# Package 'Sleuth3'

June 15, 2016

**Title** Data Sets from Ramsey and Schafer's ``Statistical Sleuth (3rd Ed)''

**Version** 1.0-2

**Date** 2016-06-15

**Author** Original by F.L. Ramsey and D.W. Schafer; modifications by Daniel W. Schafer, Jeannie Sifneos and Berwin A. Turlach; vignettes contributed by Nicholas Horton, Linda Loi, Kate Aloisio and Ruobing Zhang, with corrections by Randall Pruim

**Description** Data sets from Ramsey, F.L. and Schafer, D.W. (2013), ``The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)'', Cengage Learning.

**Maintainer** Berwin A Turlach <Berwin.Turlach@gmail.com>

**LazyData** yes

**Depends** R (>= 3.1.0)

**Suggests** CCA, Hmisc, MASS, agricolae, car, gmodels, knitr, lattice, leaps, mosaic, multcomp

**VignetteBuilder** knitr

**License** GPL (>= 2)

**URL** http://r-forge.r-project.org/projects/sleuth2/

## R topics documented:

---

`Sleuth3-package` *The R Sleuth3 package*

---

## Description

Data sets from Ramsey and Schafer's "Statistical Sleuth (3rd ed)"

## Details

This package contains a variety of datasets. For a complete list, use `library(help="Sleuth3")` or `Sleuth3Manual()`.

## Author(s)

Original by F.L. Ramsey and D.W. Schafer

Modifications by Daniel W Schafer, Jeannie Sifneos and Berwin A Turlach

Maintainer: Berwin A Turlach `<Berwin.Turlach@gmail.com>`

---

`case0101` *Motivation and Creativity*

---

## Description

Data from an experiment concerning the effects of intrinsic and extrinsic motivation on creativity. Subjects with considerable experience in creative writing were randomly assigned to on of two treatment groups.

## Usage

```
case0101
```

## Format

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

**Score** creativity score

**Treatment** factor denoting the treatment group, with levels `"Extrinsic"` and `"Intrinsic"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Amabile, T. (1985). Motivation and Creativity: Effects of Motivational Orientation on Creative Writers, *Journal of Personality and Social Psychology* **48**(2): 393–399.

## Examples

```
attach(case0101)
str(case0101)
boxplot(Score ~ Treatment)  # Basic boxplots for each level of Treatment

boxplot(Score ~ Treatment,  # Boxplots with labels
  ylab= "Average Creativity Score From 11 Judges (on a 40-point scale)",
  names=c("23 'Extrinsic' Group Students","24 'Intrinsic' Group Students"),
  main= "Haiku Creativity Scores for 47 Creative Writing Students")

detach(case0101)
```

---

# case0102 *Sex Discrimination in Employment*

---

## Description

The data are the beginning salaries for all 32 male and all 61 female skilled, entry–level clerical employees hired by a bank between 1969 and 1977.

## Usage

```
case0102
```

## Format

A data frame with 93 observations on the following 2 variables.

**Salary** starting salaries (in US$)

**Sex** sex of the clerical employee, with levels `"Female"` and `"Male"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Roberts, H.V. (1979). Harris Trust and Savings Bank: An Analysis of Employee Compensation, *Report 7946*, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

## See Also

`case1202`

## Examples

```
attach(case0102)
str(case0102)

boxplot(Salary ~ Sex,
  ylab= "Starting Salary (U.S. Dollars)",
  names=c("61 Females","32 Males"),
  main= "Harris Bank Entry Level Clerical Workers, 1969-1971")

hist(Salary[Sex=="Female"])
dev.new()
hist(Salary[Sex=="Male"])

t.test(Salary ~ Sex, var.equal=TRUE) # Equal var. version; 2-sided by default
t.test(Salary ~ Sex, var.equal=TRUE,
  alternative = "less")  # 1-sided; that group 1 (females) mean is less

detach(case0102)
```

---

# case0201 *Peter and Rosemary Grant's Finch Beak Data*

---

## Description

In the 1980s, biologists Peter and Rosemary Grant caught and measured all the birds from more than 20 generations of finches on the Galapagos island of Daphne Major. In one of those years, 1977, a severe drought caused vegetation to wither, and the only remaining food source was a large, tough seed, which the finches ordinarily ignored. Were the birds with larger and stronger beaks for opening these tough seeds more likely to survive that year, and did they tend to pass this characteristic to their offspring? The data are beak depths (height of the beak at its base) of 89 finches caught the year before the drought (1976) and 89 finches captured the year after the drought (1978).

## Usage

```
case0201
```

## Format

A data frame with 178 observations on the following 2 variables.

**Year** Year the finch was caught, 1976 or 1978

**Depth** Beak depth of the finch (mm)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Grant, P. (1986). *Ecology and Evolution of Darwin's Finches*, Princeton University Press, Princeton, N.J.

## See Also

ex0218

## Examples

```
attach(case0201)
str(case0201)

mean(Depth[Year==1978]) - mean(Depth[Year==1976])

yearFactor <- factor(Year) # Convert the numerical variable Year into a factor
# with 2 levels. 1976 is "group 1" (it comes first alphanumerically)
t.test(Depth ~ yearFactor, var.equal=TRUE) # 2-sample t-test; 2-sided by default
t.test(Depth ~ yearFactor, var.equal=TRUE,
  alternative = "less") # 1-sided; alternative: group 1 mean is less

boxplot(Depth ~ Year,
  ylab= "Beak Depth (mm)",
  names=c("89 Finches in 1976","89 Finches in 1978"),
  main= "Beak Depths of Darwin Finches in 1976 and 1978")

## BOXPLOTS FOR PRESENTATION
boxplot(Depth ~ Year,
  ylab="Beak Depth (mm)", names=c("89 Finches in 1976","89 Finches in 1978"),
  main="Beak Depths of Darwin Finches in 1976 and 1978", col="green",
  boxlwd=2, medlwd=2, whisklty=1, whisklwd=2, staplewex=.2, staplelwd=2,
  outlwd=2, outpch=21, outbg="green", outcex=1.5)

detach(case0201)
```

---

# case0202 *Anatomical Abnormalities Associated with Schizophrenia*

---

## Description

Are any physiological indicators associated with schizophrenia? In a 1990 article, researchers reported the results of a study that controlled for genetic and socioeconomic differences by examining 15 pairs of monozygotic twins, where one of the twins was schizophrenic and the other was not. The researchers used magnetic resonance imaging to measure the volumes (in $cm^3$) of several regions and subregions of the twins' brains.

## Usage

```
case0202
```

## Format

A data frame with 15 observations on the following 2 variables.

**Unaffected** volume of left hippocampus of unaffected twin (in $cm^3$)

**Affected** volume of left hippocampus of affected twin (in $cm^3$)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Suddath, R.L., Christison, G.W., Torrey, E.F., Casanova, M.F. and Weinberger, D.R. (1990). Anatomical Abnormalities in the Brains of Monozygotic Twins Discordant for Schizophrenia, *New England Journal of Medicine* **322**(12): 789–794.

## Examples

```
attach(case0202)
str(case0202)

diff <- Unaffected-Affected
summary(diff)
t.test(diff) # Paired t-test is a one-sample t-test on differences
t.test(Unaffected,Affected,pair=TRUE)  # Alternative coding for the same test

boxplot(diff,
  ylab="Difference in Hippocampus Volume (cubic cm)",
  xlab="15 Sets of Twins, One Affected with Schizophrenia",
  main="Hippocampus Difference: Unaffected Twin Minus Affected Twin")
abline(h=0,lty=2)   # Draw a dashed (lty=2) horizontal line at 0

## BOXPLOT FOR PRESENTATION:
boxplot(diff,
  ylab="Difference in Hippocampus Volume (cubic cm)",
  xlab="15 Sets of Twins, One Affected with Schizophrenia",
  main="Hippocampus Difference: Unaffected Minus Affected Twin",
  col="green", boxlwd=2, medlwd=2, whisklty=1, whisklwd=2,
  staplewex=.2, staplelwd=2, outlwd=2, outpch=21, outbg="green",
  outcex=1.5)
abline(h=0,lty=2)

detach(case0202)
```

---

# case0301 *Cloud Seeding*

---

## Description

Does dropping silver iodide onto clouds increase the amount of rainfall they produce? In a randomized experiment, researchers measured the volume of rainfall in a target area (in acre-feet) on 26 suitable days in which the clouds were seeded and on 26 suitble days in which the clouds were not seeded.

## Usage

```
case0301
```

## Format

A data frame with 52 observations on the following 2 variables.

**Rainfall** the volume of rainfall in the target area (in acre-feet)

**Treatment** a factor with levels "Unseeded" and "Seeded" indicating whether the clouds were unseeded or seeded.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Simpson, J., Olsen, A., and Eden, J. (1975). A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification. *Technometrics* **17**: 161–166.

## Examples

```
attach(case0301)
str(case0301) #Seeded is level 1 of Treatment (it's first alphabetically)

boxplot(Rainfall ~ Treatment)
boxplot(log(Rainfall) ~ Treatment) # Boxplots of natural logs of Rainfall

t.test(log(Rainfall) ~ Treatment, var.equal=TRUE,
  alternative="greater") # 1-sided t-test; alternative: level 1 mean is greater


myTest <- t.test(log(Rainfall) ~ Treatment,  var.equal=TRUE,
  alternative="two.sided") # 2-sided alternative to get confidence interval
exp(myTest$est[1] - myTest$est[2])  # Back-transform estimate on log scale
exp(myTest$conf) # Back transform endpoints of confidence interval

boxplot(log(Rainfall) ~ Treatment,
  ylab="Log of Rainfall Volume in Target Area (Acre Feet)",
  names=c("On 26 Seeded Days", "On 26 Unseeded Days"),
  main="Distributions of Rainfalls from Cloud Seeding Experiment")

## POLISHED BOXPLOTS FOR PRESENTATION:
opar <- par(no.readonly=TRUE)  # Store device graphics parameters
par(mar=c(4,4,4,4))   # Change margins to allow more space on right
boxplot(log(Rainfall) ~ Treatment, ylab="Log Rainfall (Acre-Feet)",
  names=c("on 26 seeded days","on 26 unseeded days"),
  main="Boxplots of Rainfall on Log Scale", col="green", boxlwd=2,
  medlwd=2, whisklty=1, whisklwd=2, staplewex=.2, staplelwd=2,
  outlwd=2, outpch=21, outbg="green", outcex=1.5      )
myTicks <- c(1,5,10,100,500,1000,2000,3000) # some tick marks for original scale
axis(4, at=log(myTicks), label=myTicks)   # Add original-scale axis on right
mtext("Rainfall (Acre Feet)", side=4, line=2.5) # Add right-side axis label
par(opar)  # Restore previous graphics parameter settings

detach(case0301)
```

case0302 *Agent Orange*

## Description

In 1987, researchers measured the TCDD concentration in blood samples from 646 U.S. veterans of the Vietnam War and from 97 U.S. veterans who did not serve in Vietnam. TCDD is a carcinogenic dioxin in the herbicide called Agent Orange, which was used to clear jungle hiding areas by the U.S. military in the Vietnam War between 1962 and 1970.

## Usage

```
data(case0302)
```

## Format

A data frame with 743 observations on the following 2 variables.

**Dioxin** the concentration of TCDD, in parts per trillion

**Veteran** factor variable with two levels, `"Vietnam"` and `"Other"`, to indicate the type of veteran

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Centers for Disease Control Veterans Health Studies: Serum 2,3,7,8-Tetraclorodibenzo-p-dioxin Levels in U.S. Army Vietnam-era Veterans. *Journal of the American Medical Association* **260**: 1249–1254.

## Examples

```
attach(case0302)
str(case0302)    # Note: Level 1 of Veteran is "Other" (first alphabeticall)

boxplot(Dioxin ~ Veteran)

t.test(Dioxin ~ Veteran, var.equal=TRUE,
  alternative="less") # 1-sided t-test; alternative: group 1 mean is less
t.test(Dioxin ~ Veteran, alternative="less", var.equal=TRUE,
  subset=(Dioxin < 40)) # t-test on subset for which Dioxin < 40
t.test(Dioxin ~ Veteran, alternative="less", var.equal=TRUE,
  subset=(Dioxin < 20))
t.test(Dioxin ~ Veteran, var.equal=TRUE) # 2-sided--to get confidence interval

## HISTOGRAMS FOR PRESENTATION
opar <- par(no.readonly=TRUE)  # Store device graphics parameter settings
par(mfrow=c(2,1), mar=c(3,3,1,1)) # 2 by 1 layout of plots; change margins
myBreaks <- (0:46) - .5    # Make breaks for histogram bins
hist(Dioxin[Veteran=="Other"], breaks=myBreaks, xlim=range(Dioxin),
  col="green", xlab="", ylab="", main="")
```

```
text(10,25,
  "Dioxin in 97 'Other' Veterans; Estimated mean =  4.19 ppt (95% CI: 3.72 to 4.65 ppt)",
  pos=4, cex=.75) # CI from 1-sample t-test & subset=(Veteran="Other")
hist(Dioxin[Veteran=="Vietnam"],breaks=myBreaks,xlim=range(Dioxin),
  col="green", xlab="", ylab="", main="")
text(10,160,
  "Dioxin in 646 Vietnam Veterans; Estimated mean =  4.26 ppt (95% CI: 4.06 to 4.64 ppt)",
  pos=4, cex=.75)
text(13,145,"[Estimated Difference in Means: 0.07 ppt (95% CI: -0.63 to 0.48 ppt)]",
  pos=4, cex=.75)
par(opar) # Restore previous graphics parameter settings

detach(case0302)
```

---

`case0401` *Space Shuttle*

---

## Description

The number of space shuttle O-ring incidents for 4 space shuttle launches when the air temperatures was below 65 degrees F and for 20 space shuttle launches when the air temperature was above 65 degrees F.

## Usage

```
case0401
```

## Format

A data frame with 24 observations on the following 2 variables.

**Incidents** the number of O-ring incidents

**Launch** factor variable with two levels—`"Cool"` and `"Warm"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Feynman, R.P. (1988). *What do You Care What Other People Think?* W. W. Norton.

## See Also

`ex2011`, `ex2223`

## Examples

```
str(case0401)
attach(case0401)

mCool <- mean(Incidents[Launch=="Cool"])
mWarm <- mean(Incidents[Launch=="Warm"])
mDiff <- mCool - mWarm
c(mCool,mWarm,mDiff)  # Show the values of these variables

## PERMUTATION TEST , VIA REPEATED RANDOM RE-GROUPING (ADVANCED)
numRep  <- 50 # Number of random  groupings. CHANGE TO LARGER NUMBER; eg 50,000.
rDiff   <- rep(0,numRep) # Initialize this variable to contain numRep 0s.
for (rep in 1:numRep) {  # Repeat the following commands numRep times:
  randomGroup <- rep("rWarm",24)  # Set randomGroup to have 24 values "rWarm"
  randomGroup[sample(1:24,4)]  <- "rCool"  # Replace 4 at random with "rCool"
  mW  <- mean(Incidents[randomGroup=="rWarm"]) # average of random "rWarm" group
  mC  <- mean(Incidents[randomGroup=="rCool"]) # average of random "rCool" group
  rDiff[rep] <- mC-mW  # Store difference in averages in 'rep' cell of rDiff
           }  # End of loop
hist(rDiff,  # Histogram of difference in averages from numRep random groupings
  main="Approximate Permutation Distribution",
  xlab="Possible Values of Difference in Averages",
  ylab="Frequency of Occurrence")
abline(v=mDiff)  # Draw a vertical line at the actually observed difference
pValue <- sum(rDiff >= 1.3)/numRep  # 1-sided p-value
pValue
text(mDiff,75000, paste(" -->",round(pValue,4)), adj=-0.1)

detach(case0401)
```

---

`case0402` *Cognitive Load*

---

## Description

Educational researchers randomly assigned 28 ninth-year students in Australia to receive coordinate geometry training in one of two ways: a conventional way and a modified way. After the training, the students were asked to solve a coordinate geometry problem. The time to complete the problem was recorded, but five students in the "conventional" group did not complete the solution in the five minute alloted time.

## Usage

```
case0402
```

## Format

A data frame with 28 observations on the following 3 variables.

**Time** the time (in seconds) that the student worked on the problem

**Treatment** factor variable with two levels—`"Modified"` and `"Conventional"`

**Censored** 1 if the individual did not complete the problem in 5 minutes, 0 if they did

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Sweller, J., Chandler, P., Tierney, P. and Cooper, M. (1990). Cognitive Load as a Factor in the Structuring of Technical Material, *Journal of Experimental Psychology General* **119**(2): 176–192.

## Examples

```
str(case0402) # level 1 of Treatment is "Conventional" (1st alphabetically)
attach(case0402)

boxplot(Time ~ Treatment)
median(Time[Treatment=="Conventional"])-median(Time[Treatment=="Modified"])

wilcox.test(Time ~ Treatment, exact=FALSE, correct=TRUE,
  alternative="greater")  # Rank-sum test; alternative: group 1 is greater
wilcox.test(Time ~ Treatment, exact=FALSE, correct=TRUE,
  alternative="two.sided", conf.int=TRUE)  # Use 2-sided to get confidence int.

## DOT PLOTS FOR PRESENTATION
xTreatment    <- ifelse(Treatment=="Conventional",1,2) # Make numerical values
myPointCode   <- ifelse(Censored==0,21,24)
plot(Time ~ jitter(xTreatment,.2),   # Jitter the 1's and 2's for visibility
     ylab="Completion Time (Sec.)",  xlab="Training Method (jittered)",
     main="Test Completion Times from Cognitive Load Experiment",
     axes=FALSE, pch=myPointCode, bg="green", cex=2, xlim=c(.5,2.5) )
axis(2) # Draw y-axis as usual
axis(1, tick=FALSE,  at=c(1,2),  # Draw x-axis without ticks
  labels=c("Conventional (n=14 Students)","Modified (n=14 Students)") )
legend(1.5,300, legend=c("Did not Complete in 300 sec","Completed in 300 sec."),
       pch=c(24,21), pt.cex=2, pt.bg="green")

detach(case0402)
```

---

# case0501 *Diet Restriction and Longevity*

---

## Description

Female mice were randomly assigned to six treatment groups to investigate whether restricting dietary intake increases life expectancy. Diet treatments were:

1. `"NP"`—mice ate unlimited amount of nonpurified, standard diet
2. `"N/N85"`—mice fed normally before and after weaning. After weaning, ration was controlled at 85 kcal/wk
3. `"N/R50"`—normal diet before weaning and reduced calorie diet (50 kcal/wk) after weaning
4. `"R/R50"`—reduced calorie diet of 50 kcal/wk both before and after weaning
5. `"N/R50 lopro"`—normal diet before weaning, restricted diet (50 kcal/wk) after weaning and dietary protein content decreased with advancing age
6. `"N/R40"`—normal diet before weaning and reduced diet (40 Kcal/wk) after weaning.

## Usage

```
case0501
```

## Format

A data frame with 349 observations on the following 2 variables.

**Lifetime** the lifetime of the mice (in months)

**Diet** factor variable with six levels—"NP", "N/N85", "lopro", "N/R50", "R/R50" and "N/R40"

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Weindruch, R., Walford, R.L., Fligiel, S. and Guthrie D. (1986). The Retardation of Aging in Mice by Dietary Restriction: Longevity, Cancer, Immunity and Lifetime Energy Intake, *Journal of Nutrition* **116**(4):641–54.

## Examples

```
str(case0501)
attach(case0501)

# Re-order levels for better boxplot organization:
myDiet  <- factor(Diet, levels=c("NP","N/N85","N/R50","R/R50","lopro","N/R40") )

myNames <- c("NP(49)","N/N85(57)","N/R50(71)","R/R50(56)","lopro(56)",
  "N/R40(60)")   # Make these for boxplot labeling.
boxplot(Lifetime ~ myDiet, ylab= "Lifetime (months)", names=myNames,
  xlab="Treatment (and sample size)")
myAov1   <- aov(Lifetime ~ Diet) # One-way analysis of variance
plot(myAov1, which=1) # Plot residuals versus estimated means.
summary(myAov1)
pairwise.t.test(Lifetime,Diet, pool.SD=TRUE, p.adj="none") # All t-tests

## p-VALUES AND CONFIDENCE INTERVALS FOR SPECIFIED COMPARISONS OF MEANS
if(require(multcomp)){
  diet    <- factor(Diet,labels=c("NN85", "NR40", "NR50", "NP", "RR50", "lopro"))
  myAov2  <- aov(Lifetime ~ diet - 1)
  myComparisons <- glht(myAov2,
          linfct=c("dietNR50 - dietNN85 = 0",
          "dietRR50  - dietNR50 = 0",
          "dietNR40  - dietNR50 = 0",
          "dietlopro - dietNR50 = 0",
          "dietNN85  - dietNP   = 0")   )
  summary(myComparisons,test=adjusted("none")) # No multiple comparison adjust.
  confint(myComparisons, calpha = univariate_calpha()) # No adjustment
}

## EXAMPLE 5: BOXPLOTS FOR PRESENTATION
boxplot(Lifetime ~ myDiet, ylab= "Lifetime (months)", names=myNames,
  main= "Lifetimes of Mice on 6 Diet Regimens",
  xlab="Diet (and sample size)", col="green", boxlwd=2, medlwd=2, whisklty=1,
```

```
  whisklwd=2, staplewex=.2, staplelwd=2, outlwd=2, outpch=21, outbg="green",
  outcex=1.5)

detach(case0501)
```

---

case0502 *The Spock Conspiracy Trial*

---

## Description

In 1968, Dr. Benjamin Spock was tried in Boston on charges of conspiring to violate the Selective Service Act by encouraging young men to resist being drafted into military service for Vietnam. The defence in the case challenged the method of jury selection claiming that women were underrepresented. Boston juries are selected in three stages. First 300 names are selected at random from the City Directory, then a venire of 30 or more jurors is selected from the initial list of 300 and finally, an actual jury is selected from the venire in a nonrandom process allowing each side to exclude certain jurors. There was one woman on the venire and no women on the final list. The defence argued that the judge in the trial had a history of venires in which women were systematically underrepresented and compared the judge's recent venires with the venires of six other Boston area district judges.

## Usage

```
case0502
```

## Format

A data frame with 46 observations on the following 2 variables.

**Percent** is the percent of women on the venire's of the Spock trial judge and 6 other Boston area judges

**Judge** is a factor with levels `"Spock's"`, `"A"`, `"B"`, `"C"`, `"D"`, `"E"` and `"F"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Zeisel, H. and Kalven, H. Jr. (1972). Parking Tickets and Missing Women: Statistics and the Law *in* Tanur, J.M. et al. (eds.) *Statistics: A Guide to the Unknown*, Holden-Day.

## Examples

```
str(case0502)
attach(case0502)

# Make new factor level names (with sample sizes) for boxplots
myNames <- c("A (5)", "B (6)", "C (9)", "D (2)", "E (6)", "F (9)", "Spock's (9)")

boxplot(Percent ~ Judge, ylab = "Percent of Women on Judges' Venires",
```

```
  names = myNames, xlab = "Judge (and number of venires)",
  main = "Percent Women on Venires of 7 Massachusetts Judges")
myAov1  <- aov(Percent ~ Judge)
plot(myAov1, which=1)   # Residual plot
summary(myAov1) # Initial screening. Any evidence of judge differences? (yes)

## ANALYSIS 1. TWO-SAMPLE t-TEST (ASSUMING NON-SPOCK JUDGES HAVE A COMMON MEAN)
SpockOrOther <- factor(ifelse(Judge=="Spock's","Spock","Other"))
aovFull      <- aov(Percent ~ Judge)
aovReduced   <- aov(Percent ~ SpockOrOther)
anova(aovReduced,aovFull) #Any evidence that 7 mean fits better than the 2 mean?
t.test(Percent ~ SpockOrOther, var.equal=TRUE) # Evidence that 2 means differ?

## ANALYSIS 2. COMPARE SPOCK MEAN TO AVERAGE OF OTHER MEANS
  myAov3        <- aov(Percent ~ Judge - 1)
  myContrast    <- rbind(c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6, - 1))
if(require(multcomp)){  # use multcomp library
  myComparison  <- glht(myAov3, linfct=myContrast)
  summary(myComparison, test=adjusted("none"))
  confint(myComparison)
}

## BOXPLOTS FOR PRESENTATION
boxplot(Percent ~ Judge,  ylab= "Percent of Women on Judges' Venires",
        names=myNames, xlab="Judge (and number of venires)",
        main= "Percent Women on Venires of 7 Massachusetts Judges",
        col="green", boxlwd=2,  medlwd=2,  whisklty=1,  whisklwd=2,
        staplewex=.2, staplelwd=2,  outlwd=2,  outpch=21, outbg="green",
        outcex=1.5)

detach(case0502)
```

---

`case0601` *Discrimination Against the Handicapped*

---

## Description

Study explores how physical handicaps affect people's perception of employment qualifications. Researchers prepared 5 videotaped job interviews using actors with a script designed to reflect an interview with an applicant of average qualifications. The 5 tapes differed only in that the applicant appeared with a different handicap in each one. Seventy undergraduate students were randomly assigned to view the tapes and rate the qualification of the applicant on a 0-10 point scale.

## Usage

```
case0601
```

## Format

A data frame with 70 observations on the following 2 variables.

**Score** is the score each student gave to the applicant

**Handicap** is a factor variable with 5 levels—`"None"`, `"Amputee"`, `"Crutches"`, `"Hearing"` and `"Wheelchair"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Cesare, S.J., Tannenbaum, R.J. and Dalessio, A. (1990). Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy, *Human Performance* **3**(3): 157–171.

## Examples

```
str(case0601)
attach(case0601)

## EXPLORATION
myHandicap  <- factor(Handicap,
  levels=c("None","Amputee","Crutches","Hearing","Wheelchair"))
boxplot(Score ~ myHandicap,
  ylab= "Qualification Score Assigned by Student to Interviewee",
  xlab= "Treatment Group--Handicap Portrayed (14 Students in each Group)",
  main= "Handicap Discrimination Experiment on 70 Undergraduate Students")
myAov  <- aov(Score ~ myHandicap)
plot(myAov, which=1) # Plot residuals versus estimated means
summary(myAov)

## COMPARE MEAN QUALIFICATION SCORE OF EVERY HANDICAP GROUP TO "NONE"
if(require(multcomp)){     # Use the multcomp library
  myDunnett  <- glht(myAov, linfct = mcp(myHandicap = "Dunnett"))
  summary(myDunnett)
  confint(myDunnett,level=.95)
  opar <- par(no.readonly=TRUE)  # Save current graphics parameter settings
  par(mar=c(4.1,8.1,4.1,1.1)) # Change margins
  plot(myDunnett,
    xlab="Difference in Mean Qualification Score (and Dunnet-adjusted CIs)")
  par(opar)  # Restore original graphics parameter settings
}

## COMPARE EVERY MEAN TO EVERY OTHER MEAN
if(require(multcomp)){   # Use the multcomp library
  myTukey   <- glht(myAov, linfct = mcp(myHandicap = "Tukey"))
  summary(myTukey)
}

## TEST THE CONTRAST OF DISPLAY 6.4
myAov2        <- aov(Score ~ myHandicap - 1)
myContrast    <- rbind(c(0, -1/2, 1/2, -1/2, 1/2))
if(require(multcomp)){   # Use the multcomp library
  myComparison  <- glht(myAov2, linfct=myContrast)
  summary(myComparison, test=adjusted("none"))
  confint(myComparison)
}


# BOXPLOTS FOR PRESENTATION
boxplot(Score ~ myHandicap,
  ylab= "Qualification Score Assigned by Student to Video Job Applicant",
```

```
  xlab="Handicap Portrayed by Job Applicant in Video (14 Students in each Group)",
  main= "Handicap Discrimination Experiment on 70 Undergraduate Students",
  col="green", boxlwd=2, medlwd=2, whisklty=1, whisklwd=2, staplewex=.2,
staplelwd=2, outlwd=2, outpch=21,  outbg="green", outcex=1.5)

detach(case0601)
```

---

`case0602` *Mate Preference of Platyfish*

---

## Description

Do female Platyfish prefer male Platyfish with yellow swordtails? A.L. Basolo proposed and tested a selection model in which females have a pre-existing bias for a male trait even before the males possess it. Six pairs of males were surgically given artificial, plastic swordtails—one pair received a bright yellow sword, the other a transparent sword. Females were given the opportunity to engage in courtship activity with either of the males. Of the total time spent by each female engaged in courtship during a 20 minute observation period, the percentages of time spent with the yellow-sword male were recorded.

## Usage

```
case0602
```

## Format

A data frame with 84 observations on the following 3 variables.

**Percentage** The percentage of courtship time spent by 84 females with the yellow-sword males

**Pair** Factor variable with 6 levels—`"Pair1"`, `"Pair2"`, `"Pair3"`, `"Pair4"`, `"Pair5"` and `"Pair6"`

**Length** Body size of the males

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Basolo, A.L. (1990). Female Preference Predates the Evolution of the Sword in Swordtail Fish, *Science* **250**: 808–810.

## Examples

```
str(case0602)
attach(case0602)

## EXPLORATION
plot(Percentage ~ Length,
  xlab="Length of the Two Males",
  ylab="Percentage of Time Female Spent with Yellow-Sword Male",
  main="Percentage of Time Spent with Yellow Rather than Transparent Sword Male")
abline(h=50)    # Draw a horizontal line at 50% (i.e. the "no preference" line)
```

```
myAov  <- aov(Percentage ~ Pair)
plot(myAov, which=1) # Resdiual plot
summary(myAov)

# Explore possibility of linear effect, as in Display 6.5
myAov2        <- aov(Percentage ~ Pair - 1)  # Show the estimated means.
myContrast    <- rbind(c(5, -3, 1, 3, -9, 3))
if(require(multcomp)){   # Use the multcomp library
  myComparison  <- glht(myAov2, linfct=myContrast)
  summary(myComparison, test=adjusted("none"))
}


# Simpler exploration of linear effect, via regression (Ch. 7)
myLm          <- lm(Percentage ~ Length)
summary(myLm)

# ONE-SAMPLE t-TEST THAT MEAN PERCENTAGE = 50%, IGNORING MALE PAIR EFFECT
t.test(Percentage, mu=50, alternative="greater") # Get 1-sided p-value
t.test(Percentage, alternative="two.sided")  # Get C.I.

## SCATTERPLOT FOR PRESENTATION
plot(Percentage ~ Length,
    xlab="Length of the Two Males (mm)",
    ylab="Percentage of Time Female Spent with Yellow-Sword Male",
    main="Female Preference for Yellow Rather than Transparent Sword Male",
    pch=21, lwd=2, bg="green", cex=1.5 )
abline(h=50,lty=2,col="blue",lwd=2)
text(29.5,52,"50% (no preference)", col="blue")

detach(case0602)
```

---

`case0701` *The Big Bang*

---

## Description

Hubble's initial data on 24 nebulae outside the Milky Way.

## Usage

```
case0701
```

## Format

A data frame with 24 observations on the following 2 variables.

**Velocity** recession velocity (in kilometres per second)

**Distance** distance from earth (in magaparsec)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Hubble, E. (1929). A Relation Between Distance and Radial Velocity Among Extragalactic Nebulae, *Proceedings of the National Academy of Science* **15**: 168–173.

## See Also

ex0725

## Examples

```
str(case0701)
attach(case0701)

## EXPLORATION
plot(Distance ~ Velocity)
myLm <- lm(Distance ~ Velocity)
abline(myLm)

myResiduals <- myLm$res
myFits  <- myLm$fit
plot(myResiduals ~ myFits)   # Plot residuals versus estimated means.
abline(h=0)  # Draw a horizontal line at 0.
# OR, use this shortcut...
plot(myLm, which=1) # Residual plot (red curve is a scatterplot smooother)

## INFERENCE
summary(myLm)
confint(myLm,level=.95)
myLm2 <- lm(Distance ~ Velocity - 1)   # Drop the intercept.
summary(myLm2)
confint(myLm2)

## DISPLAY FOR PRESENTATION
plot(Distance ~ Velocity, xlab="Recession Velocity (km/sec)",
  ylab="Distance from Earth (megaparsecs)",
  main="Measured Distance Versus Velocity for 24 Extra-Galactic Nebulae",
  pch=21, lwd=2, bg="green", cex=1.5 )
abline(myLm, lty=2, col="blue", lwd=2)
abline(myLm2, lty=3, col="red", lwd=2)
legend(-250,2.05,
  c("unrestricted regression line","regression through the origin"),
  lty=c(2,3), lwd=c(2,2), col=c("blue","red"))

detach(case0701)
```

---

# case0702 *Meat Processing and pH*

---

## Description

A certain kind of meat processing may begin once the pH in postmortem muscle of a steer carcass has decreased sufficiently. To estimate the timepoint at which pH has dropped sufficiently, 10 steer carcasses were assigned to be measured for pH at one of five times after slaughter.

## Usage

```
case0702
```

## Format

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

**Time** time after slaughter (hours)

**pH** pH level in postmortem muscle

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Schwenke, J.R. and Milliken, G.A. (1991). On the Calibration Problem Extended to Nonlinear Models, *Biometrics* **47**(2): 563–574.

## See Also

ex0816

## Examples

```
str(case0702)
attach(case0702)

# EXPLORATION
plot(pH ~ Time)
myLm <- lm(pH ~ Time)
abline(myLm, col="blue", lwd=2)
lines(lowess(Time,pH), col="red", lty=2, lwd=2) # Add scatterplot smoother
plot(myLm, which=1)  # Residual plot

logTime <- log(Time)
plot(pH ~ logTime)
myLm2   <- lm(pH ~ logTime)
abline(myLm2)
plot(myLm2, which=1)

## PREDICTION BAND ABOUT REGRESSION LINE
xToPredict    <- seq(1,8,length=100) # sequence from 1 to 8 of length 100
logXToPredict <- log(xToPredict)
newData       <- data.frame(logTime = logXToPredict)
myPredict     <- predict(myLm2,newData,
  interval="prediction", level=.90)
plot(pH ~ logTime)
abline(myLm2)
lines(myPredict[,3]~ logXToPredict, lty=2)
lines(myPredict[,2] ~ logXToPredict, lty=2)
# Find smallest time at which the upper endpoint of a 90% prediction
# interval is less than or equal to 6:
minTime <- min(xToPredict[myPredict[,3] <= 6.0])
minTime
```

```
abline(v=log(minTime),col="red")

# DISPLAY FOR PRESENTATION
plot(pH ~ Time, xlab="Time After Slaughter (Hours); log scale",
  ylab="pH in Muscle", main="pH and Time after Slaughter for 10 Steers",
  log="x", pch=21, lwd=2, bg="green", cex=2 )
lines(xToPredict,myPredict[,1], col="blue",  lwd=2)
lines(xToPredict, myPredict[,3], lty=2, col="blue", lwd=2)
lines(xToPredict, myPredict[,2], lty=2,   col="blue", lwd=2)
legend(3,7, c("Estimated Regression Line","90% Prediction Band"),
  lty=c(1,2), col="blue", lwd=c(2,2))
abline(h=6, lty=3, col="purple", lwd=2)
text(1.5,6.05,"Desired pH", col="purple")
lines(c(minTime,minTime),c(5,6.15), col="purple", lwd=2)
text(minTime,6.2,"4.9 hours",col="purple",cex=1.25)

detach(case0702)
```

---

`case0801` *Island Area and Number of Species*

---

## Description

The data are the numbers of reptile and amphibian species and the island areas for seven islands in the West Indies.

## Usage

```
case0801
```

## Format

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

**Area** area of island (in square miles)

**Species** number of reptile and amphibian species on island

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Wilson, E.O., 1992, *The Diversity of Life*, W. W. Norton, N.Y.

## Examples

```
str(case0801)
attach(case0801)

## EXPLORATION
logSpecies <- log(Species)
logArea  <- log(Area)
```

```
plot(logSpecies ~ logArea, xlab="Log of Island Area",
  ylab="Log of Number of Species",
  main="Number of Reptile and Amphibian Species on 7 Islands")
myLm <- lm(logSpecies ~ logArea)
abline(myLm)

## INFERENCE AND INTERPRETATION
summary(myLm)
slope     <- myLm$coef[2]
slopeConf <- confint(myLm,2)
100*(2^(slope)-1)   # Back-transform estimated slope
100*(2^(slopeConf)-1) # Back-transform confidence interval
# Interpretation: Associated with each doubling of island area is a 19% increase
# in the median number of bird species (95% CI: 16% to 21% increase).

## DISPLAY FOR PRESENTATION
plot(Species ~ Area, xlab="Island Area (Square Miles); Log Scale",
  ylab="Number of Species; Log Scale",
  main="Number of Reptile and Amphibian Species on 7 Islands",
  log="xy", pch=21, lwd=2, bg="green",cex=2 )
dummyArea <- c(min(Area),max(Area))
beta <- myLm$coef
meanLogSpecies <-  beta[1] + beta[2]*log(dummyArea)
medianSpecies  <-  exp(meanLogSpecies)
lines(medianSpecies ~ dummyArea,lwd=2,col="blue")
island <- c(" Cuba"," Hispaniola"," Jamaica", " Puerto Rico",
  " Montserrat"," Saba"," Redonda")
for (i in 1:7) {
   offset <- ifelse(Area[i] < 10000, -.2, 1.5)
   text(Area[i],Species[i],island[i],col="dark green",adj=offset,cex=.75) }

detach(case0801)
```

---

`case0802` *Breakdown Times for Insulating Fluid under different Voltage*

---

## Description

In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages until the insulating property of the fluids broke down. Seven different voltage levels were studied and the measured reponses were the times until breakdown.

## Usage

```
case0802
```

## Format

A data frame with 76 observations on the following 3 variables.

**Time** times until breakdown (in minutes)

**Voltage** voltage applied (in kV)

**Group** factor variable (group number)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Nelson, W.B., 1970, G.E. Co. Technical Report 71-C-011, Schenectady, N.Y.

## Examples

```
str(case0802)
attach(case0802)

## EXPLORATION
plot(Time ~ Voltage)
myLm <- lm(Time ~ Voltage)
plot(myLm, which=1)   # Residual plot
logTime <- log(Time)
plot(logTime ~ Voltage)
myLm <- lm(logTime ~ Voltage)
abline(myLm)
plot(myLm,which=1)  # Residual plot
myOneWay <- lm(logTime ~ factor(Voltage))
anova(myLm, myOneWay)  # Lack of fit test for simple regression (seems okay)

## INFERENCE AND INTERPREATION
beta <- myLm$coef
100*(1 - exp(beta[2]))   # Back-transform estimated slope
100*(1 - exp(confint(myLm,"Voltage")))
# Interpretation: Associated with each 1 kV increase in voltage is a 39.8%
# decrease in median breakdown time (95% CI: 32.5% decrease to 46.3% decrease).

## DISPLAY FOR PRESENTATION
options(scipen=50)  # Do this to avoid scientific notation on y-axis
plot(Time ~ Voltage, log="y", xlab="Voltage (kV)",
  ylab="Breakdown Time (min.); Log Scale",
  main="Breakdown Time of Insulating Fluid as a Function of Voltage Applied",
  pch=21, lwd=2, bg="green", cex=1.75 )
dummyVoltage <- c(min(Voltage),max(Voltage))
meanLogTime <- beta[1] + beta[2]*dummyVoltage
medianTime <- exp(meanLogTime)
lines(medianTime ~ dummyVoltage, lwd=2, col="blue")

detach(case0802)
```

---

# case0901 *Effects of Light on Meadowfoam Flowering*

---

## Description

Meadowfoam is a small plant found growing in moist meadows of the US Pacific Northwest. Researchers reported the results from one study in a series designed to find out how to elevate meadowfoam production to a profitable crop. In a controlled growth chamber, they focused on the effects of two light–related factors: light intensity and the timeing of the onset of the ligth treatment.

## Usage

```
case0901
```

## Format

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

**Flowers** average number of flowers per meadowfoam plant

**Time** time light intensity regiments started; 1=Late, 2=Early

**Intensity** light intensity (in $\mu$mol/m$^2$/sec)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(case0901)
attach(case0901)

## EXPLORATION
plot(Flowers ~ Intensity, pch=ifelse(Time ==1, 19, 21))
myLm <- lm(Flowers ~ Intensity + factor(Time) + Intensity:factor(Time))
plot(myLm, which=1)
summary(myLm)  # Note p-value for interaction term

# INFERENCE
myLm2 <- lm(Flowers ~ Intensity + factor(Time))
summary(myLm2)
confint(myLm2)

# DISPLAY FOR PRESENTATION
plot(Flowers ~ jitter(Intensity,.3),
  xlab=expression("Light Intensity ("*mu*"mol/"*m^2*"/sec)"), # Include symbols
  ylab="Average Number of Flowers per Plant",
  main="Effect of Light Intensity and Timing on Meadowfoam Flowering",
  pch=ifelse(Time ==1, 21, 22), bg=ifelse(Time==1, "orange","green"),
  cex=1.7, lwd=2)
beta <- myLm2$coef
abline(beta[1],beta[2],lwd=2, lty=2)
abline(beta[1]+beta[3],beta[2],lwd=2,lty=3)
legend(700,79,c("Early Start","Late Start"),
  pch=c(22,21),lwd=2,pt.bg=c("green","orange"),pt.cex=1.7,lty=c(3,2))

detach(case0901)
```

---

## case0902 *Why Do Some Mammals Have Large Brains for Their Size?*

---

## Description

The data are the average values of brain weight, body weight, gestation lengths (length of pregnancy) and litter size for 96 species of mammals.

## Usage

```
case0902
```

## Format

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

**Species** species

**Brain** average brain weight (in grams)

**Body** average body weight (in kilograms)

**Gestation** gestation period (in days)

**Litter** average litter size

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex0333

## Examples

```
str(case0902)
attach(case0902)

## EXPLORATION
myMatrix      <- cbind(Brain, Body, Litter, Gestation)
if(require(car)){   # Use the car library
scatterplotMatrix(myMatrix,   # Matrix of scatterplots
  smooth=FALSE,    # Omit scatterplot smoother on plots
  diagonal="histogram") # Draw histograms on diagonals
  }
myLm <- lm(Brain ~ Body + Litter + Gestation)
plot(myLm, which=1)
logBrain <- log(Brain)
logBody <- log(Body)
logGestation  <- log(Gestation)
myMatrix2 <- cbind(logBrain,logBody,Litter, logGestation)
if(require(car)){   # Use the car library
  scatterplotMatrix(myMatrix2, smooth=FALSE, diagonal="histogram")
}
myLm2 <- lm(logBrain ~ logBody + Litter + logGestation)
plot(myLm2,which=1)  # Residual plot.

if(require(car)){   # Use the car library
crPlots(myLm2)  # Partial residual plots (Sleuth Ch.11)
}
plot(logBrain ~ logBody)
identify(logBrain ~ logBody,labels=Species)   # Identify points on  scatterplot
# Place the cursor over a point of interest, then left-click.
# Continue with other points if desired. When finished, pres Esc.

## INFERENCE
```

```
summary(myLm2)
confint(myLm2)

# DISPLAYS FOR PRESENTATION
myLm3 <- lm(logBrain ~ logBody + logGestation)
beta <- myLm3$coef
logBrainAdjusted  <- logBrain - beta[2]*logBody
y <- exp(logBrainAdjusted)
ymod <- 100*y/median(y)
plot(ymod ~ Gestation, log="xy",
  xlab="Average Gestation Length (Days); Log Scale",
  ylab="Brain Weight Adjusted for Body Weight, as a Percentage of the Median",
  main="Brain Weight Adjusted for Body Weight, Versus Gestation Length, for 96 Mammal Species",
  pch=21,bg="green",cex=1.3)
identify(ymod ~ Gestation,labels=Species, cex=.7) # Identify points, as desired
# Press Esc to complete identify.
abline(h=100,lty=2) # Draw horizontal line at 100%

myLm4 <- lm(logBrain ~ logBody + Litter)
beta  <- myLm4$coef
logBrainAdjusted <- logBrain - beta[2]*logBody
y2 <- exp(logBrainAdjusted)
y2mod <- 100*y2/median(y2)
plot(y2mod ~ Litter, log="y", xlab="Average Litter Size",
  ylab="Brain Weight Adjusted for Body Weight, as a Percentage of the Median",
  main="Brain Weight Adjusted for Body Weight, Versus Litter Size, for 96 Mammal Species",
  pch=21,bg="green",cex=1.3)
identify(y2mod ~ Litter,labels=Species, cex=.7)
abline(h=100,lty=2)

detach(case0902)
```

---

`case1001` *Galileo's Data on the Motion of Falling Bodies*

---

## Description

In 1609 Galileo proved mathematically that the trajectory of a body falling with a horizontal velocity component is a parabola. His search for an experimental setting in which horizontal motion was not affected appreciably (to study inertia) let him to construct a certain apparatus. The data comes from one of his experiments.

## Usage

```
case1001
```

## Format

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

**Distance** horizontal distances (in punti)

**Height** initial height (in punti)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(case1001)
attach(case1001)

## EXPLORATION
plot(Distance ~ Height)
myLm <- lm(Distance ~ Height)
plot(myLm, which=1)
height2 <- Height^2
myLm2 <- lm(Distance ~ Height + height2)
plot(myLm2, which=1)
summary(myLm2) # Note p-value for quadratic term (it's small)
height3 <- Height^3
myLm3 <- update(myLm2, ~ . + height3)
plot(myLm3,which=1)
summary(myLm3) # Note p-value for cubic term (it's small)
height4 <- Height^4
myLm4 <- update(myLm3, ~ . + height4)
summary(myLm4) # Note p-value for quartic term (it's not small)

## DISPLAY FOR PRESENTATION
plot(Distance ~ Height, xlab="Initial Height (Punti)",
  ylab="Horizontal Distance Traveled (Punti)",
  main="Galileo's Falling Body Experiment",
  pch=21, bg="green", lwd=2, cex=2)
dummyHeight     <- seq(min(Height),max(Height),length=100)
betaQ           <- myLm2$coef
quadraticCurve  <- betaQ[1] + betaQ[2]*dummyHeight + betaQ[3]*dummyHeight^2
lines(quadraticCurve ~ dummyHeight,col="blue",lwd=3)
betaC           <- myLm3$coef # coefficients of cubic model
cubicCurve      <- betaC[1] + betaC[2]*dummyHeight + betaC[3]*dummyHeight^2 +
  betaC[4]*dummyHeight^3
lines(cubicCurve ~ dummyHeight,lty=3,col="red",lwd=3)
legend(590,290,legend=c(expression("Quadratic Fit  "*R^2*" = 99.0%"),
  expression("Cubic Fit        "*R^2*" = 99.9%")),
  lty=c(1,3),col=c("blue","red"), lwd=c(3,3))

detach(case1001)
```

---

`case1002` *The Energy Costs of Echolocation by Bats*

---

## Description

The data are on in–flight energy expenditure and body mass from 20 energy studies on three types of flying vertebrates: echolocating bats, non–echolocating bats and non–echolocating birds.

## Usage

```
case1002
```

## Format

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

**Mass** mass (in grams)

**Type** a factor with 3 levels indicating the type of flying vertebrate: non-echolocating bats, non-echolocating birds, echolocating bats

**Energy** in–flight energy expenditure (in W)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Speakman, J.R. and Racey, P.A. (1991). No cost of Echolocation for Bats in Flight, *Nature* **350**: 421–423.

## Examples

```
str(case1002)
attach(case1002)

## EXPLORATION
plot(Energy~Mass, case1002, log="xy", xlab = "Body Mass (g) (log scale)",
  ylab = "Energy Expenditure (W) (log scale)",
  pch = ifelse(Type=="echolocating bats", 19,
     ifelse(Type=="non-echolocating birds", 21, 24)))
legend(7, 50, pch=c(24, 21, 19),
     c("Non-echolocating bats", "Non-echolocating birds","Echolocating bats"))

logEnergy  <- log(Energy)
logMass <- log(Mass)
myLm2 <- lm(logEnergy ~ logMass + Type + logMass:Type)
plot(myLm2, which=1)
myLm3 <- update(myLm2, ~ . - logMass:Type)
anova(myLm3, myLm2)   # Test for interaction with extra ss F-test

## INFERENCE AND INTERPRETATION
myLm4 <- update(myLm3, ~ . - Type)  # Reduced model...with no effect of Type
anova(myLm4, myLm3)   # Test for Type effect
myType <- factor(Type,
 levels=c("non-echolocating bats","echolocating bats","non-echolocating birds"))
myLm3a <- lm(logEnergy ~ logMass + myType)
summary(myLm3a)
100*(exp(myLm3a$coef[3]) - 1)
100*(exp(confint(myLm3a,3))-1)
# Conclusion: Adjusted for body mass, the median energy expenditure for
# echo-locating bats exceeds that for echo-locating bats by an estimated
# 8.2% (95% confidence interval: 29.6% LESS to 66.3% MORE)

# DISPLAY FOR PRESENTATION
myPlotCode    <- ifelse(Type=="non-echolocating birds",24,21)
myPointColor  <- ifelse(Type=="echolocating bats","green","white")
plot(Energy ~ Mass, log="xy", xlab="Body Mass (g); Log Scale ",
```

```
  ylab="In-Flight Energy Expenditure (W); Log Scale",
  main="In-Flight Energy Expenditure Study",
  pch=myPlotCode,bg=myPointColor,lwd=2, cex=1.5)
dummyMass  <- seq(5,800,length=50)
beta       <- myLm3$coef
curve1     <- exp(beta[1] + beta[2]*log(dummyMass))
curve2     <- exp(beta[1] + beta[2]*log(dummyMass) + beta[3])
curve3     <- exp(beta[1] + beta[2]*log(dummyMass) + beta[4])
lines(curve1 ~ dummyMass)
lines(curve2 ~ dummyMass, lty=2)
lines(curve3 ~ dummyMass, lty=3)
legend(100,3,
  c("Echolocating Bats","Non-Echolocating Bats","Non-Echolocating Birds"),
  pch=c(21,21,24),lwd=2,pt.cex=c(1.5,1.5,1.5),pt.lwd=c(2,2,2),
  pt.bg=c("green","white","white"),lty=c(1,2,3))

detach(case1002)
```

---

`case1101` *Alcohol Metabolism in Men and Women*

---

## Description

These data were collected on 18 women and 14 men to investigate a certain theory on why women exhibit a lower tolerance for alcohol and develop alcohol–related liver disease more readily than men.

## Usage

```
case1101
```

## Format

A data frame with 32 observations on the following 5 variables.

**Subject** subject number in the study

**Metabol** first–pass metabolism of alcohol in the stomach (in mmol/liter-hour)

**Gastric** gastric alcohol dehydrogenase activity in the stomach (in $\mu$mol/min/g of tissue)

**Sex** sex of the subject

**Alcohol** whether the subject is alcoholic or not

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(case1101)
attach(case1101)

## EXPLORATION
library(lattice)
xyplot(Metabol~Gastric|Sex*Alcohol, case1101)

myPch <- ifelse(Sex=="Female",24,21)
myBg  <- ifelse(Alcohol=="Alcoholic","gray","white")
plot(Metabol~Gastric, pch=myPch,bg=myBg,cex=1.5)
legend(1,12, pch=c(24,24,21,21), pt.cex=c(1.5,1.5,1.5,1.5),
  pt.bg=c("white","gray", "white", "gray"),
  c("Non-alcoholic Females", "Alcoholic Females",
  "Non-alcoholic Males", "Alcoholic Males"))
identify(Metabol ~ Gastric)
# Left click on outliers to show case number; Esc when finished.

myLm1 <- lm(Metabol ~ Gastric + Sex + Gastric:Sex)
plot(myLm1, which=1)
plot(myLm1, which=4) # Show Cook's Distance; note cases 31 and 32.
plot(myLm1, which=5) # Note leverage and studentized residual for cases 31 and 32.
subject  <- 1:32  # Create ID number from 1 to 32

# Refit model without cases 31 and 32:
myLm2 <- update(myLm1, ~ ., subset = (subject !=31 & subject !=32))
plot(myLm2,which=1)
plot(myLm2,which=4)
plot(myLm2,which=5)
summary(myLm1)
summary(myLm2) # Significance of interaction terms hinges on cases 31 and 32.

myLm3 <- update(myLm2, ~ . - Gastric:Sex) #Drop interaction (without 31,32).
summary(myLm3)
if(require(car)){   # Use the car library
crPlots(myLm3) # Show partial residual (component + residual) plots.
}

## INFERENCE AND INTERPRETATION
summary(myLm3)
confint(myLm3,2:3)

## DISPLAY FOR PRESENTATION
myCol <- ifelse(Sex=="Male","blue","red")
plot(Metabol ~ Gastric,
  xlab=expression("Gastric Alcohol Dehydrogenase Activity in Stomach ("*mu*"mol/min/g of Tissue)"),
  ylab="First-pass Metabolism in the Stomach (mmol/liter-hour)",
  main="First-Pass Alcohol Metabolism and Enzyme Activity for 18 Females and 14 Males",
  pch=myPch, bg=myBg,cex=1.75, col=myCol, lwd=1)
legend(0.8,12.2, c("Females", "Males"), lty=c(1,2),
    pch=c(24,21), pt.cex=c(1.75,1.75), col=c("red", "blue"))
dummyGastric <- seq(min(Gastric),3,length=100)
beta <- myLm3$coef
curveF <- beta[1] + beta[2]*dummyGastric
curveM <- beta[1] + beta[2]*dummyGastric + beta[3]
lines(curveF ~ dummyGastric, col="red")
```

```
lines(curveM ~ dummyGastric, col="blue",lty=2)
text(.8,10,"gray indicates alcoholic",cex = .8, adj=0)

detach(case1101)
```

---

case1102 *The Blood–Brain Barrier*

---

## Description

The human brain is protected from bacteria and toxins, which course through the blood–stream, by a single layer of cells called the blood–brain barrier. These data come from an experiment (on rats, which process a similar barrier) to study a method of disrupting the barrier by infusing a solution of concentrated sugars.

## Usage

```
case1102
```

## Format

A data frame with 34 observations on the following 9 variables.

**Brain** Brain tumor count (per gm)

**Liver** Liver count (per gm)

**Time** Sacrifice time (in hours)

**Treatment** Treatment received

**Days** Days post inoculation

**Sex** Sex of the rat

**Weight** Initial weight (in grams)

**Loss** Weight loss (in grams)

**Tumor** Tumor weight (in $10^{-4}$ grams)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex1416, ex1417

## Examples

```
str(case1102)
attach(case1102)

## EXPLORATION
logRatio <- log(Brain/Liver)
logTime <- log(Time)
myMatrix <- cbind (logRatio, Days, Weight, Loss, Tumor, logTime)
if(require(car)){   # Use the car library
scatterplotMatrix(myMatrix,groups=Treatment,
  smooth=FALSE, diagonal="histogram", col=c("green","blue"), pch=c(16,17), cex=1.5)
}

myLm1 <- lm(logRatio ~ Treatment + logTime + Days + Sex + Weight + Loss + Tumor)
plot(myLm1, which=1)
if(require(car)){   # Use the car library
  crPlots(myLm1) # Draw partial resdual plots.
}

myLm2   <-  lm(logRatio ~ Treatment + factor(Time) +
  Days + Sex + Weight + Loss + Tumor)  # Include Time as a factor.
anova(myLm1,myLm2)
if(require(car)){   # Use the car library
  crPlots(myLm2) # Draw partial resdual plots.
}

summary(myLm2)  # Use backard elimination
myLm3 <- update(myLm2, ~ . - Days)
summary(myLm3)
myLm4 <- update(myLm3, ~ . - Sex)
summary(myLm4)
myLm5 <- update(myLm4, ~ . - Weight)
summary(myLm5)
myLm6 <- update(myLm5, ~ . - Tumor)
summary(myLm6)
myLm7 <- update(myLm6, ~ . - Loss)
summary(myLm7)   # Final model for inference


## INFERENCE AND INTERPRETATION
myTreatment <- factor(Treatment,levels=c("NS","BD")) # Change level ordering
myLm7a  <- lm(logRatio ~  factor(Time) + myTreatment)
summary(myLm7a)
beta <- myLm7a$coef
exp(beta[5])
exp(confint(myLm7a,5))
# Interpetation: The median ratio of brain to liver tumor counts for barrier-
# disrupted rats is estimated to be 2.2 times the median ratio for control rats
# (95% CI: 1.5 times to 3.2 times as large).

## DISPLAY FOR PRESENTATION
ratio <- Brain/Liver
jTime <- exp(jitter(logTime,.2)) # Back-transform a jittered version of logTime
plot(ratio ~ jTime, log="xy",
  xlab="Sacrifice Time (Hours), jittered; Log Scale",
  ylab="Effectiveness: Brain Tumor Count Relative To Liver Tumor Count; Log Scale",
```

```
  main="Blood Brain Barrier Disruption Effectiveness in 34 Rats",
  pch= ifelse(Treatment=="BD",21,24), bg=ifelse(Treatment=="BD","green","orange"),
  lwd=2, cex=2)
dummyTime     <- c(0.5, 3, 24, 72)
controlTerm   <- beta[1] + beta[2]*(dummyTime==3) +
  beta[3]*(dummyTime==24) + beta[4]*(dummyTime==72)
controlCurve  <- exp(controlTerm)
lines(controlCurve ~ dummyTime, lty=1,lwd=2)
BDTerm        <- controlTerm + beta[5]
BDCurve       <- exp(BDTerm)
lines(BDCurve ~ dummyTime,lty=2,lwd=2)
legend(0.5,10,c("Barrier disruption","Saline control"),pch=c(21,22),
  pt.bg=c("green","orange"),pt.lwd=c(2,2),pt.cex=c(2,2), lty=c(2,1),lwd=c(2,2))

detach(case1102)
```

---

`case1201` *State Average SAT Scores*

---

## Description

Data on the average SAT scores for US states in 1982 and possible associated factors.

## Usage

```
case1201
```

## Format

A data frame with 50 observations on the following 8 variables.

**State** US state

**SAT** state averages of the total SAT (verbal + quantitative) scores

**Takers** the percentage of the total eligible students (high school seniors) in the state who took the exam

**Income** the median income of families of test–takers (in hundreds of dollars)

**Years** the average number of years that the test–takers had formal studies in social sciences, natural sciences and humanities

**Public** the percentage of the test–takers who attended public secondary schools

**Expend** the total state expenditure on secondary schools (in hundreds of dollars per student)

**Rank** the median percentile ranking of the test–takers within their secondary school classes

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(case1201)
attach(case1201)

## EXPLORATION
logTakers  <- log(Takers)
myMatrix   <- cbind(SAT, logTakers,Income, Years, Public, Expend, Rank)
if(require(car)){   # Use the car library
scatterplotMatrix(myMatrix, diagonal="histogram", smooth=FALSE)
  }
State[Public < 50] # Identify state with low Public (Louisiana)
State[Expend > 40] # Alaska
myLm1    <- lm(SAT ~ logTakers + Income+ Years + Public + Expend + Rank)
plot(myLm1,which=1)
plot(myLm1,which=4)  # Cook's Distance
State[29] # Identify State number 29?  ([1] Alaska)
plot(myLm1,which=5)
if(require(car)){   # Use the car library
  crPlots(myLm1)  # Partial residual plot
}
myLm2 <- update(myLm1, ~ . ,subset=(State != "Alaska"))
plot(myLm2,which=1)
plot(myLm2,which=4)
if(require(car)){   # Use the car library
  crPlots(myLm2) # Partial residual plot
}
## RANK STATES ON SAT SCORES, ADJUSTED FOR Takers AND Rank
myLm3        <- lm(SAT ~ logTakers + Rank)
myResiduals  <- myLm3$res
myOrder      <- order(myResiduals)
State[myOrder]

## DISPLAY FOR PRESENTATION
dotchart(myResiduals[myOrder], labels=State[myOrder],
  xlab="SAT Scores, Adjusted for Percent Takers and HS Ranks (Deviation From Average)",
  main="States Ranked by Adjusted SAT Scores",
  bg="green", cex=.8)
abline(v=0, col="gray")

## VARIABLE SELECTION (FOR RANKING STATES AFTER ACCOUNTING FOR ALL VARIABLES)
expendSquared <- Expend^2
if(require(leaps)){   # Use the leaps library
  mySubsets   <- regsubsets(SAT ~ logTakers + Income+ Years + Public + Expend +
    Rank + expendSquared, nvmax=8, data=case1201, subset=(State != "Alaska"))
  mySummary <- summary(mySubsets)
  p <- apply(mySummary$which, 1, sum)
  plot(p, mySummary$bic, ylab = "BIC")
  cbind(p,mySummary$bic)
  mySummary$which[4,]
  myLm4 <- lm(SAT ~ logTakers + Years + Expend + Rank, subset=(State != "Alaska"))
  summary(myLm4)

## DISPLAY FOR PRESENTATION
  myResiduals2 <- myLm4$res
  myOrder2 <- order(myResiduals2)
  newState <- State[State != "Alaska"]
```

```
  newState[myOrder2]
  dotchart(myResiduals2[myOrder2], labels=State[myOrder2],
    xlab="Adjusted SAT Scores (Deviation From Average Adjusted Value)",
    main=paste("States Ranked by SAT Scores Adjusted for Demographics",
               "of Takers and Education Expenditure", sep = " "),
    bg="green", cex = .8)
  abline(v=0, col="gray")
}

detach(case1201)
```

---

`case1202` *Sex discrimination in Employment*

---

## Description

Data on employees from one job category (skilled, entry–level clerical) of a bank that was sued for sex discrimination. The data are on 32 male and 61 female employees, hired between 1965 and 1975.

## Usage

```
case1202
```

## Format

A data frame with 93 observations on the following 7 variables.

**Bsal** Annual salary at time of hire

**Sal77** Salary as of March 1975

**Sex** Sex of employee

**Senior** Seniority (months since first hired)

**Age** Age of employee (in months)

**Educ** Education (in years)

**Exper** Work experience prior to employment with the bank (months)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Roberts, H.V. (1979). Harris Trust and Savings Bank: An Analysis of Employee Compensation, *Report 7946*, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

## See Also

`case0102`

## Examples

```
str(case1202)
attach(case1202)

## EXPLORATION
logSal <- log(Bsal)
myMatrix <- cbind (logSal, Senior,Age, Educ, Exper)
if(require(car)){   # Use the car library
  scatterplotMatrix(myMatrix, smooth=FALSE, diagonal="histogram",
                    groups=Sex, col=c("red","blue") )
}
myLm1 <- lm(logSal ~ Senior + Age + Educ + Exper + Sex)
plot(myLm1, which=1)
plot(myLm1, which=4) #  Cook's Distance
if(require(car)){   # Use the car library
  crPlots(myLm1)    # Partial residual plots
}
ageSquared    <- Age^2
ageCubed      <- Age^3
experSquared  <- Exper^2
experCubed    <- Exper^3
myLm2 <- lm(logSal ~ Senior + Age + ageSquared  + ageCubed +
  Educ + Exper + experSquared + experCubed  + Sex)
plot(myLm2, which=1)  # Residual plot
plot(myLm1, which=4)  # Cook's distance

if(require(leaps)){   # Use the leaps library
  mySubsets     <- regsubsets(logSal ~ (Senior + Age + Educ + Exper +
    ageSquared  + experSquared)^2, nvmax=25, data=case1202)
  mySummary  <- summary(mySubsets)
  p  <- apply(mySummary$which, 1, sum)
  plot(mySummary$bic ~ p, ylab = "BIC")
  cbind(p,mySummary$bic)
  mySummary$which[8,]  # Note that Age:ageSquared = ageCubed
}
myLm3         <- lm(logSal ~ Age + Educ + ageSquared + Senior:Educ +
  Age:Exper + ageCubed + Educ:Exper + Exper:ageSquared)
summary(myLm3)

myLm4 <- update(myLm3, ~ . + Sex)
summary(myLm4)
myLm5 <- update(myLm4, ~ . + Sex:Age + Sex:Educ + Sex:Senior +
  Sex:Exper + Sex:ageSquared)
anova(myLm4, myLm5)

## INFERENCE AND INTERPRETATION
summary(myLm4)
beta          <- myLm4$coef
exp(beta[6])
exp(confint(myLm4,6))
# Conclusion:  The median beginning salary for males was estimated to be 12%
# higher than the median salary for females with similar values of the available
# qualification variables (95% confidence interval: 7% to 17% higher).

## DISPLAY FOR PRESENTATION
years <- Exper/12  # Change months to years
```

```
plot(Bsal ~ years, log="y", xlab="Previous Work Experience (Years)",
  ylab="Beginning Salary (Dollars); Log Scale",
  main="Beginning Salaries and Experience for 61 Female and 32 Male Employees",
  pch= ifelse(Sex=="Male",24,21), bg = "gray",
  col= ifelse(Sex=="Male","blue","red"), lwd=2, cex=1.8 )
myLm6 <- lm(logSal ~ Exper + experSquared + experCubed + Sex)
beta <- myLm6$coef
dummyExper <- seq(min(Exper),max(Exper),length=50)
curveF <- beta[1] + beta[2]*dummyExper + beta[3]*dummyExper^2 +
  beta[4]*dummyExper^3
curveM <- curveF + beta[5]
dummyYears <- dummyExper/12
lines(exp(curveF) ~ dummyYears, lty=1, lwd=2,col="red")
lines(exp(curveM) ~ dummyYears, lty = 2, lwd=2, col="blue")
legend(28,8150, c("Male","Female"),pch=c(24,21), pt.cex=1.8, pt.lwd=2,
  pt.bg=c("gray","gray"), col=c("blue","red"), lty=c(2,1), lwd=2)

detach(case1202)
```

---

`case1301` *Seaweed Grazers*

---

## Description

To study the influence of ocean grazers on regeneration rates of seaweed in the intertidal zone, a researcher scraped rock plots free of seaweed and observed the degree of regeneration when certain types of seaweed-grazing animals were denied access. The grazers were limpets (L), small fishes (f) and large fishes (F). Each plot received one of six treatments named by which grazers were allowed access. In addition, the researcher applied the treatments in eight blocks of 12 plots each. Within each block she randomly assigned treatments to plots. The blocks covered a wide range of tidal conditions.

## Usage

```
case1301
```

## Format

A data frame with 96 observations on the following 3 variables.

**Cover** percent of regenerated seaweed cover

**Block** a factor with levels `"B1"`, `"B2"`, `"B3"`, `"B4"`, `"B5"`, `"B6"`, `"B7"` and `"B8"`

**Treat** a factor indicating treatment, with levels `"C"`, `"f"`, `"fF"`, `"L"`, `"Lf"` and `"LfF"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Olson, A. (1993). Evolutionary and Ecological Interactions Affecting Seaweeds, Ph.D. Thesis. Oregon State University.

## Examples

```
str(case1301)
attach(case1301)

## EXPLORATION AND MODEL DEVELOPMENT
plot(Cover ~ Treat,xlab="Animals Present",ylab="Remaining Seaweed Coverage (%)")
myLm1 <- lm(Cover ~ Block + Treat + Block:Treat)
plot(myLm1,which=1)
ratio <- Cover/(100 - Cover)
logRatio  <- log(ratio)
myLm2 <- lm(logRatio ~ Block + Treat + Block:Treat)
plot(myLm2, which=1)
myLm3  <- lm(logRatio ~ Block + Treat)
anova(myLm3, myLm2) # Test for interaction with extra ss F-test
if(require(car)){   # Use the car library
  crPlots(myLm3)   # Partial residual plots
  myLm4  <- lm(logRatio ~ Treat)
  anova(myLm4, myLm3)     # Test for Block effect
  myLm5  <- lm(logRatio ~ Block)
  anova(myLm5, myLm3)     # Test for Treatment effect
  lmp <- factor(ifelse(Treat %in% c("L", "Lf", "LfF"), "yes", "no"))
  sml <- factor(ifelse(Treat %in% c("f", "fF", "Lf", "LfF"), "yes","no"))
  big <- factor(ifelse(Treat %in% c("fF", "LfF"), "yes","no"))
  myLm6  <- lm(logRatio ~ Block + lmp + sml + big)
  crPlots(myLm6)
  myLm7  <- lm(logRatio ~ Block + (lmp + sml + big)^2)
  anova(myLm6, myLm7)  # Test for interactions of lmp, sml, and big

## INFERENCE AND INTERPRETATION
  summary(myLm6)   # Get p-values for lmp, sml, and big effects; R^2 = .8522
  beta <- myLm6$coef
  exp(beta[9:11])
  exp(confint(myLm6,9:11) )

  myLm7 <- update(myLm6, ~ . - lmp)
  summary(myLm7)  # R^2 = .4568; Note .8522-.4580 = 39.54# (explained by limpets)
  myLm8 <- update(myLm6, ~ . - big)
  summary(myLm8)  # R^2 = .8225; Note .8522-.8255= 2.97# (explained by big fish)
  myLm9 <- update(myLm6, ~ . - sml)
  summary(myLm9)  # R^2: .8400; Note .8522-.8400 = 1.22# (explained by small fish)

## DISPLAY FOR PRESENTATION
  myYLab <- "Adjusted Seaweed Regeneration (Log Scale; Deviation from Average)"
  crPlots(myLm6, ylab=myYLab, ylim=c(-2.2,2.2),
   main="Effects of Blocks and Treatments on Log Regeneration Ratio, Adjusted for Other Factors")
}

detach(case1301)
```

---

## case1302 *Pygmalion Effect*

---

## Description

One company of soldiers in each of 10 platoons was assigned to a Pygmalion treatment group, with remaining companies in the platoon assigned to a control group. Leaders of the Pygmalion platoons were told their soldiers had done particularly well on a battery of tests which were, in fact, non-existent. In this randomised block experiment, platoons are experimental units, companies are blocks, and average Practical Specialty test score for soldiers in a platoon is the response. The researchers wished to see if the platoon response was affected by the artificially-induced expectations of the platoon leader.

## Usage

```
case1302
```

## Format

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

**Company** a factor indicating company identification, with levels `"C1"`, `"C2"`, ..., `"C10"`

**Treat** a factor indicating treatment with two levels, `"Pygmalion"` and `"Control"`

**Score** average score on practical specialty test of all soldiers in the platoon

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Eden, D. (1990). Pygmalion Without Interpersonal Contrast Effects: Whole Groups Gain from Raising Manager Expectations, *Journal of Applied Psychology* **75**(4): 395–398.

## Examples

```
str(case1302)
attach(case1302)

## EXPLORATION AND MODEL DEVELOPMENT
plot(Score ~ as.numeric(Company),cex=1.5, pch=21,
  bg=ifelse(Treat=="Pygmalion","blue","light gray"))
myLm1   <- lm(Score ~ Company + Treat + Company:Treat) # Fit with interaction.
plot(myLm1,which=1)  # Plot residuals.
myLm2   <- update(myLm1, ~ . - Company:Treat) # Refit, without interaction.
anova(myLm2, myLm1)  # Show extra-ss-F-test p-value (for interaction effect).
if(require(car)){   # Use the car library
  crPlots(myLm2)
}

## INFERENCE
myLm3 <- update(myLm2, ~ . - Company)  # Fit reduced model without Company.
anova(myLm3, myLm2)   # Test for Company effect.
summary(myLm2)   # Show estimate and p-value for Pygmalion effect.
confint(myLm2,11)  # Show 95% CI for Pygmalion effect.

## DISPLAY FOR PRESENTATION
beta        <- myLm2$coef
```

```
partialRes  <- myLm2$res + beta[11]*ifelse(Treat=="Pygmalion",1,0) # partial res
boxplot(partialRes ~ Treat,  # Boxplots of partial residuals for each treatment
  ylab="Average Test Score, Adjusted for Company Effect (Deviation from Company Average)",
   names=c("19 Control Platoons","10 Pygmalion Treated Platoons"),
   col="green", boxlwd=2, medlwd=2,whisklty=1, whisklwd=2, staplewex=.2,
   staplelwd=2, outlwd=2, outpch=21, outbg="green", outcex=1.5 )

detach(case1302)
```

---

`case1401` *Chimp Learning Times*

---

## Description

Researchers taught each of 4 chimps to learn 10 words in American sign language and recorded the learning time for each word for each chimp. They wished to describe chimp differences and word differences.

## Usage

```
case1401
```

## Format

A data frame with 40 observations on the following 4 variables.

**Minutes** learning time in minutes

**Chimp** a factor indicating chimp, with four levels `"Booee"`, `"Cindy"`, `"Bruno"` and `"Thelma"`

**Sign** a factor indicating word taught, with 10 levels

**Order** the order in which the sign was taught

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Fouts, R.S. (1973). Acquisition and Testing of Gestural Signs in Four Young Chimpanzees, *Science* **180**: 978–980.

## Examples

```
str(case1401)
attach(case1401)

## EXPLORATION AND MODEL DEVELOPMENT
plot(Minutes ~ Sign)
myLm1 <- lm(Minutes ~ Chimp + Sign)
plot(myLm1,which=1)  # Plot residuals (indicates a need for transformation).
logMinutes <- log(Minutes)
myLm2 <- lm(logMinutes ~ Chimp + Sign)
plot(myLm2, which=1)  # This looks fine.
```

```
if(require(car)){   # Use the car library
  crPlots(myLm2) # Partial residual plots
}

## INFERENCE AND INTERPRETATION
myLm3  <- update(myLm2, ~ . - Chimp)  # Fit reduced model without Chimp.
anova(myLm3, myLm2)   # Test for Chimp effect.
myLm4 <- update(myLm2, ~ . - Sign)  # Fit reduced model without Sign.
anova(myLm4, myLm2)    # Test for Sign effect.
# Fit 2-way model without intercept to order signs from easiest to hardest
myAov1 <- aov(logMinutes ~ Sign + Chimp - 1)
sort(myAov1$coef[1:10]) # Show the ordering of Signs
orderedSign <- factor(Sign,levels=c("listen","drink","shoe","key","more",
      "food","fruit","hat","look","string") )  # Re-order signs, easiest 1st
myAov2 <- aov(logMinutes ~ orderedSign + Chimp - 1)  # Refit
opar <- par(no.readonly=TRUE)  # Store current graphics parameters settings
par(mar=c(4.1,7.1,4.1,2.1)) # Adjust margins to allow room for y-axis labels

## takes too long to run
if(require(multcomp)){   # Use the multcomp library
  myMultComp    <- glht(myAov2, linfct = mcp(orderedSign = "Tukey"))
  plot(myMultComp)  # Plot Tukey-adjusted confidence intervals.
  summary(myMultComp)   # Show Tukey-adjusted p-values  pairwise comparisons
  confint(myMultComp)   # Show Tukey-adjusted 95% confidence intervals
}

par(opar) # Restore original graphics parameters settings

## DISPLAY FOR PRESENTATION
myYLab <- "Log Learning Time, Adjusted for Chimp Effect"
myXLab <- "Sign Learned"
if(require(car)){   # Use the car library
crPlots(myAov2, ylab=myYLab, xlab=myXLab,
  main="Learning Times by Sign, Adjusted for Chimp Effects",
  layout=c(1,1)) # Click on graph area to show next page (Just use first one.)
}

detach(case1401)
```

---

`case1402` *Effect of Ozone, SO2 and Drought on Soybean Yield*

---

## Description

In a completely randomized design with a 2x3x5 factorial treatment structure, researchers randomly assigned one of 30 treatment combinations to open-topped growing chambers, in which two soybean cultivars were planted. The responses for each chamber were the yields of the two types of soybean.

## Usage

```
case1402
```

## Format

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

**Stress** a factor indicating treatment, with two levels "Well-watered" and "Stressed"

**SO2** a quantitative treatment with three levels 0, 0.02 and 0.06

**O3** a quantitative treatment with five levels 0.02, 0.05, 0.07, 0.08 and 0.10

**Forrest** the yield of the Forrest cultivar of soybean (in kg/ha)

**William** the yield of the Williams cultivar of soybean (in kg/ha)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Heggestad, H.E. and Lesser, V.M. (1990). Effects of Chronic Doses of Sulfur Dioxide, Ozone, and Drought on Yields and Growth of Soybeans Under Field Conditions, *Journal of Environmental Quality* **19**: 488–495.

## Examples

```
str(case1402)
attach(case1402)

## EXPLORATION AND MODEL DEVELOPMENT; FORREST CULTIVAR
logForrest <- log(Forrest)
# Fit model without interactions first--to examine partial residual plots.
myLm1 <- lm(logForrest ~ Stress + SO2 + O3)
if(require(car)){  # Use the car library
  crPlots(myLm1)   # Partial res plots => linear effects of SO2 and O3 look ok.
}
myLm2 <- lm(logForrest ~ (Stress + SO2  + O3)^2) # all 2-factor interactions.
plot(myLm1, which=1)   # Residual plot looks ok.
anova(myLm1,myLm2) # Test for interactive effects.

## INFERENCE AND INTERPRETATION; FORREST CULTIVAR
summary(myLm1)
betaF  <- myLm1$coef
# Effect of 0.01 increase in SO2 (note coeff is negative):
100*(1 - exp(0.01*betaF[3]))
#1.655701;   a 1.65% decrease in median yield
100*(1-exp(0.01*confint(myLm1,"SO2")))
#3.772277 -0.5074294: 95% onfidence interval for effect of 0.01 increase in SO2
# Effect of 0.01 increase in O3 (note coeff is negative):
100*(1 - exp(0.01*betaF[4]))
# 5.585979   I.e. a 5.6% decrease in median yield
100*(1-exp(0.01*confint(myLm1,"O3")))
#7.445966 3.688613: 95% confidence interval for effect of 0.01 increase in O3
# Effect of water stress (note coeff is positive for effect of well-watered):
100*(1 - exp(-betaF[2]))  # Get estimate for negative of this beta
#3.220556:  a 3.2% decrease in median yield due to water stress
100*(1-exp(-confint(myLm1,2)))
#-7.875521 13.17529: 95% confidence interval
```

```
## DISPLAY FOR PRESENTATION; FORREST CULTIVAR
jO3     <- jitter(O3,factor=.25) # Jitter for plotting; jittering factor 0.25.
jS      <- jitter(SO2,factor=.25)  # Jitter SO2 for plotting.
cexfac  <- 1.75  # Use character expansion factor of 1.75 for plotting symbols.
opar <- par(no.readonly=TRUE)  # Store current graphics parameters settings
par(mfrow=c(1,2))  # Make a panel of 2 plots in 1 row.
myPointCode  <- ifelse(Stress=="Well-watered",21,24)
myPointColor <- ifelse(Stress=="Well-watered","green","orange")
par(mar=c(4.1,4.1,2.1,0.1))
plot(Forrest ~ jO3, log="y", ylab="Yield of Forrest Cultivar (kg/ha)",
xlab=expression(paste(italic("Ozone ("),mu,"L/L), jittered")),
  pch=myPointCode, lwd=2, bg=myPointColor, cex=cexfac)
legend(.02,2400, c("Well-watered","Water Stressed"), pch=c(21,24),
  pt.cex=cexfac, pt.bg=c("green","orange"), pt.lwd=2, lty=c(3,1), lwd=c(2,2))
dummyO    <- seq(min(O3), max(O3), length=2)
curve1    <- exp(betaF[1] + betaF[3]*mean(SO2) + betaF[4]*dummyO)
curve2    <- exp(betaF[1] + betaF[2] + betaF[3]*mean(SO2)+ betaF[4]*dummyO)
lines(curve1 ~ dummyO,lwd=2)
lines(curve2 ~ dummyO,lwd=2,lty=3)

# PLOT FORREST VS SO2
par(mar=c(4.1,2.1,2.1,2.1)) # Change margins
plot(Forrest ~ jS, log="y", ylab="",
xlab=expression(paste(italic("Sulfur Dioxide ("),mu,"L/L), jittered")),
  yaxt="n", pch=myPointCode, lwd=2, bg=myPointColor, cex=cexfac)
dummyS    <- seq(min(SO2),max(SO2),length=2)
curve1    <- exp(betaF[1] + betaF[3]*dummyS + betaF[4]*mean(O3))
curve2    <- exp(betaF[1] + betaF[2] + betaF[3]*dummyS + betaF[4]*mean(O3))
lines(curve1 ~ dummyS,lwd=2)
lines(curve2 ~ dummyS,lwd=2,lty=3)
par(opar) # Restore previous graphics parameter settings

detach(case1402)
```

---

`case1501` *Logging and Water Quality*

---

## Description

Data from an observational study of nitrate levels measured at three week intervals for five years in two watersheds. One of the watersheds was undisturbed and the other had been logged with a patchwork pattern.

## Usage

```
case1501
```

## Format

A data frame with 88 observations on the following 3 variables.

**Week** week after the start of the study

**Patch** natural logarithm of nitrate level (ppm) in the logged watershed (ppm)

**NoCut** natural logarithm of nitrate level in the undisturbed watershed (ppm)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learnings.

## References

Harr, R.D., Friderksen, R.L., and Rothacher, J. (1979). Changes in Streamflow Following Timber Harvests in Southwestern Oregon, USDA/USFS Research Paper PNW-249, Pacific NW Forest and Range Experiment Station, Portland, Oregon.

## Examples

```
str(case1501)
attach(case1501)

## EXPLORATION
opar <- par(no.readonly=TRUE)  # Store current graphics parameters settings
par(mfrow=c(2,1))   # Set graphics parameters: 2 row, 1 column layout
plot(NoCut ~ Week,  type="b", ylab="Log of Nitrate Concentration; NoCut")
abline(h=mean(NoCut))  # Horizontal line at the mean
plot(Patch ~ Week,  type="b", ylab="Log of Nitrate Concentration; Patch Cut")
abline(h=mean(Patch))

par(opar) # Restore previous graphics settings
lag.plot(NoCut,do.lines=FALSE)  # Lag plot for NoCut
lag.plot(Patch,do.lines=FALSE)  # Lag plot for Patch
pacf(NoCut)  # partial autocorrelation function plot; noCut
pacf(Patch)  # partial autocorrelation function plot; Patch

## INFERENCE  (2-sample comparison, accounting for first serial correlation)
diff     <- mean(Patch) - mean(NoCut)
nPatch   <- length(Patch)  # length of Patch series
nNoCut   <- length(NoCut)   # length of NoCut series
acfPatch <- acf(Patch, type="covariance")  # auto covariances for Patch series
c0Patch  <- acfPatch$acf[1]*nPatch/(nPatch-1) # variance; n-1 divisor (Patch)
c1Patch  <- acfPatch$acf[2]*nPatch/(nPatch-1) # autocov; n-1 divisor (Patch)
acfNoCut <- acf(NoCut, type="covariance") # auto covariances for NoCut series
c0NoCut  <- acfNoCut$acf[1]*nNoCut/(nNoCut - 1) # variance; n-1 divisor (NoCut)
c1NoCut  <- acfNoCut$acf[2]*nNoCut/(nNoCut - 1) # autocov; n-1 divisor (NoCut)
dfPatch  <- nPatch - 1     # DF (n-1); Patch
dfNoCut  <- nNoCut - 1     # DF (n-1); NoCut

c0Pooled   <- (dfPatch*c0Patch + dfNoCut*c0NoCut)/(dfPatch + dfNoCut)
c0Pooled   #[1] 1.413295  = pooled estimate of variance
c1Pooled   <- (dfPatch*c1Patch + dfNoCut*c1NoCut)/(dfPatch + dfNoCut)
c1Pooled   #[1] 0.9103366 = pooled estimate of lag 1 covariance

# Pooled estimate of first serial correlation coefficient:
r1 <- c1Pooled/c0Pooled                  #[1] 0.6441233
SEdiff  <- sqrt((1 + r1)/(1-r1))*sqrt(c0Pooled*(1/nPatch + 1/nNoCut))

# t-test and confidence interval
tStat      <- diff/SEdiff #[1] 0.2713923
pValue     <- 1 - pt(tStat,dfPatch + dfNoCut)     # One-sided p-value
halfWidth  <- qt(.975,dfPatch + dfNoCut)*SEdiff   # half width of 95% CI
diff + c(-1,1)*halfWidth  #95% CI -0.6557578  0.8648487
```

```
## GRAPHICAL DISPLAY FOR PRESENTATION
par(mfrow=c(1,1))                  # Reset mfrow to a single plot per page
plot(exp(Patch) ~ Week, # Use exp(Patch) to show results in original units
  log="y", type="b", xlab="Weeks After Logging",
  ylab="Nitrate Concentration in Watershed Runoff (ppm)",
  main="Nitrate Series in Patch-Cut and Undisturbed Watersheds",
  pch=21, col="dark green", lwd=3, bg="green", cex=1.3 )
points(exp(NoCut) ~ Week, pch=24, col="dark blue", lwd=3, bg="orange",cex=1.3)
lines(exp(NoCut) ~ Week, lwd=3, col="dark blue",lty=3)
abline(h=exp(mean(Patch)),col="dark green",lwd=2)
abline(h=exp(mean(NoCut)),col="dark blue", lwd=2,lty=2)
legend(205,100,legend=c("Patch Cut", "Undisturbed"),
  pch=c(21,24), col=c("dark green","dark blue"), pt.bg = c("green","orange"),
  pt.cex=c(1.3,1.3), lty=c(1,3), lwd=c(3,3))
text(-1, 8.5, "Mean",col="dark green")
text(-1,6.3,"Mean", col="dark blue")


detach(case1501)
```

---

`case1502` *Global Warming*

---

## Description

The data are the temperatures (in degrees Celsius) averaged for the northern hemisphere over a full year, for years 1850 to 2010. The 161-year average temperature has been subtracted, so each observation is the temperature difference from the series average.

## Usage

```
case1502
```

## Format

A data frame with 161 observations on the following 2 variables.

**Year** year in which yearly average temperature was computed, from 1850 to 2010

**Temperature** northern hemisphere temperature minus the 161-year average (degrees Celsius)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Jones, P.D., D. E. Parker, T. J. Osborn, and K. R. Briffa, (2011) Global and Hemispheric Temperature Anomalies and and Marine Instrumental Records, CDIAC, http://cdiac.ornl.gov/trends/temp/jonescru/jones.html, August 4, 2011.

## See Also

ex1519

## Examples

```
str(case1502)
attach(case1502)

## EXPLORATION AND MODEL BUILDING
plot(Temperature ~ Year, type="b")  # Type = "b" for *both* points and lines

yearSquared <- Year^2
yearCubed <- Year^3
myLm1 <- lm(Temperature ~ Year + yearSquared + yearCubed)
res1 <- myLm1$res
myPacf <- pacf(res1)  # Partial autocorrelation from residuals
r1 <- myPacf$acf[1]  #First serial correlation coefficient
n <- length(Temperature)   # Series length = 161
v <- Temperature[2:n]  -  r1*Temperature[1:(n-1)]   # Filtered response
ones <- rep(1-r1, n-1)   # make a variable of all 1's
u1 <- Year[2:n]          -  r1*Year[1:(n-1)]   # Filtered "ones"
u2 <- yearSquared[2:n]  -  r1*yearSquared[1:(n-1)]  # Filtered X1
u3 <- yearCubed[2:n]    -  r1*yearCubed[1:(n-1)]  # Filtered X2
myLm2 <- lm(v ~  u1 + u2 + u3 )
res2 <- myLm2$res
pacf(res2)   # Looks fine; don't worry about lag 4 marginal significance
plot(myLm2, which=1)  # Residual plot
summary(myLm2) # Cubic term isn't needed.
myLm3    <- update(myLm2, ~ . - u3) # Drop cubic term

## INFERENCE
summary(myLm3) # Everything remaining is statistically significant.

## GRAPHICAL DISPLAY FOR PRESENTATION
plot(Temperature ~ Year, xlab="Year",
  ylab=expression(paste("Annual Average Temperature (Difference From Average), ",
  degree,"C")),main="Annual Average Temperature in Northern Hemisphere; 1850-2010",
  type="b", pch=21, lwd=2, bg="green", cex=1.5)
myFits <- myLm3$fit
lines(myFits ~ Year[2:161], col="blue", lwd=2)
legend(1850,0.6,"Quadratic Regression Fit, Adjusted for AR(1) Serial Correlation",
  col="blue", lwd=2, box.lty=0)

detach(case1502)
```

---

# case1601 *Sites of Short- and Long-Term Memory*

---

## Description

Researchers taught 18 monkeys to distinguish each of 100 pairs of objects, 20 pairs each at 16, 12, 8, 4, and 2 weeks prior to a treatment. After this training, they blocked access to the hippocampal

formation in 11 of the monkeys. All monkeys were then tested on their ability to distinguish the objects. The five-dimensional response for each monkey is the number of correct objects distinguished among those taught at 16, 12, 8, 4, and 2 weeks prior to treatment.

## Usage

```
case1601
```

## Format

A data frame with 18 observations on the following 7 variables.

**Monkey** Monkey name

**Treatment** a treatment factor with levels `"Control"` and `"Treated"`

**Week2** percentage of 20 objects taught 2 weeks prior to treatment that were correctly distinguished in the test

**Week4** percentage of 20 objects taught 4 weeks prior to treatment that were correctly distinguished in the test

**Week8** percentage of 20 objects taught 8 weeks prior to treatment that were correctly distinguished in the test

**Week12** percentage of 20 objects taught 12 weeks prior to treatment that were correctly distinguished in the test

**Week16** percentage of 20 objects taught 16 weeks prior to treatment that were correctly distinguished in the test

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Sola-Morgan, S. M. and Squire, L. R. (1990). The Primate Hippocampal Formation: Evidence for a Time-limited Role in Memory Storage, *Science* **250**: 288–290.

## Examples

```
str(case1601)
attach(case1601)

## EXPLORATION
short <- (Week2 + Week4)/2
long  <- (Week8 + Week12 + Week16)/3
myPointCode <- ifelse(Treatment=="Control",15,16)
myPointColor <- ifelse(Treatment=="Control","orange","green")
plot(long ~ short, pch=myPointCode, col=myPointColor, cex=2)
abline(h=mean(long),lty=2)
abline(v=mean(short),lty=2)
identify(short,long,labels=Monkey)   # Identify outliers; press Esc when done

## INFERENCE USING HOTELLING's T-SQUARED TEST
myLm1     <- lm(cbind(short,long) ~ Treatment)  # Full model
myLm2     <- lm(cbind(short,long) ~ 1) # Reduced model, with only intercept
```

```
anova(myLm2, myLm1, test="Hotelling") # p-value for Treatment effect
# confidence intervals
n1  <- sum(Treatment=="Control")  # 7 control monkeys
n2  <- sum(Treatment=="Treated")  # 11 treated monkeys
multiplier    <- sqrt(2*((n1+n2-2)/(n1+n2-3))*qf(.95,2,n1+n2-3)) # Sleuth p. 492
summary(myLm1)
shortEffect   <- myLm1$coef[2,1] # Difference in sample averages; Short
seShortEffect <- 3.352   # Read this from summary(myLm1)
halfWidth <- multiplier*seShortEffect  # Half width of 95% confidence interval
shortEffect + c(-1,1)*halfWidth #95% CI for effect of treatment on Short
 longEffect    <- myLm1$coef[2,2] # Difference in sample averages; Long
seLongEffect  <- 3.2215  # Read this from summary(myLm1)
halfWidth     <- multiplier*seLongEffect # Half width of 95% confidence interval
longEffect + c(-1,1)*halfWidth  #95% CI for effect of treatment on Long


## GRAPHICAL DISPLAY FOR PRESENTATION
myPointCode   <- ifelse(Treatment=="Control",21,22)
myPointColor  <- ifelse(Treatment=="Control","green","orange")
plot(long ~ jitter(short),
  xlab="Short-Term Memory Score (Percent Correct)",
  ylab="Long-Term Memory Score (Percent Correct)",
  main="Memory Scores for 11 Hippocampus-Blocked and 7 Control Monkeys",
  pch=myPointCode, bg=myPointColor, cex=2.5, lwd=3)
identify(short,long,labels=Monkey)  #  Label the outliers; press Esc when done
legend(52,54,legend=c("Control","Hippocampus Blocked"), pch=c(21,22),
  pt.bg=c("green","orange"), pt.cex=c(2.5,2.5), pt.lwd=c(3,3), cex=1.5)


## ADVANCED: RANDOMIZATION TEST FOR EQUALITY OF BIVARIATE RESPONSES
myAnova <- anova(myLm2, myLm1, test="Hotelling") #Hotelling Test for Treatment
myAnova$approx[2]        #[1] 12.32109:  F-statistic
numRep <- 50 # Number of random regroupings (change to 50,000)
FStats <- rep(0,numRep)  # Initialize a variable for storing the F-statistics
myLmReduced <- lm(cbind(short,long) ~ 1)# Fit the reduced model once
for (rep in 1:numRep) {  # Do the following commands in parenthese num.rep times
  randomGroup <- rep("Group1",18) # Set randomGroup initially to all "Group1"
  randomGroup[sample(1:18,7)]  <- "Group2" # Change 7 at random to "Group2"
  randomGroup <- factor(randomGroup)  # Make the character variable a factor
  myLmFull <- lm(cbind(short,long) ~ randomGroup)  # Fit full model
  myAnova2  <- anova(myLmReduced, myLmFull, test="Hotelling") # Hotelling's test
  FStats[rep] <- myAnova2$approx[2]   # Store the F-statistic
}  # If numRep = 50,000, go get a cup of coffee while you wait for this.
hist(FStats, main="Approx. Randomizatin Dist of F-stat if No Treatment Effect")
abline(v=12.32109)    # Show actually observed Hotelling F-statistic
pValue <- sum(FStats >= 12.32109)/numRep
pValue  # Approximate randomization test p-value (no distributional assumptions)

detach(case1601)
```

---

case1602 *Oat Bran and Cholesterol*

---

## Description

In a randomized, double-blind, crossover experiment, researchers randomly assigned 20 volunteer hospital employees to either a low-fiber or low-fiber treatment group. The subjects followed the diets for six weeks. After two weeks on their normal diet, all patients crossed over to the other treatment group for another six weeks. The total serum cholesterol (in mg/dl) was measured on each patient before the first treatment, at the end of the first six week treatment, and at the end of the second six week treatment.

## Usage

```
case1602
```

## Format

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

**Baseline** total serum cholesterol before treatment

**HiFiber** total serum cholesterol after the high fiber diet

**LoFiber** total serum cholesterol after the low fiber diet

**Order** factor to identify order of treatment, with two levels `"HL"` and `"LH"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Swain, J.F., Rouse, I.L., Curley, C.B., and Sacks, F.M. (1990). Comparison of the Effects of Oat Bran and Low-fiber Wheat on Serum Lipoprotein Levels and Blood Pressure, *New England Journal of Medicine* **320**: 1746–1747.

## Examples

```
str(case1602)
attach(case1602)

## EXPLORATION
highMinusBase <- HiFiber-Baseline
highMinusLow  <- HiFiber-LoFiber
plot(highMinusBase ~ highMinusLow)
abline(h=0) # Horizontal line at 0
abline(v=0) # Vertical line at 0
# Hotelling 2-sample t-test for order effect on bivariate response:
myLm1    <- lm(cbind(highMinusBase,highMinusLow) ~ Order )   # Full model
myLm2    <- update(myLm1, ~ . - Order)  # Reduced model withour Order effect
anova(myLm2, myLm1, test="Hotelling") # p-value for Order effect


## INFERENCE: HOTELLING ONE-SAMPLE TEST THAT MEAN OF BIVARIATE RESPONSE IS (0,0)
myLm3    <- lm(cbind(highMinusBase, highMinusLow) ~ 1)   # Full model
myLm4    <- update(myLm3, ~ . - 1) # Reduced model (with both means = 0)
anova(myLm4, myLm3, test="Hotelling")  # test that the bivariate mean is (0,0)
# Confidence intervals
```

```
summary(myLm3)
HighMinusBase <-  myLm3$coef[1] # -13.850
seHighMinusBase <- 3.533  # Standard error, read from summary(myLm3)
HighMinusLow <- myLm3$coef[2] #  -0.850
seHighMinusLow <- 3.527 # Standard error, read from summary(myLm3)
n <- length(highMinusBase)  # 20: sample size
multiplier <-    sqrt(2*((n-1)/(n-2))*qf(.95,2,n-2))   # See Sleuth, page 495
HighMinusBase + c(-1,1)*multiplier*seHighMinusBase # 95% CI for High - Base
HighMinusLow + c(-1,2)*multiplier*seHighMinusLow   # 95% CI for High - Low


## GRAPHICAL DISPLAY FOR PRESENTATION
lowMinusBase  <- LoFiber - Baseline
myPointCode   <- ifelse(Order== "HL",21,22)
myPointColor  <- ifelse(Order== "HL","green","light blue")
plot(highMinusBase ~ lowMinusBase,
  xlab="Cholesterol Change (from Baseline) After High Fiber Diet (mg/dl)",
  ylab="Cholesterol Change (From Baseline) After Low Fiber Diet (mg/dl)",
  main="Cholesterol Effects of High- and Low-Cholesterol Diets on 20 Subjects",
  ylim=c(-45,20), pch=myPointCode, bg=myPointColor, cex=2.5, lwd=2)
abline(h=0)  # Horizontal line at 0
abline(v=0) # Vertical line at 0
legend(-43,22,c("High Fiber Given First","Low Fiber Given First"), pch=c(21,22),
  pt.bg =c("green","light blue"),pt.cex =c(2.5,2.5), lw = c(2,2), lty=c(0,0))

detach(case1602)
```

---

`case1701` *Magnetic Force on Printer Rods*

---

## Description

Engineers manipulated three factors (with 3, 2, and 4 levels each) in the construction and operation of printer rods, to see if they influenced the magnetic force around the rod.

## Usage

```
case1701
```

## Format

A data frame with 44 observations on the following 14 variables.

**L1, L2, L3, L4, L5, L6, L7, L8, L9, L10, L11** the magnetic force at each of the equally-spaces positions 1, 2, . . . , 11 on the printer rod

**Current** electric current passing through the rod, with three levels `"0"`, `"250"` and `"500"` (milliamperes)

**Config** a factor identifying the configuration, with two levels `"0"` and `"1"`

**Material** a factor identifying the type of metal from which the rod was made, with four levels `"1"`, `"2"`, `"3"` and `"4"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(case1701)
attach(case1701)

## EXPLORATION
MagneticForces  <- cbind(L1,L2,L3,L4,L5,L6,L7,L8,L9,L10,L11)
mfCor <- cor(MagneticForces)
round(mfCor,2)  # Show correlations, rounded to two digits

mfPca     <- prcomp(MagneticForces)  # principal components
summary(mfPca) # Show the proportion of variance explained by each PC
plot(mfPca)  # Graph proportion of variances explained by each PC (Scree Plot)
mfCoefs   <- mfPca$rotation   # Extract the coefficients
dim(mfCoefs)  #   #11 rows and 11 columns
round(mfCoefs[,1:3],3)  # Show the first 3 columns of the score matrix, rounded

# Explore a possible meaningful linear combination suggested by first PC
round(mfCoefs[,1],1)  # Show coefficients of 1st pc, rounded to 1 digit
# Coefficients are all very similar, suggesting a constant coefficient; use 1/11
mfAve     <- (L1 + L2 + L3 + L4 + L5 + L6 + L7 + L8 + L9 + L10 + L11)/11
mfScores  <- mfPca$x
pc1       <- mfScores[,1] #Values for first principal component of MagneticForces
cor(mfAve,pc1) # correlation of the average and the first PC (=0.999567)
plot(pc1 ~ mfAve)

# Explore a possible meningful linear combination suggested by second PC
round(mfCoefs[,2],1) # Show coefficients of 2nd pc, rounded to 1 digit
# Second set of coefficients are negative on the left end of the rod and
# positive on the right end. Try Ave(L9 + L10 + L11) - Ave(L1 + L2 + L3).
mfEnds    <- (L9 + L10 + L11)/3 - (L1 + L2 + L3)/3
pc2       <- mfScores[,2]
residualEnds  <- lm(mfEnds ~ mfAve)$residual # Ends with average effect removed
plot(pc2 ~ residualEnds)
cor(residualEnds, pc2)                # 0.973

# Explore a possible meaningful linear combination suggested by third PC
round(mfCoefs[,3],1)  # Show doefficients of 3rd pc, rounded to 1 digit
# Try a contrast between the first 4 positions and the 6th position
mfPeak    <- (L1 + L2 + L3 + L4)/4 - L6
pc3       <- mfScores[,3]
residualPeak <- lm(mfPeak ~ mfAve + mfEnds)$residual
plot(pc3 ~ residualPeak)
cor(residualPeak,pc3)           # 0.971
# Note: the variation explained by the third PC seems to be due almost entirely
# to one printer rod. (Keep this in mind.)


## INFERENCE:  ANALYSIS OF EXPERIMENTAL FACTORS ON 3-DIMENSIONAL RESPONSE
myResponse  <- cbind(mfAve, mfEnds, mfPeak)
cor(myResponse)
myLm1     <- lm(myResponse ~ Current + Config + Material)
```

```
anova(myLm1)  # Noticeable effect of Current but not Config or Material

plot(mfAve ~ Current)
myLm2   <- lm(mfAve ~ Current)
abline(myLm2)
summary(myLm2)  # No evidence of an effect of current on average magnetic force

plot(mfEnds ~ Current)
myLm3   <- lm(mfEnds ~ Current)
abline(myLm3)
summary(myLm3)  # Evidence that Current effects the difference in MF at the ends

plot(mfPeak ~ Current)
myLm4   <- lm(mfPeak ~ Current)
abline(myLm4)
summary(myLm4)  # No evidence of an effect of Current on peak MF in the center


## GRAPHICAL DISPLAY FOR PRESENTATION
plot(mfEnds ~ jitter(Current,.1),
  xlab="Electrical Current Used in Printer Rod Manufacture (milliamperes)" ,
  ylab="Magnetic Force at Positions 9-11 Minus Magnetic Force at Positions 1-3",
  main="Effect of Electrical Current on Magnetic Force Surrounding Printer Rod",
  col="black", pch=21, lwd=2,  bg="green", cex=2 )
abline(myLm3,
  lwd=2)

detach(case1701)
```

---

`case1702` *Love and Marriage*

---

## Description

Thirty couples participated in a study of love and marriage. Wives and husbands responded separately to four questions:

1. What is the level of passionate love you feel for your spouse?
2. What is the level of passionate love your spouse feels for you?
3. What is the level of compassionate love you feel for your spouse?
4. What is the level of compassionate love your spouse feels for you?

Each response was recorded on a five-point scale: 1=None, 2=Very Little, 3=Some, 4=A Great Deal and 5=A Tremendous Amount.

## Usage

```
case1702
```

## Format

A data frame with 30 observations on the following 9 variables.

**Couple** couple identification number

**HP** level of passionate love husband feels for spouse

**WP** level of passionate love wife feels for spouse

**HC** level of compassionate love husband feels for spouse

**WC** level of compassionate love wife feels for spouse

**PW** level of passionate love husband perceives spouse to have for him

**PH** level of passionate love wife perceives spouse to have for her

**CW** level of compassionate love husband perceives spouse to have for him

**CH** level of compassionate love husband perceives spouse to have for her

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Johnson, R.A. and Wichern, D.W. (1988). *Applied Multivariate Statistical Analysis (2nd ed)*, Prentice-Hall.

## Examples

```
str(case1702)
attach(case1702)

## EXPLORATION
x <- cbind(HP,WP,HC,WC) # 4 components of level of love you feel for spouse
y <- cbind(PW,PH,CW,CH) # 4 components of level of love you perceive from spouse

if(require(CCA)){     # Use the CCA library
  myCCA   <- cc(x,y)  # Store canonical correlation computations
  canCor   <- myCCA$cor  # Extract the canonical correlations
  canCor  #[1] 0.9506990 0.8665601 0.5571876 0.1106555

# Make a function to test the number of canonical correlations (advanced).
# Bartlett modification of likelihood ratio test
# Reference: Mardia, Kent, and Bibby, 1980, Multivariate Analysis,
myTest  <- function(xMatrix,yMatrix) {
  if(require(CCA)){     # Use the CCA library
    myCCA       <- cc(xMatrix,yMatrix) # Store CCA computations
    canCor      <- myCCA$cor   # extract the canonical correlations
    n           <- dim(x)[1] # number of rows of x,= sample size
    p           <- dim(y)[2] # number of component variables in y
    q           <- dim(x)[2] # number of component variables in x
    k           <- min(p,q) # the maximum number of canonical pairs
    testStat    <- rep(0,k) #  store the test statistics; initially set to 0
    degFr       <- rep(0,k)  # store the associated degrees of freedom
    canCor2     <- canCor[k:1] # Reverse order for the following calculations
    productTerm <-  1
    for (i in 1:k)  {
```

```
      productTerm   <-  productTerm*(1 - canCor2[i]^2)
      degFr[i]      <-  (p + 1 - i)*(q + 1 - i)
      testStat[i]   <- -(n -(p+q+3)/2)*log(productTerm)
        }
    pair <- 1:k
    testStat <- round(testStat[k:1],2)  # Revert to original order; round
    pValue  <- round(1 - pchisq(testStat,degFr),4) # p-value   to 4 digits
    canCor  <- round(canCor,4)   # Round to 4 digits
    cbind(pair,canCor, testStat,degFr, pValue) # Show the results;
    }    }
myTest(x,y)

# Explore possible meaningful linear combination suggested by first pair of CCs
round(myCCA$xcoef,1)
round(myCCA$ycoef,1)
# The 1st column of xcoef is almost entirely HC; 1st column of ycoef is CW
ccX1 <- myCCA$scores$xscores[,1]
plot(ccX1 ~ jitter(HC)) # See if HC is a good substitute for 1st X canonical var
cor(ccX1,HC)    #[1] 0.9719947

ccY1 <- myCCA$scores$yscores[,1]
plot(ccY1 ~ jitter(CW)) # See if CW is a good substitute for 1st y canonical var
cor(ccY1, CW)   #[1] 0.9975468

# Analyze the correlation of the meaningful substitute variables
cor(HC,CW)   # [1]  0.9280323
myLm1 <- lm(HC ~ CW)
summary(myLm1) # p-value < 0.0001 (test for slope= 0 equiv to test that corr = 0)


# Explore possible meaningful linear combination suggested by 2nd pair of CCs
# Suggested substitutes from 2nd columns above are WC and CH
ccX2  <- myCCA$scores$xscores[,2]
WCres  <- lm(WC ~ HC)$res  # WC with effect of HC removed
plot(ccX2 ~ WCres)
cor(ccX2,WCres)   #[1] 0.9045225

ccY2  <- myCCA$scores$yscores[,2]
CHres  <- lm(CH ~ CW)$res  # CH with effect of CW removed
plot(ccY2 ~ CHres)
cor(ccY2,CHres)   # [1] 0.9280248

cor(WC,CH)           #[1] 0.8134213
myLm2 <- lm(WC ~ CH)
summary(myLm2)       # p-value < 0.0001 for test that correlation = 0

# Explore cannonical correlations from other groupings of variables
x <- cbind(HP, HC, PW, CW)  # husband's responses
y <- cbind(WP, WC, PH, CH)  # wife's responses
myTest(x,y)    # No evidence that husbands' responses are correlated with wifes'
x <- cbind(HP,PW,WP,PH)    # passionate responses
y <- cbind(HC,CW,WC,CH)    # compassionate responses
myTest(x,y) #No evidence that passionate anc compassionate responses are correlated


##  GRAPHICAL DISPLAYS FOR PRESENTATION
jFactor <- 0.3   # Jittering factor (try different values to see what works)
```

```
jHC <- jitter(HC,factor=jFactor)
jCW <- jitter(CW,factor=jFactor)
jCH <- jitter(CH,factor=jFactor)
jWC <- jitter(WC,factor=jFactor)
opar <- par(no.readonly=TRUE)  # Store current graphical parameter settings
par(mfrow=c(2,2)) # Prepare to make a 2x2 panel of graphs
par(mar=c(1.1,4.1,1.1,1.1) )  # Adjust margins
plot(jHC ~ jCW, ylab="Husband's Compassionate Love For His Wife (Jittered)",
  xlab="", ylim=c(3,5.1), xlim=c(3,5.1), col="black", pch=21, lwd=2,
  bg="green", cex=2 )
text(3,5.1,"correlation = 0.93",adj=0)
text(3,5.0,"p-value < 0.0001",adj=0)
abline(myLm1)

par(mar=c(1.1,1.1,1.1,4.1))
plot(jHC ~ jCH, xlab="", ylab="", ylim=c(3,5.1), xlim=c(3,5.1),
  col="black", pch=21, lwd=2, bg="green",  cex=2 )
cor(HC,CH)      #[1] 0.274204
myLm3 <- lm(HC ~ CH)
summary(myLm3)# p-value = 0.143
text(3,5.1,"correlation = 0.27",adj=0)
text(3,5.0,"p-value = 0.14",adj=0)
abline(myLm3)

par(mar=c(4.1,4.1,1.1,1.1))
plot(jWC ~ jCW,
  xlab="Husband's Perceived Compassionate Love From His Wife (Jittered)",
  ylab="Wife's Compassionate Love For Her Husband (Jittered)",
  ylim=c(3,5.1), xlim=c(3,5.1), col="black", pch=21, lwd=2, bg="green", cex=2 )
cor(WC,CW)           #[1] 0.04171195
myLm4 <- lm(WC ~ CW)
summary(myLm4)       # p-value = 0.827
text(3,3.1,"correlation = 0.04",adj=0)
text(3,3,"p-value = 0.8",adj=0)
abline(myLm4)

par(mar=c(4.1,1.1,1.1,4.1))
plot(jWC ~ jCH, ylab="",
  xlab="Wife's Perceived Compassionate Love From Her Husband (Jittered)",
  ylim=c(3,5.1), xlim=c(3,5.1), col="black", pch=21,  lwd=2, bg="green", cex=2 )
text(3,3.1,"correlation = 0.81",adj=0)
text(3,3,"p-value < 0.0001",adj=0)
abline(myLm2)

par(opar) # Restore previous graphics parameter settings
}
detach(case1702)
```

---

case1801 *Obesity and Heart Disease*

---

## Description

To better understand whether the relationship between heart disease and obesity could be due to the social stigma associated with obesity, researchers examined cardiovascular disease rates of obese

and non-obese women in American Samoa, where obesity was considered socially desirable. 3,112 women were categorized according to whether they were obese or not and whether they died from cardiovascular disease (CVD).

## Usage

```
case1801
```

## Format

A data frame with 2 observations on the following 3 variables.

**Obesity** a factor with levels `"NotObese"` and `"obese"`

**Deaths** the number of women who died from CVD

**NonDeaths** the number that died from other causes

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cenage Learning.

## References

Crews, D.E. (1988). Cardiovascular Mortality in American Samoa, *Human Biology* **60**: 417–433.

## Examples

```
str(case1801)
attach(case1801)

## EXPLORATION
myTable             <- cbind(Deaths,NonDeaths)  # Form a 2 by 2 table of counts
row.names(myTable)  <- Obesity  # Assign the levels of Obesity as row names
myTable   # Show the table

## INFERENCE (4 methods for getting p-values and confidence intervals)
prop.test(myTable, alternative="greater", correct=FALSE) # Compare 2 proportions
prop.test(myTable, alternative="greater", correct=TRUE) # ...with cont. corect.
prop.test(myTable,correct=TRUE) # 2-sided alternative (default) to get CI
chisq.test(myTable) # Pearson's Chi-Squared Test
fisher.test(myTable, alternative="greater") # Fisher's exact test
fisher.test(myTable) # 2-sided alternative to get CI for odds ratio
myGlm1  <- glm(myTable ~ Obesity, family=binomial)  # Logistic reg (CH 21)
summary(myGlm1)  # Get p-value-- 0.734
beta    <- myGlm1$coef
exp(beta[2])  #Odds of death are estimated to be 17% higher for  obese women
exp(confint(myGlm1,2)) # 95% confidence interval


## GRAPHICAL DISPLAY FOR PRESENTATION
myTable
#        Deaths NonDeaths
#Obese        16      2045
#NotObese      7      1044
prop.test(16,(16+2045)) #For one proportion, est: 0.0078 95% CI: 0.0046 to 0.013
prop.test(7,(7+1044)) #For one proportion, est: 0067 95% CI: 0.0029 to 0.014
```

```
pHat    <- c(0.007763222, 0.006660324)*1000 # Get estimated deaths per 1,000 women
lower95 <- c(0.00459943, 0.002921568)*1000
upper95 <- c(0.01287243, 0.014318321)*1000

if(require(Hmisc)) { # Use Hmisc library
  myObj   <- Cbind(pHat,lower95,upper95)
  Dotplot(Obesity ~ myObj,   # Draw a dot plot of estimates and CIs
    xlab="Estimated CVD Deaths Per 1,000 Women (and 95% Confidence Intervals)",
    ylab="Weight Category", ylim=c(.5,2.5), cex=2)
}


detach(case1801)
```

---

case1802 *Vitamin C and the Common Cold*

---

## Description

In a randomized experiment, researchers assigned 407 volunteers to receive 1,000 mg of Vitamin C daily throughout the cold season and 411 to receive a placebo. A physician interviewed the volunteers at the end of the study to determine whether or not they had suffered any colds during the study period.

## Usage

```
case1802
```

## Format

A data frame with 2 observations on the following 3 variables.

**Treatment** a factor with levels `"Placebo"` and `"VitC"`

**Cold** the number of who got colds

**NoCold** the number that did not get any colds

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cenage Learning.

## References

Anderson, T.W., Reid, D.B.W. and Beaton, G. H. (1972). Vitamin C and the Common Cold, *Canadian Medial Association Journal* **107**: 503–508.

## Examples

```
str(case1802)
attach(case1802)

library(MASS)
## INFERENCE (4 methods)
myTable <- cbind(Cold,NoCold)
row.names(myTable) <- c("Placebo","Vitamin C")
myTable
prop.test(myTable, alternative="greater") # Compare 2 binomial proportions
# Alternative: pop prop. of first column (cold) in larger in first row (placebo)
prop.test(myTable, alternative="greater", correct=TRUE)
prop.test(myTable,correct=TRUE) # Use 2-sided alternative to get CI
chisq.test(myTable)   # Chi-square test
fisher.test(myTable, alternative="greater")
fisher.test(myTable) #  2-sided alternative to get CI for odds ratio
myGlm1  <- glm(myTable ~ Treatment, family=binomial) # logistic reg (Ch 21)
summary(myGlm1)
beta    <- myGlm1$coef
1 - exp(beta[2])  # 0.3474911
1 - exp(confint(myGlm1,2)) # 0.53365918 0.09042098
# Interpretation: The odds of getting a cold are 35% less on Vitamin C than
# Placebo (95% confidence interval: 9% to 53% less).

detach(case1802)
```

---

# case1803 *Smoking and Lung Cancer*

---

## Description

In a retrospective case-control study, researchers identified 86 lung cancer patients and 86 controls (without lung cancer), and categorized them according to whether they were smokers or non-smokers. The goal is to see whether the odds of lung cancer are greater for smokers than for non-smokers.

## Usage

```
case1803
```

## Format

A data frame with 2 observations on the following 3 variables.

**Smoking** a factor with levels `"NonSmokers"` and `"Smokers"`

**Cancer** the number of who were lung cancer patients

**Control** the number who were controls

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cenage Learning.

## References

Anderson, T.W., Reid, D.B.W. and Beaton, G. H. (1972). Vitamin C and the Common Cold, *Canadian Medial Association Journal* **107**: 503–508.

## Examples

```
str(case1803)
attach(case1803)

## INFERENCE
myTable   <- cbind(Cancer,Control)   # Make a 2-by-2 table of counts
row.names(myTable)  <- Smoking   # Assign the levels of Smoking as row names
myTable

fisher.test(myTable,  alternative="greater")  # Alternative: that odds of Cancer
  # in first row are greater.
fisher.test(myTable) # 2-sided alternative to get CI for odds ratio
myGlm1  <- glm(myTable ~ Smoking, family=binomial) # logistic reg (Ch 21)
summary(myGlm1)
exp(myGlm1$coef[2]) # 5.37963 : Estimated odds ratio
exp(confint(myGlm1)[2,]) #  1.675169 24.009510:  Approximate confidence interval
# Interpretation: The odds of cancer ar 5.4 times as large for smokers as for
# non-smokers (95% confidence interval: 1.7 to 24.0 times as large).

detach(case1803)
```

---

# case1901 *Sex Role Sterotypes and Personnel Decisions*

---

## Description

Researchers gave 48 male bank supervisors attending a management institute hypothetical personnel files and asked them whether they would promote the applicant based on the file. The personnel files were identical except that 24 of them listed a male and 24 listed a female applicant. The assignment of managers to receive either a male or female applicant file was carried out at random.

## Usage

```
case1901
```

## Format

A data frame with 2 observations on the following 3 variables.

**Gender** a factor with levels `"Female"` and `"Male"`

**Promoted** the number of managers who promoted the applicant

**NotPromoted** the number of managers who did not promote the applicant

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rosen, B. and Jerdee, J (1974). Influence of Sex Role Steroetypes on Personnel Decisions, *Journal of Applied Psychology* **59**: 9–14.

## Examples

```
str(case1901)
attach(case1901)

## INFERENCE
myTable             <- cbind(Promoted,NotPromoted)
row.names(myTable)  <- Gender
myTable
fisher.test(myTable, alternative="greater")
  # Alternative: that odds of Promotion in first row (Males) are greater.
fisher.test(myTable)  # Use 2-sided to get confidence interval for odds ratio
prop.test(myTable) # Compare two binomial proportions


## GRAPHICAL DISPLAY FOR PRESENTATION
myTable
#         Promoted NotPromoted
#Male         21           3
#Female       14          10
prop.test(21,(21+3)) # Est = .875; CI = .665 to .967
prop.test(14,(14+10))# Est = .583; CI = .369 to .772

pHat   <- c(0.875,0.583)
lower95 <- c(0.665, 0.369)
upper95 <- c(0.967, 0.772)
if(require(Hmisc)) {      # Use Hmisc library
  myObj<- Cbind(pHat,lower95,upper95) # Cbind: a form of cbind needed for Dotplot
  Dotplot(Gender ~ myObj,
  xlab="Probability of Promotion Based on Applicant File (and 95% Confidence Intervals)",
    ylab="Gender Listed in Applicant File", ylim=c(.5,2.5), cex=2)
}

detach(case1901)
```

---

`case1902` *Death Penalty and Race*

---

## Description

Lawyers collected data on convicted black murderers in the state of Georgia to see whether convicted black murderers whose victim was white were more likely to receive the death penalty than those whose victim was black, after accounting for aggravation level of the murder. They categorized murders into 6 progressively more serious types. Category 1 comprises barroom brawls, liquor-induced arguments lovers' quarrels, and similar crimes. Category 6 includes the most vicious, cruel, cold=blooded, unprovoked crimes.

## Usage

```
case1902
```

## Format

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

**Aggravation** the aggravation level of the crime, a numerical variable ranging from 1 to 6

**Victim** a factor indicating race of murder victim, with levels `"White"` and `"Black"`

**Death** number in the aggravation and victim category who received the death penalty

**NoDeath** number in the aggravation and victim category who did not receive the death penalty

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Woodworth, G.C. (1989). Statistics and the Death Penalty, *Stats* **2**: 9–12.

## Examples

```
str(case1902)
attach(case1902)

## EXPLORATION
proportionDeath  <- Death/(Death + NoDeath)
myPointCode <- ifelse(Victim=="White",22,24)
myPointColor <- ifelse(Victim=="White","white","black")
plot(proportionDeath ~ Aggravation, pch=myPointCode, bg=myPointColor)
oddsOfDeath <- Death/(NoDeath + .5)  # Add .5 to the demoninator to avoid 0's
plot(oddsOfDeath ~ Aggravation, pch=myPointCode, bg=myPointColor)
plot(oddsOfDeath ~ Aggravation, log="y", pch=myPointCode, bg=myPointColor)
# Use logistic regression (Ch 21) to see if the 6 odds ratios are constant
myGlm1 <- glm(cbind(Death,NoDeath) ~ Aggravation + Victim +
  Aggravation:Victim, family=binomial) # Logistic reg with interaction
myGlm2  <- update(myGlm1, ~ . - Aggravation:Victim) # without interaction
anova(myGlm2, myGlm1) # no evidence of interaction.


## INFERENCE
# Mantel Haenszel
myTable <- array(rbind(Death, NoDeath), dim=c(2,2,6),
  dimnames=list(Penalty=c("Death","No Death"), Victim=c("White","Black"),
  Aggravation=c("1","2","3","4","5","6")))
myTable   # Show the 6 2x2 tables
mantelhaen.test(myTable, alternative="greater", correct=FALSE) # 1-sided p-value
mantelhaen.test(myTable, alternative="greater") # with continuity correction
mantelhaen.test(myTable) # two.sided (default) for confidence interval

# Logistic Regression (Ch 21) (treating aggravation level as numerical)
summary(myGlm2)
beta  <- myGlm2$coef
exp(beta[3])   #   6.1144
exp(confint(myGlm2,3)) # 2.23040 18.72693
# Interpretation: The odds of death penalty for white victim murderers are
# estimated to be 6 times the odds of death penalty for black victim murderers
# with similar aggravation level(95% confidence interval: 2.2 to 18.7 times).
```

```
## GRAPHICAL DISPLAY FOR PRESENTATION
myPointColor     <- ifelse(Victim=="White","green", "orange")
plot(jitter(proportionDeath,.1) ~ jitter(Aggravation,.1),
  xlab="Aggravation Level of the Murder",
  ylab="Proportion of Murderers Who Received Death Penalty",
  pch=myPointCode, bg=myPointColor, cex=2, lwd=2)
legend(1,1, c("White Victim Murderers","Black Victim Murderers"), pch=c(21,22),
  pt.cex=c(2,2), pt.bg=c("green","orange"), pt.lw=c(2,2))
# Include logistic regression fit on plot
dummyAg <- seq(min(Aggravation),max(Aggravation),length=50)
etaB    <- beta[1] + beta[2]*dummyAg
etaW    <- etaB + beta[3]
pB      <- exp(etaB)/(1 + exp(etaB))  # Estimated prob of DP; Black victim
pW      <- exp(etaW)/(1 + exp(etaW))  # Estimated prob of DP; White victim
lines(pB ~ dummyAg,lty=1)
lines(pW ~ dummyAg,lty=2)

detach(case1902)
```

---

`case2001` *Survival in the Donner Party*

---

## Description

This data frame contains the ages and sexes of the adult (over 15 years) survivors and nonsurvivors of the Donner party.

## Usage

```
case2001
```

## Format

A data frame with 45 observations on the following 3 variables.

**Age** Age of person

**Sex** Sex of person

**Status** Whether the person survived or died

## Details

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Grayson, D.K. (1990). Donner Party Deaths: A Demographic Assessment, *Journal of Anthropological Research* **46**: 223–242.

## See Also

ex1918

## Examples

```
str(case2001)
attach(case2001)

## EXPLORATION AND MODEL BUILDING
myPointCode   <- ifelse(Sex=="Female",22,24)
myPointColor  <- ifelse(Sex=="Female","green","orange")
survivalIndicator <- ifelse(Status=="Survived",1,0)
jFactor <- 0.1   # jittering factor
plot(jitter(survivalIndicator,jFactor) ~ jitter(Age, jFactor),
  pch=myPointCode, bg=myPointColor, cex=1.5)
# Logistic regression. Start with a rich model; use backward elimination
ageSquared <- Age^2
myGlm1 <- glm(Status ~ Age + ageSquared + Sex + Age:Sex + ageSquared:Sex,
  family=binomial)
# Use backward elimination, but remove interaction and squared terms 1st
summary(myGlm1)
myGlm2 <- update(myGlm1, ~ . - ageSquared:Sex)
summary(myGlm2)
myGlm3 <- update(myGlm2, ~ . - ageSquared)
summary(myGlm3)   # Wald test p-value for interaction of Age and Sex is: 0.0865
# More accurate likelihood ratio (drop in deviance) test:
myGlm4 <-update(myGlm3, ~ . - Age:Sex)
anova(myGlm4, myGlm3)  # Drop-in-devaince chi-square stat = 3.9099 on 1 d.f.
 1 - pchisq(3.9099,1)  # 2-sided p-value = 0.048

## INFERENCE AND INTERPRETATION
# Proceed by ignoring interaction (for a casual and approximate analysis)
myGlm5  <- update(myGlm4, ~ . - Sex)
anova(myGlm5, myGlm4) # Drop-in-deviance chi-square statistic = 5.0344 on 1 d.f.
1 - pchisq(5.0344,1)  # 2-sided p-value 0.02484869: Highly suggestive
0.0248869/2  # 1-sided p-value = half the 2-sided p-value = 0.01244345
# Interpretation and confidence interval
Sex <- factor(Sex,levels=c("Male","Female")) # Reorder levels so "Male" is ref
myGlm4b <- glm(Status ~ Age + Sex, family=binomial)
beta <- myGlm4b$coef
exp(beta[3]) # 4.939645
exp(confint(myGlm4b,3))  # 25.246069  1.215435
# Interpretation: The odds of survival for females are estimated to be 5 times
# the odds of survival of similarly-aged mean (95% CI: 1.2 times to 25.2 times).
```

```
## GRAPHICAL DISPLAY FOR PRESENTATION
myPointCode <- ifelse(Sex=="Female",22,24)
myPointColor <- ifelse(Sex=="Female","green","orange")
myLineColor <- ifelse(Sex=="Female","dark green","blue")
survivalIndicator <- ifelse(Status=="Survived",1,0)
jFactor <- 0.1
plot(jitter(survivalIndicator,jFactor) ~ jitter(Age, jFactor),
  ylab="Estimated Survival Probability", xlab="Age (years)",
  main=c("Donner Party Survival by Sex and Age"), xlim=c(15,75),
  pch=myPointCode, bg=myPointColor, col=myLineColor, cex=2, lwd=3)
beta            <- myGlm4b$coef
dummyAge        <- seq(15,65,length=50)
linearMale      <- beta[1] + beta[2]*dummyAge  #log odds of survival for males
linearFemale    <- linearMale + beta[3] #log odds of survival for females
pCurveMale      <- exp(linearMale)/(1 + exp(linearMale) ) # survival prob; males
pCurveFemale    <- exp(linearFemale)/(1 + exp(linearFemale)) # females
lines(pCurveMale ~ dummyAge,lty=2,col="blue",lwd=3)
lines(pCurveFemale[dummyAge <= 50] ~ dummyAge[dummyAge <= 50],lty=1,
  col="dark green",lwd=3)

legend(63,.5,legend=c("Females","Males"), pch=c(22,24),
  pt.bg = c("green","orange"), pt.cex=c(2,2), lty=c(1,2), lwd=c(3,3),
  col=c("dark green","blue"))
text(72,1,"Survived (20)")
text(72,0,"Died (25)")

detach(case2001)
```

---

`case2002` *Birdkeeping and Lung Cancer*

---

## Description

A 1972–1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a *case–control* study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

## Usage

```
case2002
```

## Format

A data frame with 147 observations on the following 7 variables.

**LC** Whether subject has lung cancer

**FM** Sex of subject

**SS** Socioeconomic status, determined by occupation of the household's principal wage earner

**BK** Indicator for birdkeeping (caged birds in the home for more that 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (control))

**AG** Age of subject (in years)

**YR** Years of smoking prior to diagnosis or examination

**CD** Average rate of smoking (in cigarettes per day)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Holst, P.A., Kromhout, D. and Brand, R. (1988). For Debate: Pet Birds as an Independent Risk Factor for Lung Cancer, *British Medical Journal* **297**: 13–21.

## Examples

```
str(case2002)
attach(case2002)

## EXPLORATION AND MODEL BUILDING
myCode <- ifelse(BK=="Bird" & LC=="LungCancer","Bird & Cancer",
  ifelse(BK=="Bird" & LC=="NoCancer","Bird & No Cancer",
  ifelse(BK=="NoBird" & LC=="LungCancer","No Bird & Cancer", "No Bird & No Cancer")))
table(myCode)
if(require(car)){   # Use the car library
scatterplotMatrix(cbind(AG,YR,CD), groups=myCode, diagonal="none",reg.line=FALSE,
  pch=c(15,21,15,21), col=c("dark green","dark green","purple","purple"),
  var.labels=c("Age","Years Smoked","Cigarettes per Day"), cex=1.5)
}

# Reorder the levels so that the model is for log odds of cancer
LC    <- factor(LC, levels=c("NoCancer","LungCancer"))
myGlm <- glm(LC ~ FM + SS + AG + YR + CD + BK, family=binomial)
if(require(car)){   # Use the car library
  crPlots(myGlm)  }
# It appears that there's an effect of Years of Smoking and of Bird Keeping
# after accounting for other variables; no obvious effects of other variables

# Logistic regression model building using backward elimination (witholding BK)
myGlm1 <- glm(LC ~ FM + SS + AG + YR + CD, family=binomial)
summary(myGlm1)
myGlm2 <- update(myGlm1, ~ . - SS)
summary(myGlm2)
myGlm3 <- update(myGlm2, ~ . - CD)
summary(myGlm3)
myGlm4 <- update(myGlm3, ~ . - FM)
summary(myGlm4) # Everything left has a small p-value (retain the intercept)


## INFERENCE AND INTERPRETATION
myGlm5 <- update(myGlm4, ~ . + BK)    # Now add bird keeping
summary(myGlm5)
myGlm6 <- update(myGlm5, ~ . + BK:YR + AG:YR) # Try interaction terms
anova(myGlm6,myGlm5) # Drop-in-deviance = 1.61 on 2 d.f.
1 - pchisq(1.61,2)    # p-value = .45: no evidence of interaction
anova(myGlm4,myGlm5)   # Test for bird keeping effect
```

```
(1 - pchisq(12.612,1))/2  # 1-sided p-value: 0.0001916391

BK <- factor(BK, levels=c("NoBird", "Bird"))  # Make "no bird" the ref level
myGlm5b <- glm(LC ~ AG + YR + BK, family=binomial)
beta <- myGlm5b$coef  # Extract estimated coefficients
exp(beta[4])   # 3.961248
exp(confint(myGlm5b,4))   # 1.836764 8.900840
# Interpretation: The odds of lung cancer for people who kept birds were
# estimated to be 4 times the odds of lung cancer for people of similar age, sex,
# smoking history, and socio-economic status who didn't keep birds
# (95% confidence interval for this adjusted odds ratio: 1.8 times to 8.9 times).

# See bestglm library for an alternative variable selection technique.

detach(case2002)
```

---

`case2101` *Island Size and Bird Extinctions*

---

## Description

In a study of the Krunnit Islands archipelago, researchers presented results of extensive bird surveys taken over four decades. They visited each island several times, cataloguing species. If a species was found on a specific island in 1949, it was considered to be at risk of extinction for the next survey of the island in 1959. If it was not found in 1959, it was counted as an "extinction", even though it might reappear later. This data frame contains data on island size, number of species at risk to become extinct and number of extinctions.

## Usage

```
case2101
```

## Format

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

**Island** Name of Island

**Area** Area of Island

**AtRisk** Number of species at risk

**Extinct** Number of extinctions

## Details

Scientists agree that preserving certain habitats in their natural states is necessary to slow the accelerating rate of species extinctions. But they are divided on how to construct such reserves. Given a finite amount of available land, is it better to have many small reserves or a few large one? Central to the debate on this question are observational studies of what has happened in island archipelagos, where nearly the same fauna tries to survive on islands of different sizes.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Väisänen, R.A. and Järvinen, O. (1977). Dynamics of Protected Bird Communities in a Finnish Archipelago, *Journal of Animal Ecology* **46**: 891–908.

## Examples

```
str(case2101)
attach(case2101)

## EXPLORATION AND MODEL BUILDING
proportionExtinct <- Extinct/AtRisk
oddsExtinct  <- proportionExtinct/(1 - proportionExtinct)
logitExtinct <- log(oddsExtinct)   # Logit = Log Odds
logArea <- log(Area)
plot(logitExtinct ~ logArea)

binResponse  <- cbind(Extinct,AtRisk-Extinct)
myGlm1 <- glm(binResponse ~ logArea, family=binomial)
summary(myGlm1)
logArea2 <- logArea^2
myGlm2 <- update(myGlm1, ~ . + logArea2)
summary(myGlm2) # p-value for quadratic term: 0.77


## INFERENCE AND INTERPRETATION
myGlm3 <- update(myGlm1, ~ . - logArea)
anova(myGlm3, myGlm1) # Drop in deviance statistic = 33.277 on 1 d.f.
1 - pchisq(33.277,1) # p-value = 7.992234e-09
beta <- myGlm1$coef
1 - 2^beta[2]  # 0.1861153
1 - 2^confint(myGlm1,2) #0.2462041 0.1247743
# Interpretation: Associated with each doubling of island area is a 19%
# reduction in the odds of extinction (95% confidence interval: 12% to 25%
# reduction).


## GRAPHICAL DISPLAY FOR PRESENTATION
plot(oddsExtinct ~ Area, log="xy", ylab="Observed Odds of Extinction; log scale",
  xlab=expression(paste("Island Area (km"^"2","); log scale")),
  main="Extinctions of Bird Species in the Krunnit Island Archipelago",
  pch=21, lwd=2, bg="green", cex=2)   # Plot odds of extinction vs island area
dummyArea <- seq(min(Area),max(Area),length=50)
lp <- beta[1] + beta[2]*log(dummyArea)
odds <- exp(lp)
lines(odds ~ dummyArea,lwd=2)

plot(proportionExtinct ~ Area, log="xy",
  ylab="Proportions of 1949 Species not Found in 1959",
  xlab=expression(paste("Island Area (km"^"2","); log scale")),
  main="Proportions of 1949 Bird Species Extinct in 1959 on 18 Krunnit Archipelago Islands",
  pch=21, lwd=2, bg="green", cex=2)  # Plot probability of extinction vs area
dummyArea  <- seq(min(Area),max(Area),length=50)
lp <- beta[1] + beta[2]*log(dummyArea)
myProbability <- exp(lp)/(1 + exp(lp))
lines(myProbability ~ dummyArea,lwd=2,col="blue")
legend(.08,.055,legend="Estimated Probability of Extinction",lty=1,lwd=2,col="blue")
```

```
detach(case2101)
```

---

`case2102` *Moth Coloration and Natural Selection*

---

## Description

This data was collected by J.A. Bishop. Bishop selected seven locations progressively farther from Liverpool. At each location, Bishop chose eight trees at random. Equal number of dead (frozen) light (*Typicals*) and dark (*Carbonaria*) moths were glued to the trunks in lifelike positions. After 24 hours, a count was taken of the numbers of each morph that had been removed—presumably by predators.

## Usage

```
case2102
```

## Format

A data frame with 14 observations on the following 4 variables.

**Morph** Morph, a factor with levels `"light"` and `"dark"`

**Distance** Distance from Liverpool (in km)

**Placed** Number of moths placed

**Removed** Number of moths removed

## Details

Population geneticists consider clines particularly favourable situations for investigating evolutionary phenomena. A cline is a region where two colour morphs of one species arrange themselves at opposite ends of an environmental gradient, with increasing mixtures occurring between. Such a cline exists near Liverpool, England, where a dark morph of a local moth has flourished in response to the blackening of tree trunks by air pollution from the mills. The moths are nocturnal, resting during the day on tree trunks, where their coloration acts as camouflage against predatory birds. In Liverpool, where tree trunks are blackened by smoke, a high percentage of the moths are of the dark morph. One encounters a higher percentage of the typical (pepper–and–salt) morph as one travels from the city into the Welsh countryside, where tree trunks are lighter. J.A. Bishop used this cline to study the intensity of natural selection.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Bishop, J.A. (1972). An Experimental Study of the Cline of Industrial Melanism in *Biston betularia* [Lepidoptera] Between Urban Liverpool and Rural North Wales, *Journal of Animal Ecology* **41**: 209–243.

## Examples

```
str(case2102)
attach(case2102)

## EXPLORATION  AND MODEL BUILDING
proportionRemoved <- Removed/Placed
myPointCode <- ifelse(Morph=="dark",21,22)
myPointColor <- ifelse(Morph=="dark","blue","gray")
plot(proportionRemoved ~ Distance, pch=myPointCode, bg=myPointColor, cex=2, lwd=2)

binResponse <- cbind(Removed, Placed-Removed)
Morph <- factor(Morph, levels=c("light","dark"))  # Make "light" the ref  level
myGlm1 <- glm(binResponse ~ Distance + Morph + Distance:Morph, family=binomial)
summary(myGlm1)  # Residual deviance:  13.230  on 10  degrees of freedom
1 - pchisq(13.230,10)  # No evidence of overdispersion
myGlm2  <- update(myGlm1, ~ . - Distance:Morph)
anova(myGlm2, myGlm1) # Drop in deviance statistic = 11.931 on 1 d.f.
1 - pchisq(11.931,1)  # p-value = 0.0005520753 => strong evidence of interaction
# It appears that the intercepts are the same for both light and dark morphs,
# that there is no effect of Distance for light morphs, but there is an effect
# of Distance for dark morphs.


## INFERENCE AND INTERPREATION
myTerm  <- Distance*ifelse(Morph=="dark",1,0) # Create indicator var for "dark"
myGlm3 <- glm(binResponse ~ myTerm, family=binomial)
summary(myGlm3)


## GRAPHICAL DISPLAY FOR PRESENTATION
myPointCode <- ifelse(Morph=="dark",22,24)
myPointColor  <- ifelse(Morph=="dark","blue","orange")
plot(proportionRemoved ~ Distance, ylab="Proportion of Moths Taken",
  main="Proportions of Moths Taken by Predators at Seven Locations",
  xlab="Distance from Liverpool (km)", pch=myPointCode, bg=myPointColor, cex=2,
  lwd=2)
beta  <- myGlm3$coef
dummyDist <- seq(0,55,length=50)
lp <- beta[1] + beta[2]*dummyDist
propDark <- exp(lp)/(1 + exp(lp))
lines(propDark ~ dummyDist,lwd=2,col="blue")
propLight <- rep(exp(beta[1])/(1 + exp(beta[1])),length(dummyDist))
lines(propLight ~ dummyDist,lwd=2,col="orange")
legend(0,0.47,legend=c("Dark Morph","Light Morph"),
  pch=c(22,24),pt.bg=c("blue","orange"),pt.cex=c(2,2),pt.lwd=c(2,2))

detach(case2102)
```

---

# case2201 *Age and Mating Success of Male Elephants*

---

## Description

Although male elephants are capable of reproducing by 14 to 17 years of age, your adult males are usually unsuccessful in competing with their larger elders for the attention of receptive females.

Since male elephants continue to grow throughout their lifetimes, and since larger males tend to be more successful at mating, the males most likely to pass their genes to future generations are those whose characteristics enable them to live long lives. Joyce Poole studied a population of African elephants in Amboseli National Park, Kenya, for 8 years. This data frame contains the number of successful matings and ages (at the study's beginning) of 41 male elephants.

## Usage

```
case2201
```

## Format

A data frame with 41 observations on the following 2 variables.

**Age** Age of elephant at beginning of study

**Matings** Number of successful matings

## Source

Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*, Duxbury.

## References

Poole, J.H. (1989). Mate Guarding, Reproductive Success and Female Choice in African Elephants, *Animal Behavior* **37**: 842–849.

## Examples

```
str(case2201)
attach(case2201)

## EXPLORATION AND MODEL BUILDING
plot(Matings ~ Age,  log="y")
ageSquared  <- Age^2
myGlm1 <- glm(Matings ~ Age + ageSquared, family=poisson)
summary(myGlm1)  # No evidence of a need for ageSquared


## INFERENCE AND INTERPRETATION
myGlm2  <- update(myGlm1, ~ . - ageSquared)
summary(myGlm2)
beta  <- myGlm2$coef
exp(beta[2])  #1.071107
exp(confint(myGlm2,2))  #1.042558 1.100360
# Interpretation: Associated with each 1 year increase in age is a 7% increase
# in the mean number of matings (95% confidence interval 4% to 10% increase).


## GRAPHICAL DISPLAY FOR PRESENTATION
plot(Matings ~ Age, ylab="Number of Successful Matings",
  xlab="Age of Male Elephant (Years)",
  main="Age and Number of Successful Matings for 41 African Elephants",
  pch=21, bg="green", cex=2, lwd=2)
dummyAge <- seq(min(Age),max(Age), length=50)
lp <- beta[1] + beta[2]*dummyAge
```

```
curve <- exp(lp)
lines(curve ~ dummyAge,lwd=2)

detach(case2201)
```

---

case2202 *Characteristics Associated with Salamander Habitat*

---

## Description

The Del Norte Salamander (*plethodon elongates*) is a small (5–7 cm) salamander found among rock rubble, rock outcrops and moss-covered talus in a narrow range of northwest California. To study the habitat characteristics of the species and particularly the tendency of these salamanders to reside in dwindling old-growth forests, researchers selected 47 sites from plausible salamander habitat in national forest and parkland. Randomly chosen grid points were searched for the presence of a site with suitable rocky habitat. At each suitable site, a 7 metre by 7 metre search are was examined for the number of salamanders it contained. This data frame contains the counts of salamanders at the sites, along with the percentage of forest canopy and age of the forest in years.

## Usage

```
case2202
```

## Format

A data frame with 47 observations on the following 4 variables.

**Site** Investigated site

**Salamanders** Number of salamanders found in 49 $m^2$ area

**PctCover** Percentage of canopy cover

**ForestAge** Forest age

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Welsh, H.H. and Lind, A.J. (1995). *Journal of Herpetology* **29**(2): 198–210.

## Examples

```
str(case2202)
attach(case2202)

## EXPLORATION AND MODEL BUILDING
logSalamanders <- log(Salamanders + .5)
logForestAge   <- log(ForestAge + .5)
myMatrix       <- cbind(PctCover,logForestAge,logSalamanders)
if (require(car)) { # Use car library
 scatterplotMatrix(myMatrix, diagonal="histogram", reg.line=FALSE, spread=FALSE)
```

```
}

myGlm1  <- glm(Salamanders ~ PctCover + logForestAge + PctCover:logForestAge,
  family=poisson)
summary(myGlm1)   # Backward elimination...
myGlm2 <- update(myGlm1, ~ . - PctCover:logForestAge)
summary(myGlm2)
myGlm3  <- update(myGlm2, ~ . - logForestAge)
summary(myGlm3)   # PctCover is the only explanatory variable remaining

plot(Salamanders ~ PctCover)  # It appears that there are 2 distributions
# of Salamander counts; one for PctCover < 70 and one for PctCover > 70

# See if PctCover is associated Salamanders in each subset
myGlm4 <- glm(Salamanders ~ PctCover, family=poisson,subset=(PctCover > 70))
summary(myGlm4)           # No evidence of an effect for this subset
myGlm5 <- glm(Salamanders ~ PctCover, family=poisson,subset=(PctCover < 70))
summary(myGlm5)           # No evidence on this subset either

## INFERENCE  (2 means)
Group <- ifelse(PctCover > 70,"High","Low")
Group <- factor(Group, levels=c("Low","High") )  # Make "Low Cover" the ref group
myGlm6 <- glm(Salamanders ~ Group, family=poisson)
summary(myGlm6)


## GRAPHICAL DISPLAY FOR PRESENTATION
plot(Salamanders ~ PctCover, ylab="Number of Salamanders",
  xlab="Percentage of Canopy Covered",
  main="Number of Salamanders versus Percent Canopy Cover",
  pch=21,bg="green", cex=2, lwd=2)
beta <- myGlm6$coef
lines(c(0,55),exp(c(beta[1],beta[1])),lwd=2)
text(56,exp(beta[1]),paste("mean= ",round(exp(beta[1]),1)),adj=0)
lines(c(76,93),exp(c(beta[1]+beta[2],beta[1]+beta[2])),lwd=2)
text(56,exp(beta[1]+beta[2]),paste("mean=",round((beta[1]+beta[2]),1)),adj=-1)

detach(case2202)
```

---

ex0112 *Fish Oil and Blood Pressure*

---

## Description

Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet and a standard oil diet. These data are the reductions in diastolic blood pressure.

## Usage

```
ex0112
```

## Format

A data frame with 14 observations on the following 2 variables.

**BP** reduction in diastolic blood pressure (in mm of mercury)

**Diet** factor variable indicating the diet that the subject followed, with levels `"FishOil"` and `"RegularOil"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Knapp, H.R. and FitzGerald, G.A. (1989). The Antihypertensive Effects of Fish Oil, *New England Journal of Medicine* **320**: 1037–1043.

## Examples

```
str(ex0112)
```

---

# ex0116 *Gross Domestic Product (GDP) per Capita*

---

## Description

The data are the gross domestic product per capita for 228 countries in 2010.

## Usage

```
ex0116
```

## Format

A data frame with 228 observations on the following 3 variables.

**Rank** rank order of country from highest to lowest GDP

**Country** name of country

**PerCapitaGDP** per capita GDP in $US

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Central Intelligence Agency, Country Comparison: GDP per capita (PPP), **The World Factbook**, https://www.cia.gov/library/publications/theworld-factbook/rankorder/2004rank.html (retrieved June 30,2011).

## Examples

```
str(ex0116)
```

ex0125 *Zinc concentrations for two groups of rats*

## Description

The data are the zinc concentrations (in mg/ml) in the blood of rats that received a dietary supplement and rats that did not receive the supplement.

## Usage

```
ex0125
```

## Format

A data frame with 39 observations on the following 2 variables.

**Group** a factor representing the group, with levels `"A"` for the dietary supplement group and `"B"` for the control group

**Zinc** measured zinc concentration in mg/ml

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex0125)
```

ex0126 *Environmental Voting of Democrats and Republicans in the U.S. House of Representatives*

## Description

The data are the number of pro- and anti-environmental votes, according to the League of Conservation Voters, for each member of the U.S. House of Representatives in 2005, 2006, or 2007.

## Usage

```
ex0126
```

## Format

A data frame with 492 observations on the following 10 variables.

**State** the state that the member represented

**Representative** name of the representative

**Party** a factor representing political party, with levels `"R"` for Republican, `"D"` for Democratic, and `"I"` for Independent

**Pro05** the number of pro-environmental votes in 2005

**Anti05** the number of anti-environmental votes in 2005

**Pro06** the number of pro-environmental votes in 2006

**Anti06** the number of anti-environmental votes in 2006

**Pro07** the number of pro-environmental votes in 2007

**Anti07** the number of anti-environmental votes in 2007

**PctPro** the total percentage of a representative's votes between 2005 and 2007 that were deemd to be pro-environmental

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

`ex0127`

## Examples

```
str(ex0126)
```

---

# ex0127 — *Environmental Voting of Democrats and Republicans in the U.S. Senate*

---

## Description

The data are the number of pro- and anti-environmental votes, according to the League of Conservation Voters, for each member of the U.S. Senate in 2005, 2006, or 2007.

## Usage

```
ex0127
```

## Format

A data frame with 112 observations on the following 10 variables.

**State** the state that the member represented

**Senator** name of the senator

**Party** a factor representing political party, with levels `"R"` for Republican, `"D"` for Democratic, and `"I"` for Independent

**Pro2005** the number of pro-environmental votes in 2005

**Anti2005** the number of anti-environmental votes in 2005

**Pro2006** the number of pro-environmental votes in 2006

**Anti2006** the number of anti-environmental votes in 2006

**Pro2007** the number of pro-environmental votes in 2007

**Anti2007** the number of anti-environmental votes in 2007

**PctPro** the total percentage of a representative's votes between 2005 and 2007 that were deemd to be pro-environmental

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

`ex0126`

## Examples

```
str(ex0127)
```

---

`ex0211` *Lifetimes of Guinea Pigs*

---

## Description

The data are survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli.

## Usage

```
ex0211
```

## Format

A data frame with 122 observations on the following 2 variables.

**Lifetime** survival time of guinea pig (in days)

**Group** a factor with levels `"Bacilli"` and `"Control"`, indicating the group to which the guinea pig was assigned

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Doksum, K. (1974). Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two–sample Case, *Annals of Statistics* **2**: 267–277.

## Examples

```
str(ex0211)
```

---

`ex0218` *Peter and Rosemary Grant's Finch Beak Data*

---

## Description

In the 1980s, biologists Peter and Rosemary Grant caught and measured all the birds from more than 20 generations of finches on the Galapagos island of Daphne Major. In one of those years, 1977, a severe drought caused vegetation to wither, and the only remaining food source was a large, tough seed, which the finches ordinarily ignored. Were the birds with larger and stronger beaks for opening these tough seeds more likely to survive that year, and did they tend to pass this characteristic to their offspring? The data are beak depths (height of the beak at its base) of 751 finches caught the year before the drought (1976) and 89 finches captured the year after the drought (1978).

## Usage

```
ex0218
```

## Format

A data frame with 840 observations on the following 2 variables.

**Year** Year the finch was caught, 1976 or 1978

**Depth** Beak depth of the finch (mm)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Grant, P. (1986). *Ecology and Evolution of Darwin's Finches*, Princeton University Press, Princeton, N.J.

## See Also

`case0201`

## Examples

```
str(ex0218)
```

---

ex0221 *Bumpus's Data on Natural Selection*

---

## Description

As evidence in support of natural selection, Bumpus presented measurements on house sparrows brought to the Anatomical Laboratory of Brown University after an uncommonly severe winter storm. Some of these birds had survived and some had perished. Bumpus asked whether those that perished did so because they lacked physical characteristics enabling them to withstand the intensity of that particular instance of selective elimination. The data are on the the weights, in grams, for the 24 adult male sparrows that perished and for the 35 adult males that survived.

## Usage

```
ex0221
```

## Format

A data frame with 59 observations on the following 2 variables.

**Humerus** humerus length of adult male sparrows (inches)

**Status** factor variable indicating whether the sparrow perished or survived in a winter storm, with levels `Perished` and `Survived`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

`ex2016`

## Examples

```
str(ex0221)
```

ex0222 *Male and Female Intelligence*

## Description

These data are armed Forces Qualifying Test (AFQT) score percentiles and component test scores in arithmetic reasoning, word knowledge, paragraph comprehension, and mathematical knowledge for a sample of 1,278 U.S. women and 1,306 U.S. men in 1981.

## Usage

```
ex0222
```

## Format

A data frame with 2,584 observations on the following 6 variables.

**Gender** a factor with levels `"female"` and `"male"`

**Arith** score on the arithmetic reasoning component of the AFQT test

**Word** score on the word knowledge component

**Parag** score on the paragraph comprehension component

**Math** score on the mathematical knowledge component

**AFQT** percentile score on the AFQT test

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0330, ex0331, ex0524, ex0525, ex0828, ex0923, ex1033, ex1223

## Examples

```
str(ex0222)
```

ex0223 *Speed Limits and Traffic Fatalities*

## Description

The National Highway System Designation Act was signed into law in the United States on November 28, 1995. Among other things, the act abolished the federal mandate of 55 mile per hour maximum speed limits on roads in the United States and permitted states to establish their own limits. Of the 50 states (plus the District of Columbia), 32 increased their speed limits at the beginning of 1996 or sometime during 1996. These data are the percentage changes in interstate highway traffic fatalities from 1995 to 1996.

## Usage

```
ex0223
```

## Format

A data frame with 51 observations on the following 5 variables.

**State** US state

**Fatalities1995** number of traffic fatalities in 1995

**Fatalities1996** number of traffic fatalities in 1996

**PctChange** percentage change in interstate traffic fatalities between 1995 and 1996

**SpeedLimit** a factor with levels `"Inc"` and `"Ret"`, indicating whether the state increased or retained its speed limit

## Source

Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*, Duxbury.

## References

Report to Congress: The Effect of Increased Speed Limits in the Post-NMSL Era, National Highway Traffic Safety Administration, February, 1998; available in the reports library at http://www-fars.nhtsa.dot.gov/.

## Examples

```
str(ex0223)
```

## ex0321 *Umpire Life Lengths*

### Description

Researchers collected historical and current data on umpires to investigate their life expectancies following the collapse and death of a U.S. major league baseball umpire. They were investigating speculation that stress associated with the job posed a health risk. Data were found on 227 umpires who had died or had retired and were still living. The data set includes the dates of birth and death.

### Usage

```
ex0321
```

### Format

A data frame with 227 observations on the following 3 variables.

**Lifelength** observed lifetime for those umpires who had died by the time of the study or current age of those still living

**Censored** 0 for those who had died by the time of the study or 1 for those who were still living

**Expected** length from actuarial life tables for individuals who were alive at the time the person first became an umpire

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Cohen, R.S., Kamps, C.A., Kokoska, S., Segal E.M. and Tucker, J.B.(2000). Life Expectancy of Major League Baseball Umpires, *The Physician and Sportsmedicine* **28**(5): 83–89.

### Examples

```
str(ex0321)
```

## ex0323 *Solar Radiation and Skin Cancer*

### Description

Data contains yearly skin cancer rates (per 100,000 people) in Connecticut from 1938 to 1972 with a code indicating those years that came two years after higher than average sunspot activity and those years that came two years after lower than average sunspot activity.

### Usage

```
ex0323
```

## Format

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

**Year** year

**CancerRate** skin cancer rate per 100,000 people

**SunspotActivity** a factor with levels `"High"` and `"Low"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from many Fields for the Student and Research Worker*, Springer-Verlag.

## Examples

```
str(ex0323)
```

---

`ex0327` *Pollen Removal*

---

## Description

As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily.

## Usage

```
ex0327
```

## Format

A data frame with 47 observations on the following 3 variables.

**PollenRemoved** proportion of pollen removed

**DurationOfVisit** duration of visit (in seconds)

**BeeType** factor variable with levels `"Queen"` and `"Worker"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Harder, L.D. and Thompson, J.D. (1989). Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants, *American Naturalist* **133**: 323–344.

## Examples

```
str(ex0327)
```

---

ex0330 *Education and Income*

---

## Description

The data are incomes in U.S. dollars for 1,020 working Americans who had 12 years of education and 406 working Americans who had 16 years of education, in 2005.

## Usage

```
ex0330
```

## Format

A data frame with 1,426 observations on the following 3 variables.

**Subject** a subject identification number

**Educ** number of years of education–either 12 or 16

**Income2005** income, in dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0331, ex0524, ex0525, ex0828, ex0923, ex1033, ex1223

## Examples

```
str(ex0330)
```

## ex0331 *Education and Income*

### Description

The data are incomes in U.S. dollars for 406 working Americans who had 16 years of education and 374 working Americans who had more than 16 years of education, in 2005.

### Usage

```
ex0331
```

### Format

A data frame with 780 observations on the following 3 variables.

**Subject** a subject identification number

**Educ** factor with levels `"16"` and `">16"`

**Income2005** income, in dollars

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

### See Also

ex0222, ex0330, ex0524, ex0525, ex0828, ex0923, ex1033, ex1223

### Examples

```
str(ex0331)
```

## ex0332 *College Tuition*

### Description

In-state and out-of-state tuition in dollars for random samples of 25 private and 25 public U.S. colleges and universities in 2011-2012.

### Usage

```
ex0332
```

## Format

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

**College** name of the college

**Type** a factor with levels `"Private"` and `"Public"`

**InState** in-state tuition in dollars

**OutOfState** out-of-state tuition in dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

College Board: http://www.collegeboard.com/student/ (11 July 2011)

## Examples

```
str(ex0332)
```

---

# ex0333 *Brain Size and Litter Size*

---

## Description

Relative brain weights for 51 species of mammal whose average litter size is less than 2 and for 45 species of mammal whose average litter size is greater than or equal to 2.

## Usage

```
ex0333
```

## Format

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

**BrainSize** relative brain sizes (1000 * Brain weight/Body weight) for 96 species of mammals

**LitterSize** factor variable with levels `"Small"` and `"Large"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Sacher, G.A. and Staffeldt, E.F. (1974). Relation of Gestation Time to Brain Weight for Placental Mammals: Implications for the Theory of Vertebrate Growth, *American Naturalist* **108**: 593–613.

## See Also

case0902

## Examples

```
str(ex0333)
```

---

ex0428 *Darwin's Data*

---

## Description

Plant heights (inches) for 15 pairs of plants of the same age, one of which was grown from a seed from a cross-fertilized flower and the other of which was grown from a seed from a self-fertilized flower.

## Usage

```
ex0428
```

## Format

A data frame with 15 observations on the following 2 variables.

**Cross** height (inches) of cross-fertilized plant

**Self** height (inches) of self-fertilized plant

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from many Fields for the Student and Research Worker*, Springer-Verlag.

## Examples

```
str(ex0428)
```

## ex0429 *Salvage Logging*

### Description

The data are the number of tree seedlings per transect in nine logged (L) and seven unlogged (U) plots affected by the Oregon Biscuit Fire, in 2004 and 2005, and the percentage of seedlings lost between 2004 and 2005. The goal is to see whether the distribution of seedlings lost differs in logged and unlogged plots.

### Usage

```
ex0429
```

### Format

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

**Plot** an identification code for plot

**Action** a factor with levels `"L"` for logged and `"U"` for unlogged

**Seedlings2004** the number of seedlings in the plot in 2004

**Seedlings2005** the number of seedlings in the plot in 2005

**PercentLost** the percentage of 2004 seedlngs that were lost

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Donato, D.C., Fontaine, J.B., Campbell, J.L., Robinson, W.D., Kauffman, J.B., and Law, B.E. (2006). Post-Wildfire Logging Hinders Regeneration and Increases Fire Risk, *Science* **311**: 352.

### Examples

```
str(ex0429)
```

## ex0430 *Sunlight Protection Factor*

### Description

Tolerance to sunlight (in minutes) for 13 patients prior to and after treatment with a sunscreen.

### Usage

```
ex0430
```

## Format

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

**PreTreatment** tolerance to sunlight (minutes) prior to sunscreen application

**Sunscreen** tolerance to sunlight (minutes) after sunscreen application

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Fusaro, R.M. and Johnson, J.A. (1974). Sunlight Protection for Erythropoietic Protoporphyria Patients, *Journal of the American Medical Association* **229**(11): 1420.

## Examples

```
str(ex0430)
```

---

# ex0431 *Effect of Group Therapy on Survival of Breast Cancer Patients*

---

## Description

Researchers randomly assigned metastatic breast cancer patients to either a control group or a group that received weekly 90 minute sessions of group therapy and self-hypnosis, to see whether the latter treatment improved the patients' quality of life.

## Usage

```
ex0431
```

## Format

A data frame with 58 observations on the following 3 variables.

**Survival** months of survival after beginning of study

**Group** a factor with levels `"Control"` and `"Therapy"`

**Censor** 0 if entire lifetime observed, 1 if patient known to have lived at least 122 months

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Spiegel, D., Bloom, J.R., Kraemer, H.C. and Gottheil, E. (1989). Effect of Psychosocial Treatment on Survival of Patients with Metastatic Breast Cancer, *Lancet* **334**(8668): 888–891.

## Examples

```
str(ex0431)
```

## ex0432 *Therapeutic Marijuana*

### Description

To investigate the capacity of marijuana to reduce the side effects of cancer chemotherapy, researchers performed a double-blind, randomized, crossover trial. Fifteen cancer patients on chemotherapy were randomly assigned to receive either a marijuana treatment or a placebo treatment after their first three sessions of chemotherapy. They were then crossed over to the opposite treatment for their next 3 sessions.

### Usage

```
ex0432
```

### Format

A data frame with 15 observations on the following 3 variables.

**Subject** subject number 1–15

**Marijuana** total number of vomiting and retching episodes under marijuana treatment

**Placebo** total number of vomiting and retching episodes under placebo treatment

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Chang, A.E., Shiling, D.J., Stillman, R.C., Goldberg, N.H., Seipp, C.A., Barofsky, I., Simon, R.M. and Rosenberg, S.A. (1979). Delta-9-Tetrahydrocannabinol as an Antiemetic in Cancer Patients Receiving High Dose Methotrexate, *Annals of Internal Medicine* **91**(6): 819–824.

### Examples

```
str(ex0432)
```

## ex0518 *Fatty Acid*

### Description

A randomized experiment was performed to estimate the effect of a certain fatty acid CPFA on the level of a certain protein in rat livers.

### Usage

```
ex0518
```

## Format

A data frame with 30 observations on the following 4 variables.

**Protein** levels of protein (x 10) found in rat livers

**Treatment** a factor with levels `"Control"`, `"CPFA50"`, `"CPFA150"`, `"CPFA300"`, `"CPFA450"` and `"CPFA600"`

**Day** a factor with levels `"Day1"`, `"Day2"`, `"Day3"`, `"Day4"` and `"Day5"`

**TrtDayGroup** a factor with levels `"Group1"`, `"Group2"`, ..., `"Group10"`; the observed levels of the Treatment and Day interaction

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex0518)
```

---

`ex0523` *Was Tyrannosaurus Rex Warm-Blooded?*

---

## Description

Data frame with measurements of oxygen isotopic composition of vertebrate bone phosphate (per mil deviations from SMOW) in 12 bones of a singe Tyrannosaurus rex specimen

## Usage

```
ex0523
```

## Format

A data frame with 52 observations on the following 2 variables.

**Oxygen** oxygen isotopic composition

**Bone** a factor with levels `"Bone1"`, `"Bone2"`, ..., `"Bone12"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Barrick, R.E. and Showers, W.J. (1994). Thermophysiology of *Tyrannosaurus rex*: Evidence from Oxygen Isotopes, *Science* **265**(5169): 222–224.

## See Also

`ex1120`

## Examples

```
str(ex0523)
```

---

`ex0524` *IQ and Future Income*

---

## Description

These data are annual incomes in 2005 for 2,584 Americans who were selected in the National Longitudinal Study of Youth 1979, who were available for re- interview in 2006, and who had paying jobs in 2005, along with the quartile of their AFQT (IQ) test score taken in 1981. How strong is the evidence that the distributions of 2005 annual incomes differ in the four populations? By how many dollars or by what percent does the distribution of 2005 incomes for those within the highest (fourth) quartile of IQ test scores exceed the distribution for the lowest (first) quartile?

## Usage

```
ex0524
```

## Format

A data frame with 2,584 observations on the following 3 variables.

**Subject** subject identification number

**IQquartile** a factor with levels `"1stQuartile"`, `"2ndQuartile"`, `"3rdQuartile"` and `"4thQuartile"`

**Income2005** annual income in U.S. dollars, 2005

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0525, ex0828, ex0923, ex1033, ex1223

## Examples

```
str(ex0524)
```

ex0525 *IQ and Future Income*

## Description

These data are annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005. The data set also includes a code for the number of years of education that each individual had completed by 2006: <12, 12, 13–15, 16, and >16. How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others? By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

## Usage

```
ex0525
```

## Format

A data frame with 2,584 observations on the following 3 variables.

**Subject** subject identification number

**Educ** a factor for years of education category with levels "<12", "12", "13-15", "16" and ">16"

**Income2005** Annual income in 2005, in U.S. dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0524, ex0828, ex0923, ex1033, ex1223

## Examples

```
str(ex0525)
```

## ex0623 *Diet Wars*

### Description

These data are simulated to match the summary and conclusions of a real study of overweight employees who were randomly assigned to three diet groups: a low-fat diet, a low-carb diet (similar to the Atkins diet), and a Mediterranean diet. The study ran for two years, with 272 employees completing the entire protocol. Is there evidence of differences in average weight loss among these diets? If so, which diets appear to be better than which others?

### Usage

```
ex0623
```

### Format

A data frame with 272 observations on the following 3 variables.

**Subject** subject identification number

**Group** a factor with levels `"Low-Carbohydrate"`, `"Low-Fat"`, and `"Meditrranean"`

**WtLoss24** weight at the end of the 24 month study minus initial weight, in kg

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### See Also

ex1420, ex1921, ex1922

### Examples

```
str(ex0623)
```

## ex0624 *A Biological Basis for Homosexuality*

### Description

Is there a physiological basis for sexual preference? Researchers measured the volumes of four cell groups in the interstitial nuclei of the anterior hypothalamus in postmortem tissue from 41 subjects at autopsy from seven metropolitan hospitals in New York and California.

### Usage

```
ex0624
```

## Format

A data frame with 41 observations on the following 5 variables.

**Volume** volumes of INAH3 ($1000 \times \text{mm}^3$) cell clusters from 41 humans

**Group** a factor with levels

| | |
|---|---|
| `"Group1"` | heterosexual male with AIDS death |
| `"Group2"` | heterosexual male with Non-AIDS death |
| `"Group3"` | homosexual male with AIDS death |
| `"Group4"` | heterosexual female with AIDS death |
| `"Group5"` | heterosexual female with Non-AIDS death |

**Sex** a factor with levels `"Female"` and `"Male"`

**Orientation** a factor with levels `"Heterosexual"` and `"Homosexual"`

**Death** a factor with levels `"AIDS"` and `"Non-AIDS"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

LeVay, S. (1991). A Difference in Hypothalamic Structure Between Heterosexual and Homosexual Men, *Science* **253**(5023): 1034–1037.

## Examples

```
str(ex0624)
```

---

`ex0721` *Planetary Distances and Order from the Sun*

---

## Description

The first three columns are the names, orders of distance from the sun and distances from the sun (scaled so that earth is 1) of the 8 planets in our solar system and the dwarf planet, Pluto. The next three columns are the same, but also include the asteroid belt.

## Usage

```
ex0721
```

## Format

A data frame with observations on the following 6 variables.

**Name** name of object in solar system, 9 objects

**Order** order of object's distance from the sun

**Distance** distance of object from sun, with earth = 1

**Name2** name of object in solar system, including asteroid belt

**Order2** order of object's distance from the sun

**Distance2** distance of object from sun, with earth = 1

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex2226

## Examples

```
str(ex0721)
```

---

# ex0722 *Crab Claw Size and Force*

---

## Description

As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propdus heights of the claws on several crabs of three species.

## Usage

```
ex0722
```

## Format

A data frame with 38 observations on the following 3 variables.

**Force** closing strength of claw of the crab

**Height** propodus height of claw of the crab

**Species** species to which the crab belongs

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Yamada, S.B. and Boulding, E.G. (1992). Claw Morphology, Prey Size Selection and Foraging Efficiency in Generalist and Specialist Shell-Breaking Crabs. *Journal of Experimental Marine Biology and Ecolog*, **220** 191–211.

## Examples

```
str(ex0722)
```

## ex0724 *Decline in Male Births*

### Description

The data are on the proportion of male birts in Denmark, The Netherlands, Canada and the United States for a number of yeras. Notice that the proportions for Canada and the United States are only provided for the years 1970 to 1990, while Denmark and The Netherlands have data listed for 1950 to 1994.

### Usage

```
ex0724
```

### Format

A data frame with 45 observations on the following 5 variables.

**Year** year of observation

**Denmark** male birth rate of Denmark for given year

**Netherlands** male birth rate of The Netherlands for given year

**Canada** male birth rate of Canada for given year

**USA** male birth rate of the United States for given year

### Source

Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*, Duxbury.

### References

Davis, D.L., Gottlieb, M.B. and Stampnitzky, J.R. (1998). Reduced ratio of male to female births in several industrial countries, *Journal of the American Medical Association* **279**(13): 1018–1023.

### Examples

```
str(ex0724)
```

## ex0725 *The Big Bang II*

### Description

These data are measured distances and recession velocities for 10 clusters of nebulae, much farther from earth than the nebulae reported in case0701.

### Usage

```
ex0727
```

### Format

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

**Distance** distance from earth (in million parsec)

**Velocity** recession velocity (in kilometres per second)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Hubble, E. and Humason, M. (1931). The Velocity–Distance Relation Among Extra–calactic Nebulae, *Astrophysics Journal* **74**: 43–50.

### See Also

`case0701`

### Examples

```
str(ex0725)
```

---

## ex0726 *Orign of the Term Regression*

---

### Description

These data are heights of 933 adults and their parents, as measured by Karl Pearson in 1885.

### Usage

```
ex0726
```

### Format

A data frame with 933 observations on the following 5 variables.

**Gender** a factor with levels `"female"` and `"male"`

**Family** an identification number for family, 1, 2,. . . , 205

**Height** adult height of the child, inches

**Father** height of the child's father, inches

**Mother** height of the child's mother, inches

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Hubble, E. and Humason, M. (1931). The Velocity–Distance Relation Among Extra–calactic Nebulae, *Astrophysics Journal* **74**: 43–50.

## Examples

```
str(ex0725)
```

---

`ex0727` *Male Displays*

---

## Description

Black wheatears are small birds in Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. This 35–gram bird transports, on average, 3.1 kg of stones per nesting season! Different males carry somewhat different sized stones, prompting a study on whether larger stones may be a signal of higher health status. Soler et al. calculated the average stone mass (g) carried by each of 21 male black wheatears, along with T-cell response measurements reflecting their immune systems' strengths.

## Usage

```
ex0727
```

## Format

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

**Mass** average mass of stones carried by bird (in g)

**Tcell** T-cell response measurement (in mm)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Soler, M., Martín-Vivaldi, M., Marín J. and Møller, A. (1999). Weight lifting and health status in the black wheatears, *Behavioral Ecology* **10**(3): 281–286.

## Examples

```
str(ex0727)
```

ex0728 *Brain Activity in Violin and String Players*

## Description

Studies over the past two decades have shown that activity can effect the reorganisation of the human central nervous system. For example, it is known that the part of the brain associated with activity of a finger or limb is taken over for other purposes in individuals whose limb or finger has been lost. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of nine string players (six violinists, two cellists and one guitarist) and six controls who had never played a musical instrument, when the thumb and fifth finger of the left hand were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers of their left hand extensively, might show different behaviour—as a result of this extensive physical activity—than individuals who did not play stringed instruments.

## Usage

```
ex0728
```

## Format

A data frame with 15 observations on the following 2 variables.

**Years** years that the individual has been playing

**Activity** neuronal activity index

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B. and Taub E. (1995). Increased cortical representation of the fingers of the left hand in string players, *Science* **270**(5234): 305–307.

## Examples

```
str(ex0728)
```

ex0729 *Sampling Bias in Exit Polls*

## Description

These data are the number of percentage points by which exit polls over estimated the actual vote for candidate John Kerry in the 2004 U.S. presidential election, grouped according to the distance of the exit poll interviewer from the door of the polling location. How strong is the evidence that the mean Kerry overestimate increases with increasing distance of interviewer from the door (thus lending evidence to the theory that supporters of the other candidate, George W Bush, were more inclined to avoid exit pollsters)?

## Usage

```
ex0729
```

## Format

A data frame with 6 observations on the following 2 variables.

**OverEstimate** number of percentage points by which the exit poll estimate exceeded the actual percentage voting for Kerry (in all precincts with a similar distance of interviewer from the door

**Distance** distance of the interviewer from the door of the polling location, in feet

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Evaluation of Edison/Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005. http://abcnews.go.com/images/Politics/EvaluationofEdisonMitofskyElectionSystem.pdf

## See Also

ex0730

## Examples

```
str(ex0729)
```

---

# ex0730 *Sampling Bias in Exit Polls 2*

---

## Description

These data are the average proportion of voters refusing to be interviewed by exit pollsters in the 2004 U.S. presidential election, grouped gby age of the interviewer, and the approximate age of the interviewer. What evidence do the data provide that the mean refusal rate decreased with incrasing age of interviewer?

## Usage

```
ex0730
```

## Format

A data frame with 6 observations on the following 2 variables.

**Age** age of the exit poll interviewer, years

**Refusal** average proportion of voters refusing to be interviewed

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Evaluation of Edison/Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005. http://abcnews.go.com/images/Politics/EvaluationofEdisonMitofskyElectionSystem.pdf

## See Also

ex0729

## Examples

```
str(ex0730)
```

---

# ex0816 *Meat Processing*

---

## Description

The data in case0702 are a subset of the complete data on postmortum pH in 12 steer carcasses.

## Usage

```
ex0816
```

## Format

A data frame with 12 observations on the following 2 variables.

**Time** time after slaughter (hours)

**pH** pH level in postmortem muscle

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Schwenke, J.R. and Milliken, G.A. (1991). On the Calibration Problem Extended to Nonlinear Models, *Biometrics* **47**(2): 563–574.

## See Also

case0702

## Examples

```
str(ex0816)
```

## ex0817 *Biological Pest Control*

### Description

In a study of the effectiveness of biological control of the exotic weed tansy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragwort. Harvesting the plots the next season, they measured the average dry mass of ragwort remaining (grams/plant) and the flea beetle load (beetles/gram of ragwort dry mass) to see if the ragwort plants in plots with high flea beetle loads were smaller as a result of herbivory by the beetles.

### Usage

```
ex0817
```

### Format

A data frame with 15 observations on the following 2 variables.

**Load** flee beetle load (in beetles/gram of ragwort dry mass)

**Mass** dry mass of ragwort weed

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

McEvoy, P. and Cox, C. (1991). Successful Biological Control of Ragwort, *Senecio jacobaea*, by introducing insects in Oregon, *Ecological Applications* **1**(4): 430–442.

### Examples

```
str(ex0817)
```

## ex0820 *Quantifying Evidence for Outlierness*

### Description

The data are Democratic and Republican vote counts, by (a) absentee ballot and (b) voting machine, for 22 elections in Philadelphia's senatorial districts between 1982 and 1993.

### Usage

```
ex0820
```

## Format

A data frame with 22 observations on the following 2 variables.

**Year** Year of election

**District** a factor with levels `"D1"`, `"D2"`, `"D3"`, `"D4"`, `"D5"`, `"D7"`, and `"D8"`

**DemAbsenteeVotes** Number of absentee ballots indicating a vote for the Democratic candidate

**RepubAbsenteeVotes** Number of absentee ballots indicating a vote for the Republican candidate

**DemMachineVotes** Number of machine-counted ballots indicating a vote for the Democratic candidate

**RepubMachineVotes** Number of machine-coutned ballots indicating a vote for the Republican candidate

**DemPctOfAbsenteeVotes** Percentage of absentee ballots indicating a vote for the Democratic candidate

**DemPctOfMachineVotes** Percentage of machine-counted ballots indicating a vote for the Democratic candidate

**Disputed** a factor taking on the value `"yes"` for the disputed election and `"no"` for all other elections

## Details

In a special election to fill a Pennsylvania State Senate seat in 1993, the Democrat, William Stinson, received 19,127 machine–counted votes and the Republican, Bruce Marks, received 19,691. In addition, there were 1,391 absentee ballots for Stinson and 366 absentee ballots for Marks, so that the total tally showed Stinson the winner by 461 votes. The large disparity between the machine–counted and absentee votes, and the resulting reversal of the outcome due to the absentee ballots caused some concern about possible illegal influence on the absentee votes. To see whether the discrepancy in absentee votes was larger than could be explained by chance, an econometrician considered the data given in this data frame (read from a graph in *The New York Times*, 11 April 1994).

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Ashenfelter, O (1994). Report on Expected Absentee Ballots. Department of Economics, Princeton University. See also Simon Jackman (2011). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.03.10. http://pscl.stanford.edu/

## Examples

```
str(ex0820)
```

## ex0822 *Ecosystem Decay*

### Description

Data are the number of butterfly species in 16 islands of forest of various sizes in otherwise cleared areas in Brazil.

### Usage

```
ex0822
```

### Format

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

**Area** area (ha) of forest patch

**Species** number of butterfly species

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Lovejoy, T.E., Rankin, J.M., Bierregaard, Jr., R.O., Brown, Jr., K.S., Emmons, L.H. and van der Voort, M. (1984). Ecosystem decay of Amazon forest remnants *in* Nitecki, M.H. (ed.) *Extinctions*, University of Chicago Press.

### Examples

```
str(ex0822)
```

## ex0823 *Wine Consumption and Heart Disease*

### Description

The data are the average wine consumption rates (in liters per person per year) and number of ischemic heart disease deaths (per 1000 men aged 55 to 64 years) for 18 industrialized countries.

### Usage

```
data(ex0823)
```

## Format

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

**Country** a character vector indicating the country

**Wine** consumption of wine (liters per person per year)

**Mortality** heart disease mortality rate (deaths per 1,000)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

St. Leger A.S., Cochrane, A.L. and Moore, F. (1979). Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine, *Lancet*: 1017–1020.

## Examples

```
str(ex0823)
```

---

# ex0824 *Respiratory Rates for Children*

---

## Description

A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is "high" however, a physician must have a clear picture of the distribution of normal rates. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years.

## Usage

```
ex0824
```

## Format

A data frame with 618 observations on the following 2 variables.

**Age** age in months of child

**Rate** respiratory rate (breaths per minute)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rusconi, F., Castagneto, M., Porta, N., Gagliardi, L., Leo, G., Pellegatta, A., Razon, S. and Braga, M. (1994). Reference Values for Respiratory Rate in the First 3 Years of Life, *Pediatrics* **94**(3): 350–355.

## Examples

```
str(ex0824)
```

---

ex0825 *The Dramatic U.S. Presidential Election of 2000*

---

## Description

Data set shows the number of votes for Buchanan and Bush in all 67 counties in Florida during the U.S. presidential election of November 7, 2000.

## Usage

```
ex0825
```

## Format

A data frame with 67 observations on the following 3 variables.

**County** a character vector indicating the county

**Buchanan2000** votes cast for P. Buchanan

**Bush2000** votes cast for G.W. Bush

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex1222

## Examples

```
str(ex0825)
```

---

ex0826 *Kleiber's Law*

---

## Description

The data are the average mass, metabolic rate, and lifespan for 95 species of mammals. Kleiber's law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of 3/4.

## Usage

```
ex0826
```

## Format

A data frame with 95 observations on the following 5 variables.

**CommonName** the common name of the mammal species

**Species** the scientific name of the mammal species

**Mass** the average body mass in kg

**Metab** the average metabolic rate in kJ per day

**Life** the average lifespan in years

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex0826)
```

---

# ex0828 *IQ, Education, and Future Income*

---

## Description

These data are armed Forces Qualifying Test (AFQT) score percentiles, years of education, and annual income in 2005 for a subset of a random sample of 2,584 Americans selected in 1979 who were working in 2005 and re-interviewed in 2006.

## Usage

```
ex0828
```

## Format

A data frame with 2,584 observations on the following 4 variables.

**Subject** the subject identification number

**AFQT** percentile score on the AFQT test

**Educ** years of education achieved by 2005

**Income2005** annual income in 2005

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0524, ex0525, ex0923, ex1033, ex1223

## Examples

```
str(ex0828)
```

---

ex0829 *Autism Rates*

---

## Description

These data are the prevalence of autism per 10,000 ten-year old children in the United States in 1992, 1994, 1996, 1998, and 2000.

## Usage

```
ex0829
```

## Format

A data frame with 5 observations on the following 2 variables.

**Year** year

**Prevalence** the number of autism cases per 10,000 ten-year old children

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Newschaffer, C. J., Falb, M. D. and Gurney, J. G. (2005) National Autism Prevalence Trends From United States Special Education Data, *Pediatrics* **115**: e277–e282.

## Examples

```
str(ex0829)
```

## ex0914 *Pace of Life and Heart Disease*

### Description

In four regions of the US (Northeast, Midwest, South and West), in three different sized metropolitan regions, researchers measured indicators of pace of life.

### Usage

```
ex0914
```

### Format

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

**Bank** bank clerk speed

**Walk** pedestrian walking speed

**Talk** postal clerk talking speed

**Heart** age adjusted death rate due to heart disease

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Levine, R.V. (1990). The Pace of Life, *American Scientist* **78**: 450–459.

### Examples

```
str(ex0914)
```

## ex0915 *Rainfall and Corn Yield*

### Description

Data on corn yield and rainfall in six U.S. corn–producing states (Iowa, Nebraska, Illinois, Indiana, Missouri and Ohio), recorded for each year from 1890 to 1927.

### Usage

```
ex0915
```

## Format

A data frame with 38 observations on the following 3 variables.

**Year** year of observation (1890–1927)

**Yield** average corn yield for the six states (in bu/acre)

**Rainfall** average rainfall in the six states (in in/year)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Ezekiel, M. and Fox, K.A. (1959). *Methods of Correlation and Regression Analysis*, John Wiley & Sons, New York.

## Examples

```
str(ex0915)
```

---

# ex0918 *Speed of Evolution*

---

## Description

Researchers studied the development of a fly (Drosophila subobscura) that had been accidentally introduced from the Old World into North America around 1980.

## Usage

```
ex0918
```

## Format

A data frame with 21 observations on the following 8 variables.

**Continent** a factor with levels `"NA"` and `"EU"`

**Latitude** latitude (degrees)

**Females** average wing size ($10^3 \times$log mm) of female flies on log scale

**SE_Females** standard error of wing size ($10^3 \times$log mm) of female flies on log scale

**Males** average wing size ($10^3 \times$log mm) of male flies on log scale

**SE_Males** standard error of wing size ($10^3 \times$log mm) of male flies on log scale

**Ratio** average basal length to wing size ratios of female flies

**SE_Ratio** standard error of average basal length to wing size ratio of female flies

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Huey, R.B., Gilchrist, G.W., Carlson, M.L., Berrigan, D. and Serra, L. (2000). Rapid Evolution of a Geographic Cline in Size in an Introduced Fly, *Science* **287**(5451): 308–309.

## Examples

```
str(ex0918)
```

---

`ex0920` *Winning Speeds at the Kentucky Derby*

---

## Description

Data set contains the year of the Kentucky Derby, the winning horse, the condition of the track and the average speed of the winner for years 1896–2011.

## Usage

```
ex0920
```

## Format

A data frame with 116 observations on the following 8 variables.

**Year** year of Kentucky Derby

**Winner** a character vector with the name of the winning horse

**Starters** number of horses that started the race

**NetToWinner** the net winnings of the winner, in U.S. dollars

**Time** the winning time in seconds

**Speed** the winning average speed, n miles per hour

**Track** a factor indicating track condition with levels `"Fast"`, `"Good"`, `"Dusty"`, `"Slow"`, `"Heavy"`, `"Muddy"`, and `"Sloppy"`

**Conditions** a factor with with 2 levels of track condition, with levels `"Fast"` and `"Slow"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Kentucky Derby: Kentucky Derby Racing Results.

## Examples

```
str(ex0920)
```

ex0921 *Ingestion Rates of Deposit Feeders*

## Description

The data are the typical dry weight in mg, the typical ingestion rate (weight of food intake per day for one animal) in mg/day, and the percentage of the food that is composed of organic matter for 19 species of deposit feeders. The goal is to see whether the distribution of species' ingestion rates depends on the percentage of organic matter in the food, after accounting for the effect of species weight and to describe the association.

## Usage

```
ex0922
```

## Format

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

**Species** a character variable with the name of the species

**Weight** the dry weight of the species, in mg

**Ingestion** ingestion rate in mg per day

**Organic** percentage of organic matter in the food

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Cammen, L. M. (1980) Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores, *Oecologia* **44**: 303–310.

## See Also

ex1125

## Examples

```
str(ex0921)
```

ex0923 *Comparing Male and Female Incomes, Accounting for Education and IQ*

## Description

These data are a subset of the National Longitudinal Study of Youth data, with annual incomes in 2005, intelligence test scores (AFQT) measured in 1981, and years of education completed by 2006 for 1,306 males and 1,278 females who were between the ages of 14 and 22 when selected for the survey in 1979, who were available for re-interview in 2006, and who had paying jobs in 2005. Is there any evidence that the mean salary for males exceeds the mean salary for females with the same years of education and AFQT scores? By how many dollars or by what percent is the male mean larger?

## Usage

```
ex0923
```

## Format

A data frame with 2,584 observations on the following 5 variables.

**Subject** the subject identification number

**Gender** a factor with levels `"female"` and `"male"`

**AFQT** percentile score on the AFQT intelligence test

**Educ** years of education achieved by 2005

**Income2005** annual income in 2005

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0524, ex0525, ex0828, ex1033, ex1223

## Examples

```
str(ex0923)
```

## ex1014 *Toxic Effects of Copper and Zinc*

### Description

Researchers randomly allocated 25 beakers containing minnow larvae to receive one of 25 treatment combinations of 5 levels of zinc and 5 levels of copper.

### Usage

```
ex1014
```

### Format

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

**Copper** amount of copper received (in ppm)

**Zinc** amount of zinc received (in ppm)

**Protein** protein in minnow larvae exposed to copper and zinc ($\mu$g/larva)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Ryan, D.A., Hubert, J.J., Carter, E.M., Sprague, J.B. and Parrott, J. (1992). A Reduced-Rank Multivariate Regression Approach to Aquatic Joint Toxicity Experiments, *Biometrics* **48**(1): 155–162.

### Examples

```
str(ex1014)
```

## ex1026 *Thinning of Ozone Layer*

### Description

Depletion of the ozone layer allows the most damaging ultraviolet radiation to reach the Earth's surface. To measure the relationship, researchers sampled the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990.

### Usage

```
ex1026
```

## Format

A data frame with 17 observations on the following 3 variables.

**Inhibit** percent inhibition of primary phytoplankton production in water

**UVB** UVB exposure

**Surface** a factor with levels `"Deep"` and `"Surface"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Smith, R.C., Prézelin, B.B., Baker, K.S., Bidigare, R.R., Boucher, N.P., Coley, T., Karentz, D., MacIntyre, S., Matlick, H.A., Menzies, D., Ondrusek, M., Wan, Z. and Waters, K.J. (1992). Ozone Depletion: Ultraviolet Radiation and Phytoplankton Biology in Antarctic Waters, *Science* **255**(5047): 952–959.

## Examples

```
str(ex1026)
```

---

# ex1027 *Factors Affecting Extinction*

---

## Description

Data are measurements on breeding pairs of land-bird species collected from 16 islands around Britain over the course of several decades. For each species, the data set contains an average time of extinction on those islands where it appeared, the average number of nesting pairs, the size of the species and the migratory status of the species.

## Usage

```
ex1027
```

## Format

A data frame with 62 observations on the following 5 variables.

**Species** a character vector indicating the species

**Time** average extinction time in years

**Pairs** average number of nesting pairs

**Size** a factor with levels `"L"` and `"S"`

**Status** a factor with levels `"M"` and `"R"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Pimm, S.L., Jones, H.L., and Diamond, J. (1988). On the Risk of Extinction, *American Naturalist* **132**(6): 757–785.

## Examples

```
str(ex1027)
```

---

`ex1028` *El Nino and Hurricanes*

---

## Description

Data set with the numbers of Atlantic Basin tropical storms and hurricanes for each year from 1950–1997. The variable storm index is an index of overall intensity of hurricane season. Also listed are whether the year was a cold, warm or neutral El Nino year and a variable indicating whether West Africa was wet or dry that year.

## Usage

```
ex1028
```

## Format

A data frame with 48 observations on the following 7 variables.

**Year** year

**ElNino** a factor with levels `"cold"`, `"neutral"` and `"warm"`

**Temperature** numeric variable with values -1 if `ElNino` is `"cold"`, 0 if `"neutral"` and 1 if `"warm"`

**WestAfrica** numeric variable indicating whether West Africa was wet (1) or dry (0)

**Storms** number of storms

**Hurricanes** number of hurricanes

**StormIndex** index of overall intensity of hurricane season

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Data were gathered by William Gray of Colorado State University and reported on USA Today weather page: http://www.usatoday.com/weather/whurnum.htm

## Examples

```
str(ex1028)
```

ex1029 *Wage and Race*

## Description

Data set contains weekly wages in 1987 for a sample of 25,632 males between the age of 18 and 70 who worked full-time along with their years of education, years of experience, indicator variable for whether they were black, indicator variable for whether they worked in or near a city, and a code for the region in the US where they worked.

## Usage

```
ex1029
```

## Format

A data frame with 25,437 observations on the following 6 variables.

**Region** a factor with levels `"Midwest"`, `"Northeast"`, `"South"` and `"West"`

**MetropolitanStatus** a factor with levels `"MetopolitanArea"` and `"NotMetropolitanArea"`

**Exper** experience in years

**Educ** education in years

**Race** a factor with levels `"Black"` and `"NotBlack"`

**WeeklyEarnings** weekly wage in dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Bierens, H.J. and Ginther, D.K. (2001). Integrated Conditional Moment Testing of Quantile Regression Models, *Empirical Economics* **26**(1): 307–324; see http://econ.la.psu.edu/~hbierens/QUANTILE.PDF http://econ.la.psu.edu/~hbierens/MEDIAN.HTM

## Examples

```
str(ex1029)
```

ex1030 *Wage and Race 2011*

## Description

A data set with weekly earnings for 4,952 males between the age of 18 and 70 sampled in the March 2011 Current Population Survey (CPS). These males are a subset who had reported earnings and who responded as having race as either "Only White" or "Only Black." Also recorded are the region of the country (with four categories: Northeast, Midwest, South, and West), the metropolitan status of the men's employment (with three categories: Metropolitan, Not Metropolitan, and Not Identified), age, education category (with 16 categories ranging from "Less than first grade" to "doctorate Degree"), and education code, which is a numerical value that corresponds roughly to increasing levels of education (and so may be useful for plotting). What evidence do the data provide that the distributions of weekly earnings differ in the populations of white and black workers after accounting for the other variables? By how many dollars or by what percent does the White population mean (or median) exceed the Black population mean (or median)?

## Usage

```
ex1030
```

## Format

A data frame with 4,952 observations on the following 7 variables.

**Region** a factor with levels `"Midwest"`, `"Northeast"`, `"South"` and `"West"`

**MetropolitanStatus** a factor with levels `"Metopolitan"`, `"Not Metropolitan "` and `"Not Identified"`

**Age** age in years

**EducationCategory** a factor with 16 levels: `"SomeCollegeButNoDegree"`, `"AssocDegAcadem"`, `"NinthGrade"`, `"BachelorsDegree"`, `"TenthGrade"`, `"HighSchoolDiploma"`, `"AssocDegOccupVocat"`, `"DoctorateDegree"`, `"TwelthButNoDiploma"`, `"LessThanFirstGrade"`, `"EleventhGrade"`, `"ProfSchoolDegree"`, `"FifthorSixthGrade"`,`"SeventhOrEighthGrade"`, `"FirstSecondThirdOrFourthGrade"`

**EducationCode** a numerical variable indicating the approximate ordering of EducationCategory, with higher numbers indicating more education

**Race** a factor with levels `"Black"` and `"White"`

**WeeklyEarnings** weekly wage in dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.gov/cps

## Examples

```
str(ex1030)
```

ex1031 *Who Looks After the Kids*

## Description

A data set with Clutch Volume (cubic milimeters) and adult Body Mass (kg) in six different groups of animals: modern maternal-care bird species (Mat), modern paternal-care bird species (Pat), modern biparental-care bird species (BiPI), modern maternal-care crocodiles (Croc), non-avian maniraptoran dinosaurs thought to be ancestors of modern birds (Mani), and other non-avian dinosaurs (Othr). The question of interest was which group of modern creatures most closely matches the relationship in the maniraptoran dinosaurs.

## Usage

```
ex1031
```

## Format

A data frame with 443 observations on the following 6 variables.

**CommonName** the common name of the species

**Genus** species genus

**Species** species name

**Group** a factor with 6 levels corresponding to the 6 groups of animals: `"BiP"`, `"Croc"`, `"Mani"`, `"Mat"`, `"Othr"`, and `"Pat"`

**BodyMass** the average body mass of individuals in the species (kg)

**ClutchVolume** the total volume of all eggs in a clutch (average value for the species)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Varricchio, D. J., Moore, J.r., Erickson, G.M., Norell, M.A., Jackon, F.D. and Borkowski, J.J. (2008) Avian Paternal Care Had Dinosaur Origin *Science* **322**: 1826–1828

## See Also

ex1923

## Examples

```
str(ex1031)
```

ex1033 *IQ Score and Income*

## Description

This is a subset of the National Longitudinal Study of Youth (NLSY79) data, with annual incomes in 2005 (in U.S. dollars, as Recorded in a 2006 interview); scores on the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge portions of the Armed Forces Vocational Aptitude Battery (ASVAB) of tests taken in 1981; and the percentile score of the Armed Forces Qualifying Test (AFQT), which is a linear combination of the four component tests mentioned above (but note that AFQT reported here is the percentile, which is not a linear combination of the four component scores). Which of the five test scores seem to be the most important predictors of 2005 income? Is the AFQT sufficient by itself?

## Usage

```
ex1033
```

## Format

A data frame with 2,584 observations on the following 7 variables.

**Subject** the subject identification number

**Arith** score on the Arithmetic Reasoning test in 1981

**Word** score on the Word Knowledge Test in 1981

**Parag** score on the Paragraph Comprehension test in 1981

**Math** score on the Mathematics Knowledge test in 1981

**AFQT** percentile score on the AFQT intelligence test in 1981

**Income2005** annual income in 2005

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0524, ex0525, ex0828, ex0923, ex1223

## Examples

```
str(ex1033)
```

## ex1111 *Chernobyl Fallout*

### Description

The data are are the cesium concentrations (in Bq/kg) in soil and in mushrooms at 17 wooded locations in Umbria, Central Italy, from August 1986 to November 1989. Researchers wished to investigate the cesium transfer from contaminated soil to plants after the Chernobyl nuclear power plant accident in April 1986 by describing the distribution of the mushroom concentration as a function of soil concentration.

### Usage

```
ex1111
```

### Format

A data frame with 17 observations on the following 2 variables.

**Mushroom** the cesium concentration in mushrooms, Bq/kg

**Soil** the cesium concentration in soil, Bq/kg

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### Examples

```
str(ex1111)
```

## ex1120 *Was Tyrannosaurus Rex Warm-Blooded?*

### Description

Data are the isotopic composition of structural bone carbonate ($X$) and the isotopic composition of the coexisting calcite cements ($Y$) in 18 bone samples from a specimen of the dinosaur Tyrannosaurus rex. Evidence that the mean of $Y$ is positively associated with $X$ was used in an argument that the metabolic rate of this dinosaur resembled warm-blooded more than cold-blooded animals.

### Usage

```
ex1120
```

### Format

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

**Carbonate** isotopic composition of bone carbonate

**Calcite** isotopic composition of calcite cements

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Barrick, R.E. and Showers, W.J. (1994). Thermophysiology of *Tyrannosaurus rex*: Evidence from Oxygen Isotopes, *Science* **265**(5169): 222–224.

## See Also

ex0523

## Examples

```
str(ex1120)
```

---

# ex1122 *Deforestation and Debt*

---

## Description

It has been theorized that developing countries cut down their forests to pay off foreign debt. Data are debt, deforestation, and population from 11 Latin American nations.

## Usage

```
ex1122
```

## Format

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

**Country** a character vector indicating the country

**Debt** debt (millions of dollars)

**Deforest** deforestation (thousands of ha)

**Pop** population (thousands of people)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Gullison, R.R. and Losos, E.C. (1992). The Role of Foreign Debt in Deforestation in Latin America, *Conservation Biology* **7**(1): 140–7.

## Examples

```
str(ex1122)
```

ex1123 *Air Pollution and Mortality*

## Description

Does pollution kill people? Data in one early study designed to explore this issue from 5 Standard Metropolitan Statistical Areas in the U.S between 1959–1961.

## Usage

```
ex1123
```

## Format

A data frame with 60 observations on the following 7 variables.

**City** a character vector indicating the city

**Mort** total age-adjusted mortality from all causes

**Precip** mean annual precipitation (inches)

**Educ** median number of school years completed for persons 25 years or older

**NonWhite** percentage of population that is nonwhite

**NOX** relative pollution potential of oxides of nitrogen

**SO2** relative pollution potential of sulfur dioxide

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

McDonald, G.C. and Ayers, J.A. (1978). Some Applications of the "Chernoff Faces": A Technique for Graphically Representing Multivariate Data *in* Wang, P.C.C. (ed.) *Graphical Representation of Multivariate Data*, Academic Press.

## See Also

ex1217

## Examples

```
str(ex1123)
```

## ex1124 *Natal Dispersal Distances of Mammals*

### Description

An assessment of the factors affecting dispersal distances is important for understanding population spread, recolonization and gene flow which are central issues for conservation of many vertebrate species. Researchers gathered data on body weight, diet type and maximum natal dispersal distance for various animals.

### Usage

```
ex1124
```

### Format

A data frame with 64 observations on the following 4 variables.

**Species** a character vector indicating the species

**BodyMass** bodymass (kg)

**MaxDist** maximum dispersal distance (km)

**Type** a factor with levels `"C"`, `"H"` and `"O"` indicating carnivore, herbivore, or omnivore

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Sutherland, G.D., Harestad, A.S., Price, K. and Lertzman, K.P. (2000). Scaling of Natal Dispersal Distances in Terrestrial Birds and Mammals, *Conservation Ecology* **4**(1): 16.

### Examples

```
str(ex1124)
```

## ex1125 *Ingestion Rates of Deposit Feeders*

### Description

The data are the typical dry weight in mg, the typical ingestion rate (weight of food intake per day for one animal) in mg/day, and the percentage of the food that is composed of organic matter for 22 species of deposit feeders. The goal is to see whether the distribution of species' ingestion rates depends on the percentage of organic matter in the food, after accounting for the effect of species weight and to describe the association. The last three species happen to be Bivalves, so may behave differently from the others.

## Usage

```
ex1125
```

## Format

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

**Species** a character variable with the name of the species

**Weight** the dry weight of the species, in mg

**Ingestion** ingestion rate in mg per day

**Organic** percentage of organic matter in the food

**Bivalve** a factor with levels `"no"` and `"yes"` to indicate whether a species is a bivalve

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Cammen, L. M. (1980) Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores, *Oecologia* **44**: 303–310.

## See Also

ex0921

## Examples

```
str(ex1125)
```

---

# ex1217 *Pollution and Mortality*

---

## Description

Complete data set for problem introduced in ex1123. Data from early study designed to explore the relationship between air pollution and mortality.

## Usage

```
ex1217
```

## Format

A data frame with 60 observations on the following 17 variables.

**CITY** a character vector indicating the city

**Mortality** total age-adjusted mortality from all causes

**Precip** mean annual precipitation (inches)

**Humidity** percent relative humidity (annual average at 1:00pm)

**JanTemp** mean January temperature (degrees F)

**JulyTemp** mean July temperature (degrees F)

**Over65** percentage of the population aged 65 years or over

**House** population per household

**Educ** median number of school years completed for persons 25 years or older

**Sound** percentage of the housing that is sound with all facilities

**Density** population density (in persons per square mile of urbanized area)

**NonWhite** percentage of population that is nonwhite

**WhiteCol** percentage of employment in white collar occupations

**Poor** percentage of households with annual income under $3,000 in 1960

**HC** relative pollution potential of hydrocarbons

**NOX** relative pollution potential of oxides of nitrogen

**SO2** relative pollution potential of sulfur dioxide

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

McDonald, G.C. and Ayers, J.A. (1978). Some Applications of the "Chernoff Faces": A Technique for Graphically Representing Multivariate Data *in* Wang, P.C.C. (ed.) *Graphical Representation of Multivariate Data*, Academic Press.

## See Also

ex1123

## Examples

```
str(ex1217)
```

ex1220 *Galapagos Islands*

## Description

The number of species on an island is known to be related to the island's area. Of interest is what other variables are also related to the number of species, after island area is accounted for, and whether the answer differs for native and non native species.

## Usage

```
ex1220
```

## Format

A data frame with 30 observations on the following 8 variables.

**Island** a character vector indicating the island

**Total** total number of observed species

**Native** number of native species

**Area** area ($km^2$)

**Elev** elevation (m)

**DistNear** distance from nearest island (km)

**DistSc** distance from Santa Cruz (km)

**AreaNear** area of nearest island ($km^2$)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Johnson, M.P. and Raven, P.H. (1973). Species Number and Endemism: The Galapagos Archipelago Revisited, *Science* **179**(4076): 893–895.

## Examples

```
str(ex1220)
```

ex1221 *Predicting Desert Wildflower Blooms*

## Description

These data are monthly rainfalls from September to March and the subjectively rated quality of the following spring wildflower display for each of a number of years at each of four desert locations in the southwestern United States (Upland Sonoran Desert near Tucson, the lower Colorado River Valley section of the Sonoran Desert, the Baja California region of the Sonoran Desert, and the Mojave Desert). The quality of the display was judged subjectively with ordered rating categories of poor, fair, good, great, and spectacular. The variable Score is numerical variable corresponding to these ordered categories. A goal is to find an equation for predicting quality of wildflower blooms from the rainfall variables.

## Usage

```
ex1221
```

## Format

A data frame with 122 observations on the following 12 variables.

**Year** year of observed wildflower season

**Region** a factor variable with 4 levels: `"baja"`, `"colorado"`, `"mojave"`, and `"upland"`

**Sep** the September rainfall, in inches

**Oct** the October rainfall, in inches

**Nov** the November rainfall, in inches

**Dec** the December rainfall, in inches

**Jan** the January rainfall, in inches

**Feb** the February rainfall, in inches

**Mar** the March rainfall, in inches

**Total** the total rainfall from September through March, in inches

**Rating** a factor with a subjective assessment of the quality of wildflower bloom with levels `"FAIR"`, `"GOOD"`, `"GREAT"`, `"POOR"`, and `"SPECTACULAR"`

**Score** a numerical variable corresponding to the order of rating categories, with Poor=0, Fair=1, Good=2, Great=3, and Spectacular=4

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Arizona-Sonora Desert Museum, "Wildflower Flourishes and Flops: a 50–Year History," www.desertmuseum.org/programs/flw_wildflwrbloom.html (July 25, 2011).

## Examples

```
str(ex1221)
```

ex1222 *Bush Gore Ballot Controversy*

## Description

This data set contains the vote counts by county in Florida for Buchanan and for four other presidential candidates in 2000, along with the total vote counts in 2000, the presidential vote counts for three presidential candidates in 1996, the vote count for Buchanan in his only other campaign in Florida—the 1996 Republican primary, the registration in Buchanan's Reform party and the total political party registration in the county.

## Usage

```
ex1222
```

## Format

A data frame with 67 observations on the following 13 variables.

**County** a character vector indicating the county

**Buchanan2000** votes cast for Buchanan in 2000 presidential election

**Gore2000** votes cast for Gore in 2000 presidential election

**Bush2000** votes cast for Bush in 2000 presidential election

**Nader2000** votes cast for Nader in 2000 presidential election

**Browne2000** votes cast for Browne in 2000 presidential election

**Total2000** total vostes cast in 2000 presidential election

**Clinton96** votes cast for Clinton in 1996 presidential election

**Dole96** votes cast for Dole in 1996 presidential election

**Perot96** votes cast for Perot in 1996 presidential election

**Buchanan96p** votes cast for Buchanan in 1996 Republican primary

**ReformReg** the registration in Buchanan's Reform party

**TotalReg** the total political party registration

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex0825

## Examples

```
str(ex1222)
```

ex1223 *IQ Score and Income*

## Description

This is a subset of 2,584 individuals from the 1979 National Longitudinal Study of Youth (NLSY79) survey who were re-interviewed in 2006, who had paying jobs in 2005, and who had complete values for the variables listed below. A goal is to see whether intelligence (as measured by the ASVAB intelligence test score, AFQT, and its Components, Word, Parag, Math, and Arith) is a better predictor of 2005 income than education and socioeconomic status.

## Usage

```
ex1223
```

## Format

A data frame with 2,584 observations on the following 32 variables.

**Subject** the subject identification number

**Imagazine** a variable taking on the value 1 if anyone in the respondent's household regularly read magazines in 1979, otherwise 0

**Inewspaper** a variable taking on the value 1 if anyone in the respondent's household regularly read newspapers in 1979, otherwise 0

**Ilibrary** a variable taking on the value 1 if anyone in the respondent's household had a library card in 1979, otherwise 0

**MotherEd** mother's years of education

**FatherEd** father's years of education

**FamilyIncome78** family's total net income in 1978

**Race** 1 = Hispanic, 2 = Black, 3 = Not Hispanic or Black

**Gender** a factor with levels `"female"` and `"male"`

**Educ** years of education completed by 2006

**Science** score on the General Science test in 1981

**Arith** score on the Arithmetic Reasoning test in 1981

**Word** score on the Word Knowledge Test in 1981

**Parag** score on the Paragraph Comprehension test in 1981

**Numer** score on the Numerical Operations test in 1981

**Coding** score on the Coding Speed test in 1981

**Auto** score on the Automotive and Shop test in 1981

**Math** score on the Mathematics Knowledge test in 1981

**Mechanic** score on the Electronics Information test in 1981

**Elec** score on the Paragraph Comprehension test in 1981

**AFQT** percentile score on the AFQT intelligence test in 1981

**Income2005** total annual income in 2005

**Esteem1** self reported answer to 1st self esteem question, 2005

**Esteem2** self reported answer to 2md self esteem question, 2005

**Esteem3** self reported answer to 3rd self esteem question, 2005

**Esteem4** self reported answer to 4th self esteem question, 2005

**Esteem5** self reported answer to 5th self esteem question, 2005

**Esteem6** self reported answer to 6th self esteem question, 2005

**Esteem7** self reported answer to 7th self esteem question, 2005

**Esteem8** self reported answer to 8th self esteem question, 2005

**Esteem9** self reported answer to 9th self esteem question, 2005

**Esteem10** self reported answer to 10th self esteem question, 2005

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008).

## See Also

ex0222, ex0330, ex0331, ex0524, ex0525, ex0828, ex0923, ex1033

## Examples

```
str(ex1223)
```

---

# ex1225 *Gender Differences in Wages*

---

## Description

These data are weekly earnings for 9,835 Americans surveyed in the March 2011 Current Population Survey (CPS). What evidence is there from these data that males tend to receive higher earnings than females with the same values of the other variables? By how many dollars or by what percent does the male distribution exceed the female distribution?

## Usage

```
ex1225
```

## Format

A data frame with 9,835 observations on the following 9 variables.

**Region** a factor with levels "Midwest", "Northeast", "South", and "West"

**MetropolitanStatus** a a factor with levels "Metropolitan", "Not Identified", and "Not Metropolitan"

**Age** age in years

**Sex** a factor with levels "Female" and "Male"

**MaritalStatus** a factor with levels "Married" and "NotMarried"

**EdCode** a numerical variable representing educational attainment, with higher numbers corresponding to higher educational categories

**Education** a factor with 16 levels of educational category

**JobClass** a a factor with levels "FedGov", "LocalGov", "Private",and "StateGov"

**WeeklyEarnings** weekly wages in U.S. dollars

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.gov/cps/data.htm July 25, 2011.

## Examples

```
str(ex1225)
```

---

# ex1317 *Dinosaur Extinctions—An Observational Study*

---

## Description

About 65 million years ago, the dinosaurs suffered a mass extinction virtually overnight (in geologic time). Among many clues, one that all scientists regard as crucial is a layer of iridium-rich dust that was deposited over much of the earth at that time. The theory is that an event like a volcanic eruption or meteor impact caused a massive dust cloud that blanketed the earth for years killing off animals and their food sources. Dataset has Iridium depths by type of deposit.

## Usage

```
ex1317
```

## Format

A data frame with 28 observations on the following 3 variables.

**Iridium** Iridium in samples (ppt)

**Strata** a factor with levels "Limestone" and "Shale"

**DepthCat** a factor with six levels: "1", "2", . . . , "6"

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Alvarez, W. and Asaro, F. (1990). What Caused the Mass Extinction? An Extraterrestrial Impact, *Scientific American* **263**(4): 76–84.

Courtillot, E. (1990). What Caused the Mass Extinction? A Volcanic Eruption. *Scientific American* **263**(4): 85–92.

## Examples

```
str(ex1317)
```

---

# ex1319 *Nature—Nurture*

---

## Description

A 1989 study investigated the effect of heredity and environment on intelligence. Data are the IQ scores for adopted children whose biological and adoptive parents were categorized either in the highest or the lowest socioeconomic status category.

## Usage

```
ex1319
```

## Format

A data frame with 38 observations on the following 3 variables.

**IQ** IQ scores of adopted children

**Adoptive** a factor with levels `"High"` and `"Low"`; the socioeconomic status of the adoptive parents

**Biological** a factor with levels `"High"` and `"Low"`; the socioeconomic status of the biological parents

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Capron, C. and Duyme, M. (1991). Children's IQ's and SES of Biological and Adoptive Parents in a Balanced Cross-fostering Study, *European Bulletin of Cognitive Psychology* **11**(3): 323–348.

## See Also

ex1605

## Examples

```
str(ex1319)
```

---

`ex1320` *Gender Differences in Performance on Mathematics Achievement Tests*

---

## Description

Data set on 861 ACT Assessment Mathematics Usage Test scores from 1987. The test was given to a sample of high school seniors who met one of three profiles of high school mathematics course work: (a) Algebra I only; (b) two Algebra courses and Geometry; and (c) two Algebra courses, Geometry, Trigonometry, Advanced Mathematics and Beginning Calculus.

These data were generated from summary statistics for one particular form of the test as reported by Doolittle (1989).

## Usage

```
ex1320
```

## Format

A data frame with 861 observations on the following 3 variables.

**Sex** a factor with levels `"female"` and `"male"`

**Background** a factor with levels `"a"`, `"b"` and `"c"`

**Score** ACT mathematics test score

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Doolittle, A.E. (1989). Gender Differences in Performance on Mathematics Achievement Items, *Applied Measurement in Education* **2**(2): 161–177.

## Examples

```
str(ex1320)
```

ex1321 *Pygmalion*

## Description

A data set simulated to match the summary statistics and conclusions from Rosenthal and Jacobson's Pygmalion study on elementary school students. The researchers assigned students at random to a pygmalion or control treatment group. They supplied information to the teachers of those in the pygmalion group with the false information that an intelligence test had indicated that the student was likely to excel. The researchers wished to see if the change in intelligence test scores for the students tended to be larger for those students labeled as likely to excel.

## Usage

```
ex1321
```

## Format

A data frame with 320 observations on the following 5 variables.

**Student** a student identification number

**Grade** the student's grade, 1 through 6

**Class** a factor with 17 levels `"1a"`, `"1b"`, and so on, to indicate the 17 distinct teacher/classrooms.

**Treatment** a factor with levels `"pygmalion"` and `"control"` corresponding to whether the researchers had told the teacher that the student was "likely to succeed" or not

**Gain** the intelligence test score taken at the end of the school year minus the intelligence test score taken at the begging of the school year

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rosenthal, R. and Jacobson, L. 1968, *Pygmalion in the Classroom: Teacher Expectation and Pupil's Intellectual Development*, Holt, Rinehart, and Winston, Inc.

## Examples

```
str(ex1321)
```

ex1416 *Blood Brain Barrier*

## Description

Researchers designed an experiment to investigate how delivery of brain cancer antibody is influenced by tumor size, antibody molecular weight, blood-brain barrier disruption, and delivery route.

## Usage

```
ex1416
```

## Format

A data frame with 36 observations on the following 6 variables.

**Agent** a factor with levels `"AIB"`, `"DEX7"` and `"MTX"`

**Treatment** a factor with levels `"BD"` and `"NS"`

**Route** a factor with levels `"IA"` and `"IV"`

**DaysPost** days after inoculation

**BAT** concentration of antibody in the part of the brain around the tumor

**LH** concentration of antibody in the unaffected part of the brain

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Barnett, P.A., Roman-Goldstain, S., Ramsey, F., McCormick, C.I., Sexton, G., Szumowski, J. and Neuwelt, E.A. (1995). Differential Permeability and Quantitative MR Imaging of a Human Lung Carcinoma Brain Xenograft in the Nude Rat, *American Journal of Pathology* **146**(2): 436–449.

## See Also

`case1102`, `ex1417`

## Examples

```
str(ex1416)
```

ex1417 *Second Replicate of the Barrier Disruption Study*

## Description

Researchers designed an experiment to investigate how delivery of brain cancer antibody is influenced by tumor size, antibody molecular weight, blood-brain barrier disruption, and delivery route. The data for the first replicate of this study is in ex1416. This is the second replicate for the study.

## Usage

```
ex1417
```

## Format

A data frame with 36 observations on the following 6 variables.

**Agent** a factor with levels "AIB", "DEX70" and "MTX"

**Treatment** a factor with levels "BD" and "NS"

**Route** a factor with levels "IA" and "IV"

**DaysPost** days after inoculation

**BAT** concentration of antibody in the part of the brain around the tumor

**LH** concentration of antibody in the unaffected part of the brain

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Barnett, P.A., Roman-Goldstain, S., Ramsey, F., McCormick, C.I., Sexton, G., Szumowski, J. and Neuwelt, E.A. (1995). Differential Permeability and Quantitative MR Imaging of a Human Lung Carcinoma Brain Xenograft in the Nude Rat, *American Journal of Pathology* **146**(2): 436–449.

## See Also

case1102, ex1416

## Examples

```
str(ex1417)
```

ex1419 *Clever Hans Effect*

## Description

These data were simulated to match the summary statistics and conclusions of Rosenthal and Fode's Clever Hans experiment. Each of 12 students trained rats to run a maze. The data set contains their number of successful runs out of 50 on each of 5 days. It also shows two summarizing statistics for each student: the overall success rate on all 5 days and the slope in the least squares regression of daily success rate (number of successes in a day divided by 50) on day. Also included are the student's response to the prior expectation of success question and the student's response to a post- experiment question about how relaxed they felt handling their rats (with higher values corresponding to more relaxed). The treatment variable shows whether or not the students were supplied with the fictitious information about whether their rats were bright or not.

## Usage

```
ex1419
```

## Format

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

**Student** a student identification number

**PriorExp** the student's prior expectation of rat-training success, on a scale from -10 to 10

**Block** a numerical variable for pairs of students grouped according to their values of PriorExp

**Treatment** a factor with levels `"bright"` and `"dull"` corresponding to whether students were told (falsely) that their rats were bright or not

**Day1** the number of successful rat mazed runs on day 1

**Day2** the number of successful rat mazed runs on day 2

**Day3** the number of successful rat mazed runs on day 3

**Day4** the number of successful rat mazed runs on day 4

**Day5** the number of successful rat mazed runs on day 5

**Relax** degree of relaxation students felt in handling their rats, on a scale from 0 to 10

**Success** the total proportion of successful maze runs in 5 days

**Slope** the slope in the least squares regression of mean daily success as a function of day, estimated for each student individually

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rosenthal, R. and Fode, K.L. (1963) The Effect of Experimenter Bias on the Performance of the Albino Rat *Behavioral Science* **8:3**: 183–189.

## See Also

ex2120

## Examples

```
str(ex1419)
```

---

ex1420 *Diet Wars*

---

## Description

These data are the weight losses of subjects randomly assigned to one of three diets, and these additional covariates sex, initial age, and body mass index. Is there any evidence from these data that the mean weight loss differs for the different diets, after accounting for the effect of the covariates? How big are the difference?

## Usage

```
ex1420
```

## Format

A data frame with 272 observations on the following 6 variables.

**Subject** a subject identification number

**Diet** a factor with levels "Low-Carbohydrate", "Low-Fat"and "Mediterranean"

**Sex** a factor with levels "F" and "M"

**Age** subject's age in years

**BMI** body mass index in kg/squared meter

**WtLoss24** weight at the end of the 24 month study minus initial weight, in kg

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex0623, ex1921, ex1922

## Examples

```
str(ex1420)
```

## ex1507 *Global Warming, Southern Hemisphere*

### Description

The data are the temperatures (in degrees Celsius) averaged for the southern hemisphere over a full year, for years 1850 to 2010. The 161-year average temperature has been subtracted, so each observation is the temperature difference from the series average.

### Usage

```
ex1507
```

### Format

A data frame with 161 observations on the following 2 variables.

**Year** year in which yearly average temperature was computed, from 1850 to 2010

**Temperature** southern hemisphere temperature minus the 161-year average (degrees Celsius)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Jones, P.D., D. E. Parker, T. J. Osborn, and K. R. Briffa, (2011) Global and Hemispheric Temperature Anomalies and and Marine Instrumental Records, CDIAC, http://cdiac.ornl.gov/trends/temp/jonescru/jones.html, Aug 4, 2011

### Examples

```
str(ex1507)
```

## ex1509 *Sunspots*

### Description

The data are the annual sunspot counts in each year from 1700 to 2010.

### Usage

```
ex1507
```

### Format

A data frame with 311 observations on the following 2 variables.

**Year** year in which sunspots were counted, from 1700 to 2010

**Sunspots** the number of sunspots observed in a year

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

SIDC-Solar Influences Dta Center, http://sidc.oma.be/sunspot-data/ (July 15, 2011).

## Examples

```
str(ex1509)
```

---

ex1514 *Melanoma and Sunspot Activity—An Observational Study*

---

## Description

Several factors suggest that the incidence of melanoma is related to solar radiation. These data are the age-adjusted melanoma incidence among males in the Connecticut Tumor Registry and the sunspot activity, 1936–1972 .

## Usage

```
ex1514
```

## Format

A data frame with 37 observations on the following 3 variables.

**Year** year

**Melanoma** male melanoma incidence in number of cases per 100,000 population

**Sunspot** sunspot relative number

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Houghton, A., Munster, E.W. and Viola, M.V. (1978). Increased Incidence of Malignant Melanoma After Peaks of Sunspot Activity, *Lancet*: 759–760.

## Examples

```
str(ex1514)
```

## ex1515 *Lynx Trappings and Sunspots*

### Description

Data on the annual numbers of lynx trapped in the Mackenzie River district of northwest Canada from 1821–1934.

### Usage

```
ex1515
```

### Format

A data frame with 114 observations on the following 3 variables.

**Year** year

**Lynx** number of lynx trapped

**Sunspots** number of sunspots

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Elston, C. and Nicholson, M. (1942). The Ten Year Cycle in Numbers of the Lynx in Canada, *Journal of Animal Ecology* **11**(2): 215–244.

### Examples

```
str(ex1515)
```

## ex1516 *Trends in Firearm and Motor Vehicle Deaths in the U.S.*

### Description

Data shows the number of deaths due to firearms and the number of deaths due to motor vehicle accidents in the United States between 1968 and 1993.

### Usage

```
ex1516
```

## Format

A data frame with 26 observations on the following 3 variables.

**Year** year

**FirearmDeaths** deaths due to firearms (in thousands per year)

**MotorVehicleDeaths** deaths due to motor vehicles (in thousands per year)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Data read from a Centers for Disease Control and Prevention graph reported in *The Oregonian*, June 17, 1997.

## Examples

```
str(ex1516)
```

---

# ex1517 *S&P 500*

---

## Description

The Standard and Poor's 500 stock index (S&P 500) is a benchmark of stock market performance, based on 400 industrial firms, 40 financial stocks, 40 utilities, and 20 transportation stocks. These data include the value of a $1 investment in 1871 at the end of each year from 1871 to 1999, according to the S&P 500, assuming all dividends are reinvested. Describe the distribution of the S&P value as a function of year.

## Usage

```
ex1517
```

## Format

A data frame with 129 observations on the following 2 variables.

**Year** year

**S.P500Return** Value of Stock at the end of the year

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex1517)
```

## ex1518 *Effectiveness of Measles Vaccine*

### Description

The data are the number of measles cases reported in the United States for each year from 1950 to 2008. A goal is to explore the effect of the introduction of the measles vaccine in 1963 on the series mean.

### Usage

```
ex1518
```

### Format

A data frame with 59 observations on the following 3 variables.

**Year** year

**Cases** number of measles cases

**Vaccine** a factor with levels `"no"` and `"yes"` indicating whether the measles vaccine had been licensed or not (yes for every year starting with 1963)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Center for Disease Control,
http://www.cdc.gov/vaccines/pubs/pinkbook/downloads/appendices/G/cases-deaths.pdf
retrieved on July 23, 2009

### Examples

```
str(ex1518)
```

## ex1519 *El Nino and the Southern Oscillation*

### Description

The data are the Sea Surface Temperatures (SST) and Southern Oscillation Index (SOI) measurements from 1950 to 2010.

### Usage

```
ex1519
```

## Format

A data frame with 732 observations on the following 4 variables.

**Year** year

**Month** a numerical variable for month, with 1 = January

**SOI** the Southern Oscillation Index)

**SST** the Sea Surfact Temperature)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

U.S. National Oceanographic and Atmospheric Administration (http://www.cpc.ncep.noaa.gov/data/indices/).

## See Also

case1502

## Examples

```
str(ex1519)
```

---

# ex1605 *Nature—Nurture*

---

## Description

Data are a subset from an observational, longitudinal, study on adopted children. Is child's intelligence related to intelligence of the biological mother and the intelligence of the adoptive mother?

## Usage

```
ex1605
```

## Format

A data frame with 62 observations on the following 6 variables.

**FMED** adoptive (foster) mother's years of education

**TMIQ** biological mother's score on IQ test

**Age2IQ** IQ of child at age 2

**Age4IQ** IQ of child at age 4

**Age8IQ** IQ of child at age 8

**Age13IQ** IQ of child at age 13

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Skodak, M. and Skeels, H.M. (1949). A Final Follow-up Study of One Hundred Adopted Children, *Journal of Genetic Psychology* **75**: 85–125.

## See Also

ex1319

## Examples

```
str(ex1605)
```

---

# ex1611 *Religious Competition*

---

## Description

Adam Smith, in Wealth of Nations, observed that even religious monopolies become weak when they are not challenged by competition. Data to illustrate this point is from 21 countries in which the percentages of Catholics in the populations varied from a low 1.2% to a high 97.6%.

## Usage

```
ex1611
```

## Format

A data frame with 21 observations on the following 4 variables.

**Country** a character vector indicating the country

**PctCatholic** percent Catholics in the population

**PriestParishRatio** priest to parishioner ratio

**PctIndigenous** percent clergy indigenous

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Gill, A.J. (1994). Rendering unto Caesar? Religious Competition and Catholic Political Strategy in Latin America, 1962–79, *American Journal of Political Science* **38**(2): 403–425.

## Examples

```
str(ex1611)
```

## ex1612 *Wastewater*

### Description

Samples of effluent were divided and sent to two laboratories for testing. Data are measurements of biochemical oxygen demand and suspended solid measurements obtained for 2 sample splits from the two laboratories.

### Usage

```
ex1612
```

### Format

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

**ComBOD** biochemical oxygen demand measurements from commercial laboratory

**ComSS** suspended solids measurements from commercial laboratory

**StaBOD** biochemical oxygen demand measurements from state laboratory

**StaSS** suspended solids measurements from state laboratory

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Johnson, R.A. and Wichern, D.W. (1988). *Applied Multivariate Statistical Analysis*, Prentice-Hall.

### Examples

```
str(ex1612)
```

## ex1613 *Flea Beetle Distinction*

### Description

Data are the measurements from two very similar species of flea beetle.

### Usage

```
ex1613
```

## Format

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

**Specimen** specimen identification number

**Jnt1** measurement of first joint in micrometers

**Jnt2** measurement of second joint in micrometers

**Species** a factor with levels `"conc"` and `"heik"`

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Lubischew, A.A. (1962). On the Use of Discriminant Functions in Taxonomy, *Biometrics* **18**: 455–477.

## Examples

```
str(ex1613)
```

---

# ex1614 *Pschoimmunology*

---

## Description

Recent studies in the field of psychoimmunology suggest a link exists between behavioral events and the functioning of one's immune system. Data shows the results of a study on 12 subjects who were monitored during three distinct activities. The first activity consisted of neutral activity such as reporting tasks. During the second activity, subjects listened to audiotape exercises relating to images of heaviness, warmth in the body, relaxation, suggestions to remember happy events, etc. The third activity included a nonaudio tape follow up stimulus consisting of continued relaxation as in activity 2 and a verbal discussion of the positive aspects of the audiotape.

## Usage

```
ex1614
```

## Format

A data frame with 12 observations on the following 3 variables.

**Subject** subject identification number

**PhaseA** Interleukin-1 levels (counts per minute) from blood samples taken during activity A

**PhaseB** Interleukin-1 levels (counts per minute) from blood samples taken during activity B

**PhaseC** Interleukin-1 levels (counts per minute) from blood samples taken during activity C

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Keppel, W. (1993). Effects of Behavioral Stimuli on Plasma Interleukin-1 Activity in Humans at Rest, *Journal of Clinical Psychology* **49**(6): 777–785.

## Examples

```
str(ex1614)
```

---

# ex1615 *Trends in SAT Scores*

---

## Description

Data shows a partial listing of a data set with ratios of average math to average verbal SAT scores in the United States and the District of Columbia for 1989 and 1996–1999.

## Usage

```
ex1615
```

## Format

A data frame with 51 observations on the following 6 variables.

**State** a character vector indicating the state

**M.V.89** average MATH SAT scores divided by average VERBAL SAT score in 1989

**M.V.96** average MATH SAT scores divided by average VERBAL SAT score in 1996

**M.V.97** average MATH SAT scores divided by average VERBAL SAT score in 1997

**M.V.98** average MATH SAT scores divided by average VERBAL SAT score in 1998

**M.V.99** average MATH SAT scores divided by average VERBAL SAT score in 1999

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex1615)
```

## ex1620 *Differential Gene Expression with RNA Sequencing*

### Description

In an experiment to identify genes of the plant Arabidopsis that react to a particular pathogen, researchers used RNA sequencing to produce gene profiles for a number of plants not subjected to the pathogen and several plants subjected to the pathogen. Tests comparing the distribution of gene expression in the two groups were performed for each gene individually. The data are the p-values from all these tests. The goal is to use a identify a set of genes that differentially express in the two groups, subject to some specified value for expected false discovery rate, such as 5%.

### Usage

```
ex1620
```

### Format

A data frame with 20,245 observations on the following 3 variables.

**Gene** an identification number for genes

**GeneName** a character variable with the name of the gene

**pValue** the p-value from a test that the mean expression level for the gene differs in the two groups (ignoring multiple testing)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Chang, J, Department of Botany and Plant Pathology, Oregon State University, personnal communication.

### Examples

```
str(ex1620)
```

## ex1708 *Pig Fat*

### Description

Actual pig fat and measurements of pig fat from magnetic resonance images at 13 locations for 12 pigs.

### Usage

```
ex1708
```

## Format

A data frame with 12 observations on the following 14 variables.

**Fat** actual pig fat (in percent)

**M1** magnetic resonance image at location 1

**M2** magnetic resonance image at location 2

**M3** magnetic resonance image at location 3

**M4** magnetic resonance image at location 4

**M5** magnetic resonance image at location 5

**M6** magnetic resonance image at location 6

**M7** magnetic resonance image at location 7

**M8** magnetic resonance image at location 8

**M9** magnetic resonance image at location 9

**M10** magnetic resonance image at location 10

**M11** magnetic resonance image at location 11

**M12** magnetic resonance image at location 12

**M13** magnetic resonance image at location 13

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Glasbey, C.A and Flowler, P.A. (1992). Regression Models Fitted Using Conditional Independence to Estimate Pig Fatness from Magnetic Resonance Images, *The Statistician* **41**(2): 179–184.

## Examples

```
str(ex1708)
```

---

# ex1715 *Church Distinctiveness*

---

## Description

Data show measures that differ among denominations of American Protestant and Catholic churches.

## Usage

```
ex1715
```

### Format

A data frame with 18 observations on the following 6 variables.

**Denomination** a character vector indicating the church denomination

**Distinct** distinctiveness (strictness of discipline on a seven point scale)

**Attend** average percentage of weeks that individuals attended a church meeting (% weekly)

**NonChurch** average number of secular organisations to which members belong

**PctStrong** average percentage of members that describe themselves as being strong church members (%)

**AnnInc** average income of members (US$)

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Iannaccone, L.R. (1994). Why Strict Churches Are Strong, *American Journal of Sociology* **99**(5): 1180–1211.

### Examples

```
str(ex1715)
```

---

## ex1716 *Insurance*

---

### Description

In the 1970's the U.S. Commission on Civil Rights investigated charges that insurance companies were attempting to redefine Chicago "neighborhoods" in order to cancel existing homeowner insurance policies or refuse to issue new ones. Dataset has data on homeowner and residential fire insurance policy issuances from 47 zip codes in the Chicago area.

### Usage

```
ex1716
```

### Format

A data frame with 47 observations on the following 8 variables.

**ZIP** last 2 digits of zip code

**Fire** fires per 1000 housing units

**Theft** thefts per 1000 population

**Age** percentage of housing units built prior to 1940

**Income** median family income

**Race** percentage minority

**Vol** number of new policies per 100 housing units

**Invol** number of FAIR plan policies and renewals per 100 housing units

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from many Fields for the Student and Research Worker*, Springer-Verlag.

## Examples

```
str(ex1716)
```

---

`ex1914` *Mantel-Haenszel Test for Censored survival Times: Lymphoma and Radiation Data*

---

## Description

Survival times for two groups of lymphoma patients.

## Usage

```
ex1914
```

## Format

A data frame with 34 observations on the following 4 variables.

**Months** months after diagnosis

**Group** a factor with levels `"no"` and `"radiation"`

**Survived** number of patients known to survive beyond this month

**Died** number of patients known to die after this many months

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Neuwelt, E.A., Goldman, D.L., Dahlborg, S.A., Crossen, J., Ramsey, F., Roman-Goldstein, S., Braziel, R. and Dana, B. (1991). Primary CNS Lymphoma Treated with Osmotic Blood-brain Barrier Disruption: Prolonged Survival and Preservation of Cognitive Function, *Journal of Clinical Oncology* **9**(9): 1580–1590.

## Examples

```
str(ex1914)
```

ex1916 *Vitamin C and Colds*

## Description

Fictitious data set based on results of an experiment where subjects were randomly divided into two groups and given a placebo or vitamin c to take during the cold season. At the end of the cold season, the subjects were interviewed by a physician who determined whether they had or had not suffered a cold during the period. Skeptics interviewed the 800 subjects to determine who knew and who did not know to which group they had been assigned. Vitamin C has a bitter taste and those familiar with it could recognize whether their pills contained it.

## Usage

```
ex1916
```

## Format

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

**Knew** a factor with levels `"no"` and `"yes"`

**Treatment** a factor with levels `"placebo"` and `"vitC"`

**Cold** number of people who got a cold

**NoCold** number of people who did not get a cold

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
str(ex1916)
```

ex1917 *Alcohol Consumption and Breast Cancer—A Retrospective Study*

## Description

Dataset from a study which investigated the added risk of breast cancer due to alcohol consumption. A sample of confirmed breast cancer patients were compared with a sample of cancer free women who were close in age and from the same neighborhood as the cases. Data was collected on the alcohol consumption and body mass of both sets of women.

## Usage

```
ex1917
```

## Format

A data frame with 6 observations on the following 4 variables.

**BodyMass** a factor with levels `"high"`, `"low"` and `"medium"`

**Drinking** a factor with levels `"high"` and `"low"`

**Cases** number of women with breast cancer

**Controls** number of women without breast cancer

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rosenberg, L., Palmer, J.R., Miller, D.R., Clarke, E.A. and Shapiro, S. (1990). A Case-Control Study of Alcoholic Beverage Consumption and Breast Cancer, *American Journal of Epidemiology* **131**(1): 6–14.

## Examples

```
str(ex1917)
```

---

# ex1918 *The Donner Party*

---

## Description

In 1846 the Donner party became stranded while crossing the Sierra Nevada Mountains near Lake Tahoe. The data frame has the counts for male and female survivors for six age groups.

## Usage

```
ex1918
```

## Format

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

**AgeCat** a numerical code corresponding to six age categories, with 1 = `"15-19"`, 2 = `"20-29"`, 3 = `"30-39"`, 4 = `"40-49"`, 5 = `"50-59"` and 6 = `"60-69"`

**Sex** a factor with levels `"female"` and `"male"`

**Lived** number that lived

**Died** number that died

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Grayson, D.K. (1990). Donner Party Deaths: A Demographic Assessment, *Journal of Anthropological Research* **46**: 223–242.

## See Also

case2001

## Examples

```
str(ex1918)
```

---

# ex1919 *Tire-Related Fatal Accidents and Ford Sports Utility Vehicles*

---

## Description

Data shows the numbers of compact sports utility vehicles involved in fatal accidents in the U.S. between 1995 and 1999, categorized according to travel speed, make of car (Ford or other), and cause of accident (tire-related or other).

## Usage

```
ex1919
```

## Format

A data frame with 8 observations on the following 4 variables.

**SpeedCat** a numerical code corresponding to 4 categories of speed (in miles per hour), with 1 = "0-40", 2 = "41-55", 3 = "56-65" and 4 = ">65"

**Make** a factor with levels "Ford" and "Other"

**Other** cause of accident was other than tire-related

**Tire** cause of accident was tire-related

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex2018

## Examples

```
str(ex1919)
```

## ex1921 *Diet Wars II*

### Description

In the study of different diets for losing weight (ex0623, ex1420 and ex1922), there appear to have been many more experimental subjects that dropped out from the low carbohydrate diet group than from the other two diet groups. This data set contains the numbers who did and didn't drop out in each diet group. Is there any evidence that the drop out rate differs in the three groups?

### Usage

```
ex1921
```

### Format

A data frame with 3 observations on the following 4 variables.

**Diet** a factor with levels `"LowCarb"`, `"LowFat"`, and `"Medit"`

**DroppedOut** the number of subjects who dropped out of the study

**Completed** the number of subjects who completed the study

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### See Also

ex0623, ex1420, ex1922

### Examples

```
str(ex1921)
```

## ex1922 *Diet Wars III*

### Description

For the study of different diets for losing weight (ex0623, ex1420 and ex1921), it is desired to see whether women were more or less likely to drop out of the study than men (after accounting for the apparent differential drop out rates associated with diet). This data set includes the numbers that dropped out and completed the study for each combination of Sex and Diet.

### Usage

```
ex1922
```

## Format

A data frame with 6 observations on the following 4 variables.

**Diet** a factor with levels `"LowCarb"`, `"LowFat"`, and `"Medit"`

**Gender** a factor with levels `"Men"` and `"Women"`

**DroppedOut** the number of subjects who dropped out of the study

**Completed** the number of subjects who completed the study

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

`ex0623`, `ex1420`, `ex1921`

## Examples

```
str(ex1922)
```

---

`ex1923` *Who Looks After the Kids?*

---

## Description

One issue concerning the validity of the clutch volume and parental care study of `ex1031` is the selection of the bird species in the set of currently living animals. Was the selection just as good as a random sample of species from each of the groups? One way to study this for birds, at least, is to compare the numbers of species from each of the 29 orders of birds in the study with the known total number of species in each of the orders. If the selection of birds had been at random, the expected proportion of species in the study from one particular order, n, is the proportion of all species in that order (N=9,866) times the total number of species in the sample (414). That is, the expected number in each sample, if random sampling were used, is (N/9,866)x 414. Calculate the expected numbers and compare the observed numbers with them using Pearson's chi-square statistic.

## Usage

```
ex1923
```

## Format

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

**Order** a character variable with the name of the order

**N** the known number of species in the order

**n** the number of sampled species from the order

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex1031

## Examples

```
str(ex1923)
```

---

ex2011 *Space Shuttle*

---

## Description

This data frame contains the launch temperatures (degrees Fahrenheit) and an indicator of O-ring failures for 24 space shuttle launches prior to the space shuttle *Challenger* disaster of January 28, 1986.

## Usage

```
ex2011
```

## Format

A data frame with 24 observations on the following 2 variables.

**Temperature** Launch temperature (in degrees Fahrenheit)

**Failure** Indicator of O-ring failure

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

case0401, ex2223

## Examples

```
str(ex2011)
```

ex2012 *Muscular Dystrophy*

## Description

Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Boys with the disease usually die at a young age; but affected girls usually do not suffer symptoms, may unknowingly carry the disease and may pass it to their offspring. It is believed that about 1 in 3,300 women are DMD carriers. A woman might suspect she is a carrier when a related male child develops the disease. Doctors must rely on some kind of test to detect the presence of the disease. This data frame contains data on two enzymes in the blood, creatine kinase (CK) and hemopexin (H) for 38 known DMD carriers and 82 women who are not carriers. It is desired to use these data to obtain an equation for indicating whether a women is a likely carrier.

## Usage

```
ex2012
```

## Format

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

**Group** Indicator whether the woman has DMD (`"Case"`) or not (`"Control"`)

**CK** Creatine kinase reading

**H** Hemopexin reading

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems From Many Fields For The Student And Research Worker*, Springer-Verlag, New York.

## Examples

```
str(ex2012)
```

`ex2015` *Spotted Owl Habitat*

## Description

A study examined the association between nesting locations of the Northern Spotted Owl and availability of mature forests. Wildlife biologists identified 30 nest sites. The researchers selected 30 other sites at random coordinates in the same forest. On the basis of aerial photographs, the percentage of mature forest (older than 80 years) was measured in various rings around each of the 60 sites.

## Usage

```
ex2015
```

## Format

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

**Site** Site, a factor with levels `"Random"` and `"Nest"`

**PctRing1** Percentage of mature forest in ring with outer radius 0.91 km

**PctRing2** Percentage of mature forest in ring with outer radius 1.18 km

**PctRing3** Percentage of mature forest in ring with outer radius 1.40 km

**PctRing4** Percentage of mature forest in ring with outer radius 1.60 km

**PctRing5** Percentage of mature forest in ring with outer radius 1.77 km

**PctRing6** Percentage of mature forest in ring with outer radius 2.41 km

**PctRing7** Percentage of mature forest in ring with outer radius 3.38 km

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Ripple W.J., Johnson, D.H., Thershey, K.T. and Meslow E.C. (1991). Old–growth and Mature Forests Near Spotted Owl Nests in Western Oregon, *Journal of Wildlife Management* **55**(2): 316–318.

## Examples

```
str(ex2015)
```

ex2016 *Bumpus Natural Selection Data*

## Description

Hermon Bumpus analysed various characteristics of some house sparrows that were found on the ground after a severe winter storm in 1898. Some of the sparrows survived and some perished. This data set contains the survival status, age, the length from tip of beak to tip of tail (in mm), the alar extent (length from tip to tip of the extended wings, in mm), the weight in grams, the length of the head in mm, the length of the humerus (arm bone, in inches), the length of the femur (thigh bones, in inches), the length of the tibio–tarsus (leg bone, in inches), the breadth of the skull in inches and the length of the sternum in inches.

## Usage

```
ex2016
```

## Format

A data frame with 87 observations on the following 11 variables.

**Status** Survival status, factor with levels `"Perished"` and `"Survived"`

**AG** a numerical code corresponding to two categories of age, with 1 = `"adult"` and 2 = `"juvenile"`

**TL** total length (in mm)

**AE** alar extent (in mm)

**WT** weight (in grams)

**BH** length of beak and head (in mm)

**HL** length of humerus (in inches)

**FL** length of femur (in inches)

**TT** length of tibio–tarsus (in inches)

**SK** width of skull (in inches)

**KL** length of keel of sternum (in inches)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex0221

## Examples

```
str(ex2016)
```

ex2017 *Catholic stance*

## Description

The Catholic church has explicitly opposed authoritarian rule in some (but not all) Latin American countries. Although such action could be explained as a desire to counter repression or to increase the quality of life of its parishioners, A.J. Gill supplies evidence that the underlying reason may be competition from evangelical Protestant denominations. He compiled the data given in this data frame.

## Usage

```
ex2017
```

## Format

A data frame with 12 observations on the following 5 variables.

**Stance** Catholic church stance, factor with levels `"Pro"` and `"Anti"`

**Country** Latin American country

**PQLI** Physical Quality of Life Index in the mid-1970s; Average of live expectancy at age 1, infant mortality and literacy at age 15+.

**Repression** Average civil rights score for the period of authoritarian rule until 1979

**Competition** Percentage increase of competitive religious groups during the period 1900–1970

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Gill, A.J. (1994). Rendering unto Caesar? Religious Competition and Catholic Strategy in Latin America, 1962–1979, *American Journal of Political Science* **38**(2): 403–425.

## Examples

```
str(ex2017)
```

ex2018 *Fatal Car Accidents Involving Tire Failures on Ford Explorers*

## Description

This data frame contains data on 1995 and later model compact sports utility vehicles involved in fatal accidents in the United States between 1995 and 1999, excluding those that were struck by another car and excluding accidents that, according to police reports, involved alcohol.

## Usage

```
ex2018
```

## Format

A data frame with 2,321 observations on the following 4 variables.

**Make** Type of sports utility vehicle, factor with levels `"Other"` and `"Ford"`

**VehicleAge** Vehicle age (in years); surrogate for age of tires

**Passengers** Number of passengers

**Cause** Cause of fatal accident, factor with levels `"NotTire"` and `"Tire"`

## Details

The Ford Explorer is a popular sports utility vehicle made in the United States and sold throughout the world. Early in its production concern arose over a potential accident risk associated with tires of the prescribed size when the vehicle was carrying heavy loads, but the risk was thought to be acceptable if a low tire pressure was recommended. The problem was apparently exacerbated by a particular type of Firestone tire that was overly prone to separation, especially in warm temperatures. This type of tire was a common one used on Explorers in model years 1995 and later. By the end of 1999 more than 30 lawsuits had been filed over accidents that were thought to be associated with this problem. U.S. federal data on fatal car accidents were analysed at that time, showing that the odds of a fatal accident being associated with tire failure were three times as great for Explorers as for other sports utility vehicles.

Additional data from 1999 and additional variables may be used to further explore the odds ratio. It is of interest to see whether the odds that a fatal accident is tire-related depend on whether the vehicle is a Ford, after accounting for age of the car and number of passengers. Since the Ford tire problem may be due to the load carried, there is some interest in seeing whether the odds associated with a Ford depend on the number of passengers.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

ex1919

## Examples

```
str(ex2018)
```

## ex2019 *Missile Defenses*

### Description

Following a successful test of an anti-ballistic missile (ABM) in July 1999, many prominent U.S. politicians called for the early deployment of a full ABM system. The scientific community was less enthusiastic about the efficacy of such a system. This data set contains the success or failure of ABM tests between March 1983 and May 1995. Do these data suggest any improvement in ABM test success probability over time?

### Usage

```
ex2019
```

### Format

A data frame with 17 observations on the following 3 variables.

**Date** date of an ABM test

**Months** number of months after March 1983

**Result** a factor with levels `"Failure"` and `"Success"`

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Lewis, G. N., Postol, T. A. and Pike, J. (1999) Why National Missile Defense Won't Work, *Scientific American* **281(2)**: 36–41.

### Examples

```
str(ex2019)
```

## ex2113 *Vitamin C and Colds*

### Description

These data are from a randomized experiment to asses the effect of large doses of vitamin C on the incidence of colds.

### Usage

```
ex2113
```

## Format

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

**Dose** the daily dose of vitamin C, in g

**Number** the number of subjects given that dose of vitamin C

**WithoutIllness** the number of subjects who did not become ill

**ProportionWithout** the proportion of subjects who did not become ill

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Anderson, T.W., Suranyi,G., and Beaton, G.H. (1974) The Effect on Winter Illness of Large Doses of Vitamin C *Canadian Medical Association Journal* **111** 31–36.

## Examples

```
str(ex2113)
```

---

# ex2115 *Belief Accessibility*

---

## Description

The study the effect of *context questions* prior to *target questions*, researchers conducted a poll involving 1,054 subjects selected randomly from the Chicago phone directory. To include possibly unlisted phones, selected numbers were randomly altered in the last position. This data frame contains the responses to one of the questions asked concerning continuing U.S. aid to the Nicaraguan Contra rebels. Eight different versions of the interview were given, representing all possible combinations of three factors at each of two levels. The experimental factors were Context, Mode and Level.

`Context` refers to the type of context questions preceding the question about Nicaraguan aid. Some subjects received a context question about Vietnam, designed to elicit reticence about having the U.S. become involved in another foreign war in a third–world country. The other context question was about Cuba, designed to elicit anti–communist sentiments.

`Mode` refers to whether the target question immediately followed the context question or whether there were other questions scattered in between.

`Level` refers to two versions of the context question. In the `"high"` level the question was worded to elicit a higher level of agreement than in the `"low"` level wording.

## Usage

```
ex2115
```

## Format

A data frame with 8 observations on the following 7 variables.

**Context** Factor referring to the context of the question preceding the target question about U.S. aid to the Nicaraguan Contra rebels

**Mode** Factor with levels `"not"` and `"scattered"`, `"scattered"` is used if the target question was not asked directly after the context question

**Level** Factor with levels `"low"` and `"high"`, refers to the wording of the question

**Number** Number of people interviewed

**InFavor** Number of people in favor of Contra Aid

**NotInFavor** Number of people not in favor of Contra Aid

**PercentInFavor** Percentage in favour of Contra aid

## Details

Increasingly, politicians look to public opinion surveys to shape their public stances. Does this represent the ultimate in democracy? Or are seemingly scientific polls being rigged by the manner of questioning? Psychologists believe that opinions—expressed as answers to questions—are usually generated at the time the question is asked. Answers are based on a quick sampling of relevant beliefs held by the subject, rather than a systematic canvas of all such beliefs. Furthermore, this sampling of beliefs tends to overrepresent whatever beliefs happen to be most accessible at the time the question is asked. This aspect of delivering opinions can be abused by the pollster. Here, for example, is one sequence of questions:

**(1)** "Do you believe the Bill of Rights protects personal freedom?"

**(2)** "Are you in favor of a ban on handguns?"

Here is another:

**(1)** "Do you think something should be done to reduce violent crime?"

**(2)** "Are you in favor of a ban on handguns?"

The proportion of yes answers to question 2 may be quite different depending on which question 1 is asked first.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Tourangeau, R., Rasinski, K.A., Bradburn, N. and D'Andrade, R. (1989). Belief Accessibility and Context Effects in Attitude Measurement, *Journal of Experimental Social Psychology* **25**: 401–421.

## Examples

```
str(ex2115)
```

## ex2116 *Aflatoxicol and Liver Tumors in Trout*

### Description

An experiment at the Marine/Freshwater Biomedical Sciences Center at Oregon State University investigated the carcinogenic effects of aflatoxicol, a metabolite of Aflatoxin B1, which is a toxic by-product produced by a mold that infects cottonseed meal, peanuts and grains. Twenty tanks of rainbow trout embryos were exposed to one of five doses of Aflatoxicol for one hour. The data represent the numbers of fish in each tank and the numbers of these that had liver tumours after one year.

### Usage

```
ex2116
```

### Format

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

**Dose** Dose (in ppm)

**Tumor** Number of trout with liver tumours

**Total** Number of trout in tank

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### Examples

```
str(ex2116)
```

## ex2117 *Effect of Stress During Conception on Odds of a Male Birth*

### Description

The probability of a male birth in humans is about .51. It has previously been noticed that lower proportions of male births are observed when offspring is conceived at times of exposure to smog, floods or earthquakes. Danish researchers hypothesised that sources of stress associated with severe life events may also have some bearing on the sex ratio. To investigate this theory they obtained the sexes of all 3,072 children who were born in Denmark between 1 January 1980 and 31 December 1992 to women who experienced the following kind of severe life events in the year of the birth or the year prior to the birth: death or admission to hospital for cancer or heart attack of their partner or of their other children. They also obtained sexes on a sample of 20,337 births to mothers who did not experience these life stress episodes. This data frame contains the data that were collected. Noticed that for one group the exposure is listed as taking place during the first trimester of pregnancy. The rationale for this is that the stress associated with the cancer or heart attack of a family member may well have started before the recorded time of death or hospital admission.

## Usage

```
ex2117
```

## Format

A data frame with 5 observations on the following 4 variables.

**Group** Indicator for groups to which mothers belong

**Time** Indicator for time at which severe life event occurred

**Number** Number of births

**PctBoys** Percentage of boys born

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Hansen, D., Møller, H. and Olsen, J. (1999). Severe Periconceptional Life Events and the Sex Ratio in Offspring: Follow Up Study based on Five National Registers, *British Medical Journal* **319**(7209): 548–549.

## Examples

```
str(ex2117)
```

---

# ex2118 *HIV and Circumcision*

---

## Description

Researchers in Kenya identified a cohort of more that 1,000 prostitutes who were known to be a major reservoir of sexually transmitted diseases in 1985. It was determined that more than 85% of them were infected with human immunodeficiency virus (HIV) in February, 1986. The researchers identified men who acquired a sexually-transmitted disease from this group of women after the men sought treatment at a free clinic. The data frame contains data on the subset of those men who did not test positive for HIV on their first visit and who agreed to participate in the study. The men are categorised according to whether they later tested positive for HIV during the study period, whether they had one or multiple sexual contacts with the prostitutes and whether they were circumcised.

## Usage

```
ex2118
```

## Format

A data frame with 4 observations on the following 5 variables.

**Contact** Whether men had single or multiple contact with prostitutes

**Circumcised** Whether the men are circumcised, factor with levels `"No"` and `"Yes"`

**HIV** Number of men that tested positive for HIV

**Number** Number of men

**NoHIV** Number of men that did not test positive for HIV (should be Number-HIV)

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Cameron, D.W., D'Costa, L.J., Maitha, G.M., Cheang, M., Piot, P., Simonsen, J.N., Ronald, A.R., Gakinya, M.N., Ndinya-Achola, J.O., Brunham, R.C. and Plummer, F. A. (1989). Female to Male Transmission of Human Immunodeficiency Virus Type I: Risk Factors for Seroconversion in Men, *The Lancet* **334**(8660): 403–407.

## Examples

```
str(ex2118)
```

---

`ex2119` *Meta–Analysis of Breast Cancer and Lactation Studies*

---

## Description

This data frame gives the results of 10 separate case–control studies on the association of breast cancer and whether a woman had breast–fed children.

## Usage

```
ex2119
```

## Format

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

**Study** Factor indicating the study from which data was taken

**Lactate** Whether women had breast–fed children (lactated)

**Cancer** Number of women with breast cancer

**NoCancer** Number of women without breast cancer

## Details

Meta–analysis refers to the analysis of analyses. When the main results of studies can be cast into $2\times2$ tables of counts, it is natural to combine individual odds ratios with a logistic regression model that includes a factor to account for different odds from the different studies. In addition, the odds ratio itself might differ slightly among studies because of different effects on different populations or different research techniques. One approach for dealing with this is to suppose an underlying common odds ratio and to model between–study variability as extra–binomial variation.

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Data gathered from various sources by Karolyn Kolassa as part of a Master's project, Oregon State University.

## Examples

```
str(ex2119)
```

---

`ex2120` *Clever Hans Effect*

---

## Description

These data were simulated to match the summary statistics and conclusions of Rosenthal and Fode's Clever Hans experiment. Each of 12 students trained rats to run a maze. The data set contains their number of successful runs out of 50 on each of 5 days, the student's prior expectation of success (on a scale from -10 to 10), and a variable indicating treatment–whether or not the students were supplied with the fictitious information that their rights were bright.

## Usage

```
2120
```

## Format

A data frame with 60 observations on the following 5 variables.

**Student** a student identification number

**PriorExp** the student's prior expectation of rat-training success, on a scale from -10 to 10

**Treatment** a factor with levels `"bright"` and `"dull"` corresponding to whether students were told (falsely) that their rats were bright or not

**Day** day of the study, ranging from 1 to 5

**Success** the number of successful maze runs on a day, out of 50

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Rosenthal, R. and Fode, K.L. (1963) The Effect of Experimenter Bias on the Performance of the Albino Rat *Behavioral Science* **8:3**: 183–189.

## See Also

ex1419

## Examples

```
str(ex2120)
```

---

ex2216 *Murder–Suicides by Deliberate Plane Crash*

---

## Description

Some sociologist suspect that highly publicised suicides may trigger additional suicides. In one investigation of this hypothesis, D.P. Phillips collected information about 17 airplane crashes that were known (because of notes left behind) to be murder–suicides. For each of these crashes, Phillips reported an index of the news coverage (circulation of nine newspapers devoting space to the crash multiplied by length of coverage) and the number of multiple-fatality plane crashes during the week following the publicised crash. This data frame contains the collected data.

## Usage

```
ex2216
```

## Format

A data frame with 17 observations on the following 2 variables.

**Index** Index for the amount of newspaper coverage given the murder–suicide

**Crashes** Multiple-fatality crashes in the week following a murder–suicide by plane crash

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Phillips, D.P. (1978). Airplane Accident Fatalities Increase Just After Newspaper Stories About Murder and Suicide, *Science* **201**: 748–750.

## Examples

```
str(ex2216)
```

## ex2220 *Cancer Deaths of Atomic Bomb Survivors*

### Description

The data are the number of cancer deaths among survivors of the atomic bombs dropped on Japan during World War II, categorized by time (years) after the bomb that death occurred and the amount of radiation exposure that the survivors received from the blast. Also listed in each cell is the person-years at risk, in 100's. This is the sum total of all years spent by all persons in the category. The data can be analyzed by supposing the number of cancer deaths in each cell is Poisson with mean = risk x rate, where risk is the person-years at risk and rate is the rate of cancer deaths per person per year. How does the rate depend on the radiation exposure, after accounting for years after exposure?

### Usage

```
ex2220
```

### Format

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

**Exposure** radiation exposure, in rads

**YearsAfter** years after the exposure

**AtRisk** number of survivors in the group

**Deaths** number of survivors in the group who died of Cancer

### Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

### References

Pierce, D.A., personal communication

### Examples

```
str(ex2220)
```

## ex2222 *Emulating Jane Austen's Writing Style*

### Description

When she died in 1817, the English novelist Jane Austen had not yet finished the novel *Sanditon*, but she did leave notes on how she intended to conclude the book. The novel was completed by a ghost writer, who attempted to emulate Austen's style. In 1978, a researcher reported counts of some words found in chapters of books written by Austen and in chapters written by the emulator. These data are given in this data frame.

## Usage

```
ex2222
```

## Format

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

**Count** Number of occurrences of a word in various chapters of books written by Jane Austen and the ghost writer

**Book** Title of books used

**Word** Words used

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Morton, A.Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*, Charles Scribner's Sons, New York.

## Examples

```
str(ex2222)
```

---

# ex2223 *Space Shuttle O-Ring Failures*

---

## Description

On January 27, 1986, the night before the space shuttle *Challenger* exploded, an engineer recommended to the National Aeronautics and Space Administration (NASA) that the shuttle not be launched in the cold weather. The forecasted temperature for the *Challenger* launch was 31 degrees Fahrenheit—the coldest launch ever. After an intense 3-hour telephone conference, officials decided to proceed with the launch. This data frame contains the launch temperatures and the number of O-ring problems in 24 shuttle launches prior to the *Challenger*.

## Usage

```
ex2223
```

## Format

A data frame with 24 observations on the following 2 variables.

**Temp** Launch temperatures (in degrees Fahrenheit)

**Incidents** Numbers of O-ring incidents

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## See Also

`case0401`, `ex2011`

## Examples

```
str(ex2223)
```

---

`ex2224` *Valve Failure in Nuclear Reactors*

---

## Description

This data frame contains data on characteristics and numbers of *failures* observed in valve types from one pressurised water reactor.

## Usage

```
ex2224
```

## Format

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

**System** a numerical code corresponding to 5 categories of system (1 = containment, 2 = nuclear, 3 = power conversion, 4 = safety, 5 = process auxiliary)

**Operator** a numerical code corresponding to 4 different operator types (1 = air, 2 = solenoid, 3 = motor=driven, 4 = manual)

**Valve** a numerical code corresponding to 6 different valve types (1 = ball, 2 = butterfly, 3 = diaphram, 4 = gate, 5 = globe, 6 = directional control)

**Size** a numerical code corresponding to 3 head size categories (1 = less than 2 inches, 2 = 2–10 inches, 3 = 10–30 inches)

**Mode** a numerical code corresponding to two categories of operation mode (1 = normally closed, 2 = normally open)

**Failures** Number of failures observed

**Time** Lengths of observation time

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Moore, L.M. and Beckman, R.J. (1988). Appropriate One-Sided Tolerance Bounds on the Number of Failures using Poisson Regression, *Technometrics* **30**: 283–290.

## Examples

```
str(ex2224)
```

---

ex2225 *Body Size and Reproductive Success in a Population of Male Bullfrogs*

---

## Description

As an example of field observation in evidence of theories of sexual selection, S.J. Arnold and M.J. Wade presented the following data set on size and number of mates observed in 38 bullfrogs.

## Usage

```
ex2225
```

## Format

A data frame with 38 observations on the following 2 variables.

**BodySize** Body size (in mm)

**Mates** Number of mates

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Arnold, S.J. and Wade, M.J. (1984). On the Measurement of Natural and Sexual Selection: Aplications, *Evolution* **38**: 720–734.

## Examples

```
str(ex2225)
```

---

ex2226 *Number of Moons*

---

## Description

Apparently, larger planets have more moons, but is it the volume (as indicated by diameter) or mass that are more relevant, or is it both? These data include the diameter, mass, distance from the sun, and number of moons for 13 planets, gas giants, and dwarf planets in our solar system. Which size variable best explains mean number of moons (possible after accounting for distance from sun). (Consider negative binomial regression.)

## Usage

```
ex2226
```

## Format

A data frame with 13 observations on the following 5 variables.

**Name** a character variable with the name of the planet, gas giant, or dwarf planet)

**Distance** distance from sun, relative to earth's

**Diameter** diameter of the planet, relative to earth's

**Mass** mass, relative to earth's

**Moons** number of moons

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Wikipedia: http://en.wikipedia.org/wiki/Planet August 10, 2011

## See Also

ex0721

## Examples

```
str(ex2226)
```

---

# ex2414 *Amphibian Crisis and UV-B*

---

## Description

Data frame contains the percentage of unsuccessful hatching from enclosures containing 150 eggs each in a study to investigate whether UV-B is responsible for low hatch rates.

## Usage

```
ex2414
```

## Format

A data frame with 71 observations on the following 4 variables.

**Percent** percentage of frog eggs failing to hatch

**Treatment** factor variable with levels "NoFilter", "UV-BTransmitting" and "UV-BBlocking"

**Location** factor variable with levels "ThreeCreeks", "SparksLake", "SmallLake" and "LostLake"

**Photolyase** Photolyase activity

## Source

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## References

Blaustein, A.R., Hoffman, P.D., Hokit, D.G., Kiesecker, J.M., Walls, S.C. and Hays, J.B. (1994). UV Repair and Resistance to Solar UV-B in Amphibian Eggs: A Link to Population Declines? *Proceedings of the National Academy of Science, USA* **91**: 1791–1795.

## Examples

```
str(ex2414)
```

---

`Sleuth3Manual` *Manual of the R Sleuth3 package*

---

## Description

If the option "pdfviewer" is set, this command will display the PDF version of the help pages.

## Usage

```
Sleuth3Manual()
```

## Author(s)

Berwin A Turlach <Berwin.Turlach@gmail.com>

## References

Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning.

## Examples

```
## Not run: Sleuth3Manual()
```

# Index