race</th>
<th>Race of firefighter (B=black, H=Hispanic, or W=white)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Promotion desired (Captain or Lieutenant)</td>
</tr>
<tr>
<td>Oral</td>
<td>Oral exam score</td>
</tr>
<tr>
<td>Written</td>
<td>Written exam score</td>
</tr>
<tr>
<td>Combine</td>
<td>Combined score (written exam gets 60% weight)</td>
</tr>
</tbody>
</table>
Details

The city of New Haven, Connecticut administered exams (both written and oral) in November and December of 2003 to firefighters hoping to qualify for promotion to either Lieutenant or Captain in the city fire department. A final score consisting of a 60% weight for the written exam and a 40% weight for the oral exam was computed for each person who took the exam. For each person who took the exams, there are measurements on their race (black, white, or Hispanic), which position they were trying for (Lieutenant, Captain), scores on the oral and written exams, and the combined score. These data were used as part of a court case (Ricci v. DeStefano) dealing with racial discrimination

Source

Data (RicciData.csv) and documentation (Ricci.txt) downloaded from http://www.amstat.org/publications/jse/jse_data_archive.htm


RiverElements

Elements in River Water Samples

Description

Concentrations of elements in river water samples from upstate NY

Format

A dataset with 12 observations on the following 27 variables.

<table>
<thead>
<tr>
<th>River</th>
<th>One of four rivers: Grasse, Oswegatchie, Raquette, or St. Regis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>Location: 1=UpStream, 2=MidStream, 3=Downstream</td>
</tr>
<tr>
<td></td>
<td>Al Aluminum</td>
</tr>
<tr>
<td></td>
<td>Ba Barium</td>
</tr>
<tr>
<td></td>
<td>Br Bromine</td>
</tr>
<tr>
<td></td>
<td>Ca Calcium</td>
</tr>
<tr>
<td></td>
<td>Ce Cerium</td>
</tr>
<tr>
<td></td>
<td>Cu Copper</td>
</tr>
<tr>
<td></td>
<td>Dy Dysprosium</td>
</tr>
<tr>
<td></td>
<td>Er Erbium</td>
</tr>
<tr>
<td></td>
<td>Fe Iron</td>
</tr>
<tr>
<td></td>
<td>Gd Gadolinium</td>
</tr>
<tr>
<td></td>
<td>Ho Holmum</td>
</tr>
<tr>
<td></td>
<td>K Potassium</td>
</tr>
<tr>
<td></td>
<td>La Lathanum</td>
</tr>
<tr>
<td></td>
<td>Li Lithium</td>
</tr>
<tr>
<td></td>
<td>Mg Magnesium</td>
</tr>
</tbody>
</table>
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 concentrations (parts per million) of a variety of elements in those water samples. The dataset RiverIron contains the information for iron (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

<table>
<thead>
<tr>
<th>River</th>
<th>Site</th>
<th>Iron</th>
<th>LogIron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grasse</td>
<td>Downstream</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oswegatchie</td>
<td>Midstream</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raquette</td>
<td>Upstream</td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. Regis</td>
<td>Downstream</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 concentrations 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  Field Goal Attempts in the NFL

Description

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

Format

A dataset with 30 observations on the following 13 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>ID number</td>
</tr>
<tr>
<td>PlayerID</td>
<td>Code for player</td>
</tr>
<tr>
<td>LastName</td>
<td>Last name</td>
</tr>
<tr>
<td>FirstName</td>
<td>First name</td>
</tr>
<tr>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>Team</td>
<td>Abbreviation for team name</td>
</tr>
<tr>
<td>Date</td>
<td>Code for date: mmddyy</td>
</tr>
<tr>
<td>FGAttempts</td>
<td>Field goals attempted by the kicker that game</td>
</tr>
<tr>
<td>FGMade</td>
<td>Field goals made by the kicker that game</td>
</tr>
<tr>
<td>Attempt</td>
<td>Which attempt during the game?</td>
</tr>
<tr>
<td>Result</td>
<td>1=made the field goal or 0=missed</td>
</tr>
<tr>
<td>Yards</td>
<td>Number of yards for the field goal attempt</td>
</tr>
<tr>
<td>Block</td>
<td>1=attempt blocked or 0=not blocked</td>
</tr>
</tbody>
</table>
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

Description

Ant counts on samples of different kinds of sandwiches

Format

A dataset with 48 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial</td>
<td>Trial number</td>
</tr>
<tr>
<td>Bread</td>
<td>Type of bread: Multigrain, Rye, White, or Wholemeal</td>
</tr>
<tr>
<td>Filling</td>
<td>Type of filling: HamPickles, PeanutButter, or Vegemite</td>
</tr>
<tr>
<td>Butter</td>
<td>Butter on the sandwich? no or yes</td>
</tr>
<tr>
<td>Ants</td>
<td>Number of ants on the sandwich</td>
</tr>
</tbody>
</table>

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 counted 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

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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MathSAT</td>
<td>Score (out of 800) on the mathematics portion of</td>
</tr>
<tr>
<td></td>
<td>the SAT exam</td>
</tr>
<tr>
<td>VerbalSAT</td>
<td>Score (out of 800) on the verbal portion of the</td>
</tr>
<tr>
<td></td>
<td>SAT exam</td>
</tr>
<tr>
<td>GPA</td>
<td>Grade point average (0.0-4.0 scale)</td>
</tr>
</tbody>
</table>

**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 and Math portions 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.

---

SeaIce  Arctic Sea Ice (1979-2015)

**Description**

Area of sea ice in the Arctic measured yearly in September (1979 to 2015)

**Format**

A data frame with 37 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Year (1979 - 2015)</td>
</tr>
<tr>
<td>Extent</td>
<td>Extent of arctic sea ice (in million square km)</td>
</tr>
<tr>
<td>Area</td>
<td>Area of arctic sea ice (in million square km)</td>
</tr>
<tr>
<td>t</td>
<td>Index for year (t=1 in 1979)</td>
</tr>
</tbody>
</table>
Details

Climatologists have been measuring the amount of sea ice in both the Arctic and Antarctic regions for a number of years. This datafile gives information about the amount of sea ice in the arctic region as measured in September (the time when the amount of ice is at its least) since 1979. The basic research question is to see if we can use time to model the amount of sea ice.

In fact, there are two ways to measure the amount of sea ice: Area and Extent. Area measures the actual amount of space taken up by ice. Extent measures the area inside the outer boundaries created by the ice. If there are areas inside the outer boundaries that are not ice (think about a slice of swiss cheese), then the Extent will be a larger number than the Area. In fact, this is almost always true.

Source

Data from ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/Sep/N_09_areaV2.txt updated data from


<table>
<thead>
<tr>
<th>SeaSlugs</th>
<th>Sea Slug Larvae</th>
</tr>
</thead>
</table>

Description

Metamorphose rates for sea slugs exposed to different water samples

Format

A dataset with 36 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Time</th>
<th>Minutes after tide come in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>Proportion of 15 sea slug larvae that metamorphose</td>
</tr>
</tbody>
</table>

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).
SleepingShrews

Source


References


---

SleepingShrews  Shrew Heart Rates at Stages of Sleep

Description

Heart rates for a sample of six tree shrews at each of three stages of sleep.

Format

A data frame with 18 observations on the following 4 variables.

- ID  Row ID
- Shrew  Shrew ID (A through F)
- Phase  Phase of sleep (DSW=deep wave, LSW=light wave, or REM=dreaming)
- Rate  Heart rate (beats per minute)

Details

Heart rates were recorded for a sample of six tree shrews at each of three stages of sleep.

Source

sluacf

Computes autocorrelations (ACF) for a time series

Description
This function computes autocorrelations for various lags of a time series.

Usage
`sluacf(series, lags = 1, maxlag = NULL, ndiff = 0, sdiff = 0)`

Arguments
- `series`: a time series object
- `lags`: a multiplier for the lag. For example, use lag=12 for monthly data.
- `maxlag`: maximum number of lags to compute
- `ndiff`: number of regular differences to take before finding the ACF
- `sdiff`: number of seasonal differences (with seasonal period specified by lags)

Details
This is a wrapper for the `acf` function which allows for specifying regular (ndiff) and seasonal (sdiff) differences. The lags parameter specifies the seasonal lag and adjusts the default lags in the returned acf object to go 1, 2, ..., rather than showing fractional lags (for example, 1/12, 2/12, ... for monthly data). The lag 0 autocorrelation is set to NA (rather than 1) so that it won’t show when acf is plotted.

Value
An object of class "acf"

Examples
```r
data(SeaIce)
ExtentY=ts(SeaIce$Extent,start=1979)
sluacf(ExtentY)
sluacf(ExtentY, maxlag=8, ndiff=1)

data(Inflation)
CPIts=ts(Inflation$CPI,start=c(2009,1),frequency=12)
PILacf=sluacf(CPIts, maxlag=36, lags=12)
plot(PILacf, lwd=2, ci.type="ma", xlab=c(1,36), xaxp=c(0,36,6), main="")
```
### Sparrows

**Sparrow Measurements**

**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

**Land Area and Mammal Species**

**Description**

Land area and number of mammal species for islands 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

<table>
<thead>
<tr>
<th>Speed</th>
<th>Highway Fatality Rates (Yearly)</th>
</tr>
</thead>
</table>

Description
Highway fatality rates 1987-2007

Format
A dataset with 21 observations on the following 3 variables.

Year | Year (1987-2007)
FatalityRate | Number of fatalities 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 Safety Administration website at http://www-fars.nhtsa.dot.gov/Main/index.aspx

SugarEthanol | Effects of Oxygen on Sugar Metabolism

Description
Experiment on the effects of oxygen on sugar metabolism by bacteria
**Format**

A data frame with 16 observations on the following 3 variables.

- **Sugar**  Type of sugar (Galactose or Glucose)
- **Oxygen** 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**


The original article is Yamada T., Takahashi-Abbe S., Abbe K. (1985) "Effects of oxygen concentration on pyruvate formatelyase in situ and sugar metabolism of Streptococcus mutans and Streptococcus sanguis," Infection and Immunity, pp. 129-134.

---

**Description**

Data on serious suicide attempts in Shandong, China

**Format**

A data frame with 2571 observations on the following 11 variables.

- **Person_ID**  ID number
- **Hospitalised** Hospitalised? (no or yes)
- **Died** Died? (no or yes)
- **Urban** Urban area? (no, unknown, or yes)
- **Year**  Year (2009, 2010, or 2011)
- **Month** Month (1=Jan through 12=December)
- **Sex**  Sex (female or male)
- **Age**  Age (years)
- **Education** Education level (iliterate, primary, Secondary, Tertiary, or unknown)
- **Occupation** One of ten occupation categories
- **method** One of nine possible methods
Details

Data from a study of serious suicide attempts over three years in a predominantly rural population in Shandong, China.

Source

Data downloaded via Dryad Digital Repository. https://doi.org/10.5061/dryad.r0v35

Swahili

Attitudes Towards Swahili in Kenyan Schools

Description

Attitudes towards the Swahili language among Kenyan school children

Format

A dataset with 480 observations on the following 4 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Province</td>
<td>NAIROBI or PWANI</td>
</tr>
<tr>
<td>Sex</td>
<td>female or male</td>
</tr>
<tr>
<td>Attitude.Score</td>
<td>Score (out a possible 200 points) on a survey of attitude towards the Swahili language</td>
</tr>
<tr>
<td>School</td>
<td>Code for the school: A through L</td>
</tr>
</tbody>
</table>

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.
**Tadpoles**  
**Effects of a Fungus on Tadpoles**

**Description**
Comparing intestine lengths for tadpoles with and without exposure to Bd fungus

**Format**
A data frame with 27 observations on the following 4 variables.

- **Treatment**  Exposed to fungus (Bd=yes or Control=no)
- **Body**  Length of body (in mm)
- **GutLength**  Length of intestine (in mm)
- **MouthpartDamage**  Measure of damage to the mouth (e.g. missing teeth)

**Details**
Biologists wondered whether tadpoles can adjust the relative length of their intestines if they are exposed to a fungus called Batrachochytrium dendrobatidis (Bd).

**Source**

---

**TechStocks**  
**Daily Prices of Three Tech Stocks**

**Description**
Daily closing prices of Apple, Google, and Microsoft stocks (12/1/2015 to 12/1/2017)

**Format**
A data frame with 504 observations on the following 5 variables.

- **Date**  Date (coded as mm/dd/yyyy)
- **AAPL**  Apple Inc. closing price
- **GOOG**  Alphabet Inc. (Google) closing price
- **MSFT**  Microsoft Corp. closing price
- **t**  Time index (1 to 505)
Details

Closing price of Apple (AAPL), Google/Alphabet (GOOG) and Microsoft (MSFT) stocks for each trading day in a two-year period from 12/1/2015 to 12/1/2017.

Source

Data downloaded using the Quandl R package (12/2/2017)

---

<table>
<thead>
<tr>
<th>TeenPregnancy</th>
<th>State Teen Pregnancy Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>CivilWar, Church, Teen</td>
</tr>
</tbody>
</table>

Description

State teen pregnancy rates, Civil War participation, and church attendance.

Format

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

State State abbreviation
CivilWar Role in Civil War (B=border, C=Confederate, O=other, or U=union)
Church Percentage who attended church in previous week (from a state survey)
Teen Number of pregnancies per 1000 teenage girls in state

Details

State level data on teen pregnancies, church attendance, and role in the U.S. Civil War.

Source

2010 teen pregnancy rate, per 1000 teenage women, per year. Source: Guttmacher Institute, via Tanya Lewis (5 May 2014) "Teen pregnancy rates by state," https://www.livescience.com

---

<table>
<thead>
<tr>
<th>TextPrices</th>
<th>Textbook Prices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>Price</td>
</tr>
</tbody>
</table>

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

---

US Senate Votes on Clarence Thomas Confirmation

Description

Votes in the US Senate on Clarence Thomas nomination for the US Supreme Court

Format

A data frame with 100 observations on the following 6 variables.

State  State name
Senator  Senator name
Party  Party affiliation (D=Democrat or R=Republican)
ConfVote  Confirmation vote (Nay or Yea)
StateOpinion  Percentage of state residents supporting the choice
Vote  Numeric coding for vote (1=for or 0=against)

Details

Data from the U.S. Senate vote on October 15, 1991 to confirm Clarence Thomas to a position on the Supreme Court.

Source

These numbers are taken from Kastellec, J.P., Lax, J.R., and Phillips, J. (2010), "Public Opinion and Senate Confirmation of Supreme Court Nominees," Journal of Politics, 72(3): 767-84. In this paper the authors used opinion polls and an advanced statistical method known as multilevel regression and poststratification to determine the StateOpinion levels.
ThreeCars2017

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.

<table>
<thead>
<tr>
<th>CarType</th>
<th>BMW, Jaguar, or Porsche</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>Asking price (in $1,000’s)</td>
</tr>
<tr>
<td>Age</td>
<td>Age of the car (in years)</td>
</tr>
<tr>
<td>Mileage</td>
<td>previous miles driven (in 1,000’s)</td>
</tr>
<tr>
<td>Car</td>
<td>0=Porsche, 1=Jaguar, 2=BMW</td>
</tr>
<tr>
<td>Porsche</td>
<td>Indicator with 1= Porsche and 0=otherwise</td>
</tr>
<tr>
<td>Jaguar</td>
<td>Indicator with 1= Jaguar and 0=otherwise</td>
</tr>
<tr>
<td>BMW</td>
<td>Indicator with 1= BMW and 0=otherwise</td>
</tr>
</tbody>
</table>

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 (from the first edition) has only the Porsche data and the PorscheJaguar dataset has the data for those two models.

This dataset has been updated (with different car models) to ThreeCars2017 for the second edition.

Source

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

ThreeCars2017

Description

Data from cars.com for a sample of three different models of used cars in 2017

Format

A data frame with 90 observations on the following 7 variables.

| CarType   | Model (Accord, Maxima, or Mazda6) |
**TipJoke**

Age  Age of used car (years)
Price  Price (in thousands of dollars)
Mileage  Mileage (in thousands of miles)
Mazda6  Is the car a Mazda6? (1=yes or 0=no)
Accord  Is the car an Accord? (1=yes or 0=no)
Maxima  Is the car a Maxima? (1=yes or 0=no)

**Details**

Data for a sample of cars from three models (Mazda6, Honda Accord, Toyota Maxima) from a website. The dataset AccordPrice is a subset of this file.

**Source**

Data obtained from cars.com, February 2017 using zip code 44107, Lakewood, Ohio.

---

**TipJoke**  

*Improve Chances of Getting a Tip?*

---

**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 (1=ad card left or 0=no ad card)
- Joke  Indicator for Joke card (1=joke card left or 0=no joke card)
- None  Indicator for no card (1=no card left or 0=ad or joke card left)

**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


Titanic

Passengers on the Titanic

Description

List and outcomes for passengers on the Titanic

Format

A dataset with 1313 observations on the following 6 variables.

<table>
<thead>
<tr>
<th>Name</th>
<th>Passenger name</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClass</td>
<td>Passenger class: *missing, 1st, 2nd, or 3rd</td>
</tr>
<tr>
<td>Age</td>
<td>Age (in years)</td>
</tr>
<tr>
<td>Sex</td>
<td>female or male</td>
</tr>
<tr>
<td>Survived</td>
<td>1=survived or 0=died</td>
</tr>
<tr>
<td>SexCode</td>
<td>1=female or 0=male</td>
</tr>
</tbody>
</table>

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


TMS

Migraines and TMS

Description

Effects of transcranial magnetic stimulation (TMS) on migraine headaches
Format

A dataset with 2 observations on the following 4 variables.
<table>
<thead>
<tr>
<th>Group</th>
<th>Treatment group: Placebo or TMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Count of number of patients that were pain-free</td>
</tr>
<tr>
<td>No</td>
<td>Count of number of patients that had pain</td>
</tr>
<tr>
<td>Trials</td>
<td>Number of patients in the group</td>
</tr>
</tbody>
</table>

**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. This dataset renamed as Migraines in second edition.

**Source**


---

**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.

<table>
<thead>
<tr>
<th>Game</th>
<th>Week number in the 2006 season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opponent</td>
<td>Name of opposing team</td>
</tr>
<tr>
<td>Attempts</td>
<td>Number of rushing attempts</td>
</tr>
<tr>
<td>Yards</td>
<td>Total yards gained rushing for the game</td>
</tr>
</tbody>
</table>

**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.

This data set from the first edition was replaced by BreesPass in the second edition.
TukeyNonaddPlot

Source

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

Description

This function produces a Tukey nonadditivity plot for a two-way ANOVA model.

Usage

TukeyNonaddPlot(formula, data, out = "n", main = "Tukey Nonadditivity Plot", ylab = "Residuals")

Arguments

formula A formula for a two-way ANOVA in the form Response=FactorA+FactorB (or FactorA*FactorB)
data A dataframe
out Control what is returned. Default is "n"=nothing. Other options are "comp" for the comparisons, "line" for the equation of the line, and "resid" for the cell residuals.
main Add a title, default is "Tukey Nonadditivity Plot"
ylab Label vertical axis, default is "Residuals"

Details

More details need to be written

Value

Depends on the option set with out.

Examples

data(Dinosaurs)
TukeyNonaddPlot(Iridium~Source*factor(Depth),data=Dinosaurs)
**TwinsLungs**  
*Comparing Twins Ability to Clear Radioactive Particles*

**Description**

Experiment comparing twins (one urban, one rural) ability to clear airborne radioactive particles from their lungs.

**Format**

A dataset with 14 observations on the following 3 variables.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Code for the twin pair: A - G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environ</td>
<td>Living environment: Rural or Urban</td>
</tr>
<tr>
<td>Percent</td>
<td>Percentage of radioactivity remaining in lungs</td>
</tr>
</tbody>
</table>

**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.

This dataset was renamed as RadioactiveTwins for the second edition.

**Source**


---

**Undoing**  
*Defense of Undoing OCD Symptoms in Psychotherapy*

**Description**

Ratings of an OCD symptom in psychotherapy sessions.
USstamps

Format

A data frame with 44 observations on the following 3 variables.

Group  Time frame of the session (I=early through VI=late)
Score  Rating of OCD symptom on a 1 to 4 scale
Symbol  Indicator for groups I, III, and IV

Details

A patient had been diagnosed with OCD (obsessive/compulsive disorder) and underwent a series of psychotherapy sessions. Notes from the sessions were presented to three different experienced therapists who rated sessions with a particular OCD symptom (defense of undoing) on a 1 to 4 scale (smaller values indicating worse symptoms). If all three judges agreed on the stage of a session, that determined the category. Otherwise, they discussed until they reached a consensus. The sessions were also grouped into six groups with I being the earliest sessions and VI being the latest.

Source


<table>
<thead>
<tr>
<th>USstamps</th>
<th>Price of US Stamps</th>
</tr>
</thead>
</table>

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

### VisualVerbal

**Visual versus Verbal Performance**

**Description**

Experiment to compare visual and verbal performance

**Format**

A data frame with 80 observations on the following 5 variables.

- **Subject**  Subject number (s1 to s20)
- **Task**  Follow a letter (Visual) or a sentence (Verbal)
- **Report**  Point response (Visual) or say response (Verbal)
- **Group**  Combination of Task+Report (Letter Point, Letter Say, Sentence Point, or Sentence Say)
- **Time**  Response time (in seconds)

**Details**

Subjects carried out two kinds of tasks, one visual (identify letters), one verbal (identify sentences); and to report the results in either of two ways, one visual (pointing at a response), one verbal (speaking a response). Time to complete each task was recorded in seconds.

**Source**


---

### Volts

**Voltage Drop for a Discharging Capacitor**

**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)
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 0.02 seconds.

Measurements recorded by one of the authors.

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

A dataset with 24 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Group</th>
<th>Treatments: exercise control, final report, special exercises, or weekly report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age (in months) when first walking</td>
</tr>
</tbody>
</table>

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.

WalkTheDogs  

Did the Author Walk the Dogs Today?

Description

Daily pedometer data for one of the authors

Format

A data frame with 223 observations on the following 7 variables.

- StepCount: Number of steps taken in the day
- Kcal: Calories burned (according to pedometer)
- Miles: Miles walked
- Weather: cold, rain, or shine
- Day: Day of week (F=Friday, M=Monday, R=Thursday, S=Saturday, T=Tuesday, U=Sunday, W=Wednesday)
- Walk: Were the dogs walked? (1=yes or 0=no)
- Steps: Steps in units of 1,000 (so StepCount/1000)

Details

One of the authors recorded daily pedometer data, the weather, and whether or not he walked the dogs.

Source

One of the author’s pedometer records.

WeightLossIncentive  

Do Financial Incentives Improve Weight Loss?

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


WeightLossIncentive4  
Do Financial Incentives Improve Weight Loss? (4 Months)

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


---

WeightLossIncentive7  Do Financial Incentives Improve Weight Loss? (7 Months)

Description

Weight loss after seven months with/without a financial incentive

Format

A dataset with 33 observations on the following 2 variables.

<table>
<thead>
<tr>
<th>Group</th>
<th>Treatment group: Control or Incentive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month7Loss</td>
<td>Weight loss (in pounds) after seven months</td>
</tr>
</tbody>
</table>

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


---

Whickham2  Whickham Health Study

Description

Morality data over 20 years for 1314 women from Whickham, England
**Format**

A data frame with 1314 observations on the following 5 variables.

- **Outcome** Status at 20-year follow-up (Alive or Dead)
- **Smoker** Smoker at baseline? (No or Yes)
- **Age** Age (in years at baseline)
- **AgeGroup** Age group (18–64 or 65+)
- **Alive** Numeric code for Outcome (1=alive or 0=dead)

**Details**

 Twenty-year mortality, smoking status, and age for 1314 women in Whickham, England. We have named this Whickham2 to distinguish it from Whickham, which is a file in the mosaicData package.

**Source**

A version of these data are in the mosaicData package but originally are from:


---

**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).
This dataset appears in the first edition, but is not used in the second edition.

Source

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

<table>
<thead>
<tr>
<th>WordsWithFriends</th>
<th>Words with Friends Scores</th>
</tr>
</thead>
</table>

Description

Results from the online game Words with Friends (solo play)

Format

A data frame with 444 observations on the following 11 variables.

- Points: Number of points scored by the author
- OppPoints: Number of points scored by opponent ("solo")
- WinMargin: Points minus OppPoints, so margin of victory (or loss)
- Start: Did the author go first or pass? (first or pass)
- Ss: Number of S tiles (0 to 5)
- BlanksNumber: Number of Blank tiles (0 to 2)
- J: Did the author get the J tile? (1=yes, 0=no)
- Q: Did the author get the Q tile? (1=yes, 0=no)
- X: Did the author get the X tile? (1=yes, 0=no)
- Z: Did the author get the Z tile? (1=yes, 0=no)
- Blanks: Number of Blank tiles (0blanks, 1blank, or 2blanks)

Details

Results collected from one of the authors playing the "solo" mode of Words with Friends.

Source

Author’s iPhone
Wrinkle
Moving Wet Objects with Wrinkled Fingers

Description
Results from an experiment to move wet/dry objects with wrinkled/dry fingers

Format
A data frame with 80 observations on the following 7 variables.

- Participant: Participant ID (p1 to p20)
- Time: Time to move objects (seconds)
- Condition: non-wrinkled/dry, non-wrinkled/wet, wrinkled/dry, or wrinkled/wet
- Fingers: Status of fingers (non or wrinkled)
- Objects: Status of objects (dry or wet)
- WrinkledThenNon: Wrinkled first? (1=yes or 1=no)
- DryThenWet: Dry first? (1=yes or 1=no)

Details
Each of 20 participants were measured doing a “transfer task” several times under each of four conditions. The transfer task was to pick up an item with the right hand thumb and index finger, pass the item through a small hole and grab it with the left hand, and then put the item into a box that had a hole in the lid. Sometimes the participant’s fingers were wrinkled; sometimes the items were sitting in water.

Source

YouthRisk
Annual survey of health-risk youth behaviors

Description
Data from the Youth Risk Behavior Surveillance System
YouthRisk2007

**Format**

A data frame with 13387 observations on the following 6 variables.

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

**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 dataset focuses on whether or not youths have recently (in past 30 days) ridden with a drunk driver.

**Source**

http://www.cdc.gov/HealthyYouth/yrbs/index.htm
This dataset renamed as YouthRisk for the second edition.

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.

Description

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

Format

A dataset with 500 observations on the following 6 variables.

Sleep Average hours sleep on school night (10 or more hours, 9 hours, down to 4 or less hours)
Sleep7 Seven or more hours of sleep? (0=no or 1=yes)
SmokeLife Ever smoked? (No or Yes)
SmokeDaily Regular smoker? (No or Yes)
MarijuanaEver 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).
This data set is from the first edition, but not used in the second edition.

Source

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

Stand Your Ground Simpson’s Paradox

**Description**

Data from 220 cases in Florida where a "Stand your ground" defense was used.

**Format**

A data frame with 220 observations on the following 5 variables.

- **Convicted**: Was the defendant Convicted? (No or Yes)
- **IndWhiteVictim**: Was the victim white? (1=yes or 0=no)
- **IndWhiteDefendant**: Was the defendant white? (1=yes or 0=no)
- **VictimRace**: Race of the victim (Minority or White)
- **DefendantRace**: Race of the defendant (Minority or White)

**Details**

Inspired by the Travon Martin case, combined fatal and non-fatal cases of assault in Florida for which the defendant used the Stand Your Ground law in defense. These data show Simpson’s Paradox. Race of the victim is more important than race of the defendant.

**Source**

Data from Tampa Bay Times, male plus female cases, as of 2/8/15 – final posted data http://www.tampabay.com/stand-your-ground-law/nonfatal-cases http://www.tampabay.com/stand-your-ground-law/fatal-cases
Index

* datasets
  AccordPrice, 7
  AHCAvote2017, 7
  Airlines, 8
  Alfalfa, 8
  AlitoConfirmation, 9
  Amyloid, 10
  AppleStock, 10
  ArcheryData, 11
  AthleteGrad, 12
  AudioVisual, 12
  AutoPollution, 13
  Backpack, 14
  BaseballTimes, 14
  BaseballTimes2017, 15
  BeeStings, 16
  BirdCalcium, 16
  BirdNest, 17
  Blood1, 18
  BlueJays, 18
  BrainpH, 19
  BreesPass, 20
  BritishUnions, 20
  ButterfliesBc, 21
  CAFE, 21
  CalciumBP, 22
  CanadianDrugs, 23
  CancerSurvival, 24
  Caterpillars, 24
  CavsShooting, 25
  Cereal, 26
  ChemOThC, 26
  ChildSpeaks, 27
  ClintonSanders, 28
  Clothing, 28
  CloudSeeding, 29
  CloudSeeding2, 30
  CO2, 31
  CO2Germany, 32
  CO2Hawaii, 32
  CO2SouthPole, 33
  Contraceptives, 33
  CountyHealth, 35
  CrabShip, 36
  CrackerFiber, 36
  CreditRisk, 37
  Cuckoo, 37
  Day1Survey, 38
  DiabeticDogs, 39
  Diamonds, 40
  Diamonds2, 40
  Dinosaurs, 41
  Election08, 41
  Election16, 42
  ElephantsFB, 43
  ElephantsMF, 44
  Ethanol, 48
  Eyes, 48
  Faces, 49
  FaithfulFaces, 50
  FantasyBaseball, 50
  FatRats, 51
  Fertility, 52
  FGByDistance, 52
  Film, 53
  FinalFourIzzo, 54
  FinalFourIzzo17, 54
  FinalFourLong, 55
  FinalFourLong17, 56
  FinalFourShort, 57
  FinalFourShort17, 57
  Fingers, 58
  FirstYearGPA, 59
  FishEggs, 59
  Fitch, 60
  FlightResponse, 61
  FloridaDP, 61
  Fluorescence, 62
FranticFingers, 63
FruitFlies, 63
FruitFlies2, 64
FunnelDrop, 65
GlowWorms, 66
Goldenrod, 66
GrinnellHouses, 67
Grocery, 68
Gunnels, 69
Handwriting, 69
Hawks, 70
HawkTail, 71
HawkTail2, 72
HearingTest, 72
HeatingOil, 73
HighPeaks, 74
Hoops, 74
HorsePrices, 75
Houses, 76
HousesNY, 76
ICU, 77
InfantMortality2010, 78
Inflation, 78
InsuranceVote, 79
IQGuessing, 80
Jurors, 80
Kershaw, 81
KeyWestWater, 82
Kids198, 83
Leafhoppers, 83
LeafWidth, 84
Leukemia, 85
LeveeFailures, 85
LewyBody2Groups, 86
LewyDLBad, 87
LongJumpOlympics, 87
LongJumpOlympics2016, 88
LosingSleep, 88
LostLetter, 89
Marathon, 90
Markets, 90
MathEnrollment, 91
MathPlacement, 91
MedGPA, 92
Meniscus, 93
MentalHealth, 94
MetabolicRate, 94
MetroCommutes, 95
MetroHealth83, 96
Migraines, 96
Milgram, 97
MLB2007Standings, 98
MLBStandings2016, 99
MothEggs, 100
MouseBrain, 101
MusicTime, 102
NCDistances, 102
NFL2007Standings, 103
NFLStandings2016, 104
Nursing, 105
OilDeabsorption, 105
Olives, 106
Orings, 107
Overdrawn, 108
Oysters, 108
PalmBeach, 109
PeaceBridge2003, 110
PeaceBridge2012, 110
Pedometer, 111
Perch, 111
PigFeed, 112
Pines, 113
PKU, 114
Political, 114
Pollster08, 115
Popcorn, 116
PorscheJaguar, 116
PorschePrice, 117
Pulse, 118
Putts1, 118
Putts2, 119
Putts3, 119
RacialAnimus, 120
RadioactiveTwins, 121
RailsTrails, 122
Rectangles, 123
ReligionGDP, 124
RepeatedPulse, 124
ResidualOil, 125
Retirement, 126
Ricci, 126
RiverElements, 127
RiverIron, 128
SampleFG, 129
SandwichAnts, 130
SATGPA, 131
INDEX

SeaIce, 131
SeaSlugs, 132
SleepingShrews, 133
Sparrows, 135
SpeciesArea, 135
Speed, 136
SugarEthanol, 136
SuicideChina, 137
Swahili, 138
Tadpoles, 139
TechStocks, 139
TeenPregnancy, 140
TextPrices, 140
ThomasConfirmation, 141
ThreeCars, 142
ThreeCars2017, 142
TipJoke, 143
Tobacco, 144
TMS, 144
TomlinsonRush, 146
TwinsLungs, 148
Undoing, 148
USStamps, 149
VisualVerbal, 150
Volts, 150
WalkingBabies, 151
WalkTheDogs, 152
WeightLossIncentive, 152
WeightLossIncentive4, 153
WeightLossIncentive7, 154
Whickham2, 154
WordMemory, 155
WordsWithFriends, 156
Wrinkle, 157
YouthRisk, 157
YouthRisk2007, 158
YouthRisk2009, 159
Zimmerman, 160

* package
  Stat2Data-package, 6

AccordPrice, 7
AHCAvote2017, 7
Airlines, 8
Alfalfa, 8
AlitoConfirmation, 9
Amyloid, 10
AppleStock, 10
ArcheryData, 11

AthleteGrad, 12
AudioVisual, 12
AutoPollution, 13
Backpack, 14
BaseballTimes, 14
BaseballTimes2017, 15
BeeStings, 16
BirdCalcium, 16
BirdNest, 17
Blood1, 18
BlueJays, 18
BrainpH, 19
BreesPass, 20
BritishPass, 20
ButterfliesBc, 21

CAFE, 21
CalciumBP, 22
CanadianDrugs, 23
CancerSurvival, 24
Caterpillars, 24
CavsShooting, 25
Cereal, 26
ChemoTHC, 26
ChildSpeaks, 27
ClintonSanders, 28
Clothing, 28
CloudSeeding, 29
CloudSeeding2, 30
CO2, 31
CO2Germany, 32
CO2Hawaii, 32
CO2SouthPole, 33
Contraceptives, 33
cooksplot, 34
CountyHealth, 35
CrabShip, 36
CrackerFiber, 36
CreditRisk, 37
Cuckoo, 37

Day1Survey, 38
DiabeticDogs, 39
Diamonds, 40
Diamonds2, 40
Dinosaurs, 41

Election08, 41
Election16, 42
ElephantsFB, 43
ElephantsMF, 44
emplogitplot1, 44
emplogitplot2, 46
Ethanol, 48
Eyes, 48
Faces, 49
FaithfulFaces, 50
FantasyBaseball, 50
FatRats, 51
Fertility, 52
FGByDistance, 52
Film, 53
FinalFourIzzo, 54
FinalFourIzzo17, 54
FinalFourLong, 55
FinalFourLong17, 56
FinalFourShort, 57
FinalFourShort17, 57
Fingers, 58
FirstYearGPA, 59
FishEggs, 59
Fitch, 60
FlightResponse, 61
FloridaDP, 61
Fluorescence, 62
FranticFingers, 63
FruitFlies, 63
FruitFlies2, 64
FunnelDrop, 65
GlowWorms, 66
Goldenrod, 66
GrinnellHouses, 67
Grocery, 68
Gunnels, 69
Handwriting, 69
Hawks, 70
HawkTail, 71
HawkTail2, 72
HearingTest, 72
HeatingOil, 73
HighPeaks, 74
Hoops, 74
HorsePrices, 75
Houses, 76
HousesNY, 76
ICU, 77
InfantMortality2010, 78
Inflation, 78
InsuranceVote, 79
IQGuessing, 80
Jurors, 80
Kershaw, 81
KeyWestWater, 82
Kids198, 83
Leafhoppers, 83
LeafWidth, 84
Leukemia, 85
LeveeFailures, 85
LewyBody2Groups, 86
LewyDLBad, 87
LongJumpOlympics, 87
LongJumpOlympics2016, 88
LosingSleep, 88
LostLetter, 89
Marathon, 90
Markets, 90
MathEnrollment, 91
MathPlacement, 91
MedGPA, 92
Meniscus, 93
MentalHealth, 94
MetabolicRate, 94
MetroCommutes, 95
MetroHealth83, 96
Migraines, 96
Milgram, 97
MLB2007Standings, 98
MLBStandings2016, 99
MothEggs, 100
MouseBrain, 101
MusicTime, 102
NCbirths, 102
NFL2007Standings, 103
NFLStandings2016, 104
Nursing, 105
OilDeapsorbtion, 105
Olives, 106
INDEX

Orings, 107
Overdrawn, 108
Oysters, 108

PalmBeach, 109
PeaceBridge2003, 110
PeaceBridge2012, 110
Pedometer, 111
Perch, 111
PigFeed, 112
Pines, 113
PKU, 114
Political, 114
Pollster08, 115
Popcorn, 116
PorscheJaguar, 116
PorschePrice, 117
Pulse, 118
Putts1, 118
Putts2, 119
Putts3, 119
RacialAnimus, 120
RadioactiveTwins, 121
RailsTrails, 122
Rectangles, 123
ReligionGDP, 124
RepeatedPulse, 124
ResidualOil, 125
Retirement, 126
Ricci, 126
RiverElements, 127
RiverIron, 128
SampleFG, 129
SandwichAnts, 130
SATGPA, 131
SeaIce, 131
SeaSlugs, 132
SleepingShrews, 133
sluacf, 134
Sparrows, 135
SpeciesArea, 135
Speed, 136
Stat2Data (Stat2Data-package), 6
Stat2Data-package, 6
SugarEthanol, 136
SuicideChina, 137
Swahili, 138

Tadpoles, 139
TechStocks, 139
TeenPregnancy, 140
TextPrices, 140
ThomasConfirmation, 141
ThreeCars, 142
ThreeCars2017, 142
TipJoke, 143
Titanic, 144
TMS, 144
TomlinsonRush, 146
TukeyNonaddPlot, 147
TwinsLungs, 148
_undoing, 148
USstamps, 149
VisualVerbal, 150
Volts, 150
WalkingBabies, 151
WalkTheDogs, 152
WeightLossIncentive, 152
WeightLossIncentive4, 153
WeightLossIncentive7, 154
Whickham2, 154
WordMemory, 155
WordsWithFriends, 156
Wrinkle, 157
YouthRisk, 157
YouthRisk2007, 158
YouthRisk2009, 159
Zimmerman, 160