

Package ‘datarium’

July 23, 2026

Type Package

Title Data Bank for Statistical Analysis and Visualization

Version 0.2.0

Date 2026-07-23

Description Contains data organized by topics: categorical data, regression model, means comparisons, independent and repeated measures ANOVA, mixed ANOVA and ANCOVA. The package code is under GPL-2; the data sets are covered by the terms in the 'LICENSE.note' file (the author's data sets under CC0 1.0, and the 'heartdisease' data under CC BY 4.0, see the 'heartdisease' help page).

License GPL-2

Encoding UTF-8

LazyData true

Depends R (>= 3.1.0)

URL <https://rpkgs.datanovia.com/datarium/>

BugReports <https://github.com/kassambara/datarium/issues>

RoxygenNote 7.3.3

Suggests knitr

NeedsCompilation no

Author Alboukadel Kassambara [aut, cre]

Maintainer Alboukadel Kassambara <alboukadel.kassambara@gmail.com>

Repository CRAN

Date/Publication 2026-07-23 10:50:35 UTC

Contents

AirPassengersDf	2
antismoking	3
anxiety	4
depression	5

genderweight	6
headache	7
heartattack	8
heartdisease	9
housetasks.raw	11
infections	12
jobsatisfaction	13
marketing	14
mice	15
mice2	16
performance	17
properties	18
renalstone	19
selfesteem	20
selfesteem2	21
stress	22
taskachievement	23
titanic.raw	24
weightloss	25
Index	27

AirPassengersDf	<i>Air Passengers</i>
-----------------	-----------------------

Description

This dataset contains the number of monthly air passengers from 1949 to 1960. It is taken from the R datasets package and formatted into long data frame format.

Usage

```
data("AirPassengersDf")
```

Format

A data frame with 144 rows and 2 columns.

Month the month of the observation, as a Date (the first day of each month from January 1949 to December 1960).

Passengers the monthly total of international airline passengers.

Source

The AirPassengers time series in base R (the classic Box and Jenkins airline data), reshaped into a long data frame.

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1976) Time Series Analysis, Forecasting and Control. San Francisco: Holden-Day.

Examples

```
data("AirPassengersDf")
head(AirPassengersDf)

# Monthly international airline passengers over time
plot(Passengers ~ Month, data = AirPassengersDf, type = "l")
```

antismoking

Anti-Smoking Emotive Communication Data for McNemar Test

Description

Paired nominal data providing the smoking status 62 individuals before and after emotive video communications showing the danger of smoking. This a demo dataset for McNemar test.

It is the data set used in the Datanovia tutorial “McNemar’s Test in R: Compare Paired Proportions (Before/After)” (<https://www.datanovia.com/learn/biostatistics/categorical/mcnemar-test-in-r>).

Usage

```
data("antismoking")
```

Format

A data frame with 62 rows and 3 columns (stored as a tibble).

id participant identifier (1 to 62).

before the smoking status before the communication, "non.smoker" or "smoker".

after the smoking status after the communication, "non.smoker" or "smoker".

Source

A simulated dataset for teaching McNemar’s test.

See Also

Datanovia tutorial: [McNemar’s Test in R: Compare Paired Proportions \(Before/After\)](#).

Examples

```
data(antismoking)
xtabs(~before + after, data = antismoking)

# McNemar test of the change in smoking status
mcnemar.test(table(antismoking$before, antismoking$after))
```

anxiety

Anxiety Data for Two-Way Mixed ANOVA

Description

The data provide the anxiety score, measured at three time points, of three groups of individuals practicing physical exercises at different levels (grp1: basal, grp2: moderate and grp3: high)

Two-way mixed ANOVA can be used to evaluate if there is interaction between group and time in explaining the anxiety score.

It is the data set used in the Datanovia tutorial “Mixed ANOVA in R: One Between- and One Within-Subjects Factor” (<https://www.datanovia.com/learn/biostatistics/anova/mixed-anova-in-r>).

Usage

```
data("anxiety")
```

Format

A data frame with 45 rows and 5 columns (stored as a tibble).

id participant identifier (1 to 45).

group the physical exercise level, "grp1" (basal), "grp2" (moderate) or "grp3" (high); a between-subjects factor.

t1 the anxiety score at the first time point.

t2 the anxiety score at the second time point.

t3 the anxiety score at the third time point.

The anxiety score is a simulated measure on an arbitrary scale with no real-world units.

Source

A simulated dataset created for teaching two-way mixed ANOVA.

See Also

Datanovia tutorial: [Mixed ANOVA in R: One Between- and One Within-Subjects Factor](#).

Examples

```
data(anxiety)
head(anxiety)

# Two-way mixed ANOVA: group (between) x time (within)
anxiety_long <- reshape(anxiety, varying = c("t1", "t2", "t3"),
                        v.names = "score", timevar = "time",
                        direction = "long")
summary(aov(score ~ group * factor(time) + Error(factor(id)/factor(time)),
           data = anxiety_long))
```

depression

Depression Data for Two Way Mixed ANOVA

Description

The data correspond to an experiment in which a treatment for depression is studied. Two groups of patients - control ("ctr") and treatment ("treated") - have been followed at four different time points: t0 (pre-test) and three post-test follow-up measurements (t1, t2, t3). The dependent variable is a depression score.

Repeated measures ANOVA can be performed in order to determine the effect of the treatment and the effect of time on the depression score.

For a worked tutorial on this analysis, see the Datanovia lesson “Mixed ANOVA in R: One Between- and One Within-Subjects Factor” (<https://www.datanovia.com/learn/biostatistics/anova/mixed-anova-in-r>).

Usage

```
data("depression")
```

Format

A data frame with 24 rows and 6 columns (stored as a tibble).

id participant identifier (1 to 24).

treatment the treatment group, "ctr" (control) or "treated".

t0 the depression score at pre-test.

t1 the depression score at the first post-test follow-up.

t2 the depression score at the second post-test follow-up.

t3 the depression score at the third post-test follow-up.

The depression score is a simulated measure on an arbitrary scale with no real-world units.

Source

A simulated dataset created for teaching two-way mixed ANOVA.

See Also

Datanovia tutorial: [Mixed ANOVA in R: One Between- and One Within-Subjects Factor](#).

Examples

```
data(depression)
head(depression)

# Two-way mixed ANOVA: treatment (between) x time (within)
depression_long <- reshape(depression, varying = c("t0", "t1", "t2", "t3"),
                           v.names = "score", timevar = "time",
                           direction = "long")
summary(aov(score ~ treatment * factor(time) + Error(factor(id)/factor(time)),
            data = depression_long))
```

genderweight

Weight Data By Gender for Two-Samples Mean Test

Description

Contains the weights by sex (M for male; F for female). The question is whether the average women's weight differs from the average men's weight?

A two-samples independent t-test can be performed to answer to this question.

It is the data set used in the Datanovia tutorial "T-Test in R: One-Sample, Independent (Student & Welch) & Paired" (<https://www.datanovia.com/learn/biostatistics/two-groups/t-test-in-r>).

Usage

```
data("genderweight")
```

Format

A data frame with 40 rows and 3 columns (stored as a tibble).

id participant identifier (1 to 40).

group the participant's sex, "F" (female) or "M" (male).

weight the participant's body weight.

Source

A simulated dataset for teaching the two-samples (independent) t-test.

See Also

Datanovia tutorial: [T-Test in R: One-Sample, Independent \(Student & Welch\) & Paired](#).

Examples

```
data(genderweight)
head(genderweight)

# Two-samples independent t-test: do mean weights differ by sex?
t.test(weight ~ group, data = genderweight)
```

headache

Headache Data for Three Way ANOVA

Description

A pharmaceutical company tested three treatments for migraine headache sufferers. 72 participants were enrolled in the experiments. The aim is to examine the potential of new class of treatments in lowering the pain score associated with the migraine headache episode.

The participants include 36 males and 36 females. Males and females were further (equally) subdivided into whether they were at low or high risk of migraine headache.

This data set is suited for three way Anova test.

It is the data set used in the Datanovia tutorial “Three-Way ANOVA in R: Three Factors and Their Interaction” (<https://www.datanovia.com/learn/biostatistics/anova/three-way-anova-in-r>).

Usage

```
data("headache")
```

Format

A data frame with 72 rows and 5 columns (stored as a tibble).

id participant identifier (1 to 72).

gender the participant’s gender, "male" or "female".

risk the participant’s migraine risk, "high" or "low".

treatment the treatment received, "X", "Y" or "Z".

pain_score a simulated migraine pain score (an arbitrary scale with no real-world units).

Source

A simulated dataset for teaching three-way ANOVA.

See Also

[heartattack](#), the companion three-way ANOVA data set. Datanovia tutorial: [Three-Way ANOVA in R: Three Factors and Their Interaction](#).

Examples

```
data(headache)
head(headache)

# Three-way ANOVA of the pain score
summary(aov(pain_score ~ gender * risk * treatment, data = headache))
```

heartattack

Heart Attack Data for Three Way ANOVA

Description

Measures of cholesterol concentration in 72 participants treated with three different drugs. The aim is to examine the potential of new class of drugs in lowering the cholesterol concentration and consequently reducing heart attack.

The participants include 36 males and 36 females. Males and females were further (equally) subdivided into whether they were at low or high risk of heart attack.

This data set is suited for three way Anova test.

For a worked tutorial on this analysis, see the Datanovia lesson “Three-Way ANOVA in R: Three Factors and Their Interaction” (<https://www.datanovia.com/learn/biostatistics/anova/three-way-anova-in-r>).

Usage

```
data("heartattack")
```

Format

A data frame with 72 rows and 5 columns (stored as a tibble).

gender the participant’s gender, "male" or "female".

risk the participant’s heart attack risk, "high" or "low".

drug the drug received, "A", "B" or "C".

cholesterol the blood cholesterol concentration.

id participant identifier (1 to 72).

Source

A simulated dataset for teaching three-way ANOVA.

See Also

[headache](#), the companion three-way ANOVA data set. Datanovia tutorial: [Three-Way ANOVA in R: Three Factors and Their Interaction](#).

Examples

```
data(heartattack)
head(heartattack)

# Three-way ANOVA of the cholesterol concentration
summary(aov(cholesterol ~ gender * risk * drug, data = heartattack))
```

heartdisease

Heart Disease Data (UCI, Four Sites, with Real Missing Values)

Description

A real, deliberately messy clinical dataset: 920 patients assessed for coronary artery disease at four institutions (Cleveland, Hungary, Switzerland and the Long Beach VA). Unlike the simulated teaching sets in `datarium`, it ships close to raw — integer-coded categorical variables, substantial and uneven missingness across sites, and a multi-level outcome — so it can be used to teach the "clean the data before you analyse it" step ahead of logistic regression, chi-square tests and group comparisons. The values are the original UCI codes and are **not** relabelled; see the `data-raw/heartdisease/heart-disease.names` codebook.

For a worked tutorial on this analysis, see the Datanovia lesson "Logistic Regression in R: Model a Binary Outcome" (<https://www.datanovia.com/learn/biostatistics/regression/logistic-regression-in-r>).

Usage

```
data("heartdisease")
```

Format

A data frame with 920 rows and 15 columns. The missing marker "?" in the source has been converted to NA; missingness is heavy and very uneven across sites (see `colSums(is.na(heartdisease))`).

age age, in years.

sex sex, integer-coded: 1 = male, 0 = female.

cp chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic.

trestbps resting blood pressure (mm Hg on admission to hospital).

chol serum cholesterol (mg/dl). Note: 172 patients are recorded with 0 — all 123 of the Switzerland site's patients and a further 49 at the Long Beach VA site — where 0 means "not measured" rather than a true value: a real missing-data trap to handle before analysis.

fbs fasting blood sugar > 120 mg/dl: 1 = true, 0 = false.

restecg resting electrocardiographic results: 0 = normal, 1 = ST-T wave abnormality, 2 = probable or definite left ventricular hypertrophy (Estes' criteria).

thalach maximum heart rate achieved.

exang exercise-induced angina: 1 = yes, 0 = no.

oldpeak ST depression induced by exercise relative to rest.

slope slope of the peak exercise ST segment: 1 = upsloping, 2 = flat, 3 = downsloping.

ca number of major vessels (0–3) coloured by fluoroscopy.

thal thallium stress-test result: 3 = normal, 6 = fixed defect, 7 = reversible defect.

num diagnosis of heart disease (angiographic disease status): 0 = < 50% diameter narrowing (no disease); 1–4 = > 50% narrowing (disease present). It is commonly binarised to absence (0) vs presence (> 0).

dataset the collecting site, a factor with levels "cleveland", "hungarian", "switzerland" and "va".

Source

Heart Disease, UCI Machine Learning Repository ([doi:10.24432/C52P4X](https://doi.org/10.24432/C52P4X)). Real patient data collected by the Hungarian Institute of Cardiology, Budapest; University Hospital, Zurich; University Hospital, Basel; and the V.A. Medical Center, Long Beach, and shared by the Cleveland Clinic Foundation. This dataset combines the four processed `<site>.data` files (the 14 commonly-used attributes of the original 76), converting the "?" marker to NA and keeping the raw values otherwise unchanged; the source files and codebook are in `data-raw/heartdisease/`. Distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence; when using the data please cite Janosi et al. (1989) (see references).

References

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. [doi:10.24432/C52P4X](https://doi.org/10.24432/C52P4X).

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304–310.

See Also

Datanovia tutorial: [Logistic Regression in R: Model a Binary Outcome](#).

Examples

```
data("heartdisease")

# Real data is messy: missingness is heavy and very uneven across sites
colSums(is.na(heartdisease))
table(heartdisease$dataset)

# The Cleveland site is the most complete; use it for a first model
cleveland <- subset(heartdisease, dataset == "cleveland")

# Binarise the 0-4 angiographic outcome into disease present/absent
cleveland$disease <- as.integer(cleveland$num > 0)

# Logistic regression of disease on age, sex and maximum heart rate
fit <- glm(disease ~ age + sex + thalach, family = binomial, data = cleveland)
summary(fit)
```

```
# Chi-square test: is disease associated with sex (all sites)?
heartdisease$disease <- as.integer(heartdisease$num > 0)
chisq.test(table(heartdisease$sex, heartdisease$disease))
```

housetasks.raw

Housetasks

Description

A data frame containing the frequency of execution of 13 house tasks in the couple.

It is the data set used in the Datanovia tutorial “Chi-Square Test of Independence in R: Are Two Variables Associated?” (<https://www.datanovia.com/learn/biostatistics/categorical/chi-square-test-of-independence-in-r>).

Usage

```
data("housetasks.raw")
```

Format

A data frame with 1744 rows and 2 columns (stored as a tibble). Each row is one recorded performance of a household task, recovered from the original contingency table.

tasks the household task, one of 13 categories: "Laundry", "Main_meal", "Dinner", "Breakfeast", "Tidying", "Dishes", "Shopping", "Official", "Driving", "Finances", "Insurance", "Repairs" and "Holidays". ("Breakfeast" is a misspelling of "Breakfast" carried over from the source data.)

status who performs the task: "Partner1", "Alternating", "Parter2" or "Jointly". The two partner categories were relabeled from the source data's "Wife" and "Husband"; "Parter2" is a misspelling of "Partner2" retained as-is in the stored data.

Source

Derived from the housetasks data set in the **factoextra** package, with its "Wife" and "Husband" columns relabeled "Partner1" and "Partner2", expanded to one row per recorded case.

See Also

Datanovia tutorial: [Chi-Square Test of Independence in R: Are Two Variables Associated?](https://www.datanovia.com/learn/biostatistics/categorical/chi-square-test-of-independence-in-r).

Examples

```
data(housetasks.raw)
table(housetasks.raw)
```

```
# Chi-square test of association between task and who performs it
chisq.test(table(housetasks.raw$tasks, housetasks.raw$status))
```

infections

Infection Count Data for Poisson Regression

Description

A simulated dataset giving the number of infections recorded for 100 patients over a one-year follow-up in a placebo-controlled study. It is designed for teaching Poisson regression of a count outcome on a treatment group and a continuous covariate.

Usage

```
data("infections")
```

Format

A data frame with 100 rows and 4 columns.

id patient identifier (1 to 100).

treatment the treatment group, "placebo" or "treated".

age the patient's age, in years.

count the number of infections recorded during the one-year follow-up (a count outcome).

Source

A simulated dataset created for teaching Poisson regression. The counts were generated as $\text{count} \sim \text{Poisson}(\mu)$ with $\log(\mu) = 1.6 - 0.8 * (\text{treatment} == \text{"treated"}) + 0.25 * ((\text{age} - 40) / 10)$ (see `data-raw/infections.R`), so the treated group has a lower infection rate.

Examples

```
data(infections)
head(infections)

# Poisson regression of the infection count on treatment and age
fit <- glm(count ~ treatment + age, family = poisson, data = infections)
summary(fit)
```

`jobsatisfaction`*Job Satisfaction Data for Two-Way ANOVA*

Description

Contains the job satisfaction score organized by gender and education level.

It is the data set used in the Datanovia tutorial “Two-Way ANOVA in R: Two Factors and Their Interaction” (<https://www.datanovia.com/learn/biostatistics/anova/two-way-anova-in-r>).

Usage

```
data("jobsatisfaction")
```

Format

A data frame with 58 rows and 4 columns (stored as a tibble).

id participant identifier (1 to 58).

gender the participant’s gender, "male" or "female".

education_level the participant’s education level, "school", "college" or "university".

score a simulated job satisfaction score (an arbitrary scale with no real-world units).

Source

A simulated dataset for teaching two-way ANOVA.

See Also

Datanovia tutorial: [Two-Way ANOVA in R: Two Factors and Their Interaction](#).

Examples

```
data(jobsatisfaction)
head(jobsatisfaction)

# Two-way ANOVA of satisfaction by gender and education level
summary(aov(score ~ gender * education_level, data = jobsatisfaction))
```

marketing

Marketing Data Set

Description

A data frame containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales (in thousands of units). The advertising experiment has been repeated 200 times. This is a simulated data.

It is the data set used in the Datanovia tutorial “Multiple Linear Regression in R: Several Predictors” (<https://www.datanovia.com/learn/biostatistics/regression/multiple-linear-regression-in-r>).

Usage

```
data("marketing")
```

Format

A data frame with 200 rows and 4 columns.

youtube the advertising budget spent on YouTube, in thousands of dollars.

facebook the advertising budget spent on Facebook, in thousands of dollars.

newspaper the advertising budget spent on newspaper, in thousands of dollars.

sales the sales generated, in thousands of units.

Source

A simulated dataset for teaching multiple linear regression.

See Also

Datanovia tutorial: [Multiple Linear Regression in R: Several Predictors](#).

Examples

```
data(marketing)
res.lm <- lm(sales ~ youtube*facebook, data = marketing)
summary(res.lm)
```

`mice`*Mice Weight Data for One Sample Mean Test*

Description

Contains the weight of 10 mice. The question is whether the average weight of the mice differs from 25g.

A one sample t-test can be performed to answer to this question.

It is the data set used in the Datanovia tutorial “T-Test in R: One-Sample, Independent (Student & Welch) & Paired” (<https://www.datanovia.com/learn/biostatistics/two-groups/t-test-in-r>).

Usage

```
data("mice")
```

Format

A data frame with 10 rows and 2 columns (stored as a tibble).

name mouse identifier, "M_1" to "M_10".

weight the mouse weight, in grams.

Source

A simulated dataset for teaching the one-sample t-test.

See Also

[mice2](#), the paired-samples companion data set. Datanovia tutorial: [T-Test in R: One-Sample, Independent \(Student & Welch\) & Paired](#).

Examples

```
data(mice)
head(mice)

# One-sample t-test: does the mean weight differ from 25 g?
t.test(mice$weight, mu = 25)
```

`mice2`*Mice Weight Data for Paired-Samples Mean Test*

Description

contains the weight of 10 mice before and after the treatment.

A paired-samples t-test can be performed to answer to this question.

It is the data set used in the Datanovia tutorial “T-Test in R: One-Sample, Independent (Student & Welch) & Paired” (<https://www.datanovia.com/learn/biostatistics/two-groups/t-test-in-r>).

Usage

```
data("mice2")
```

Format

A data frame with 10 rows and 3 columns.

id mouse identifier (1 to 10).

before the mouse weight before the treatment.

after the mouse weight after the treatment.

Source

A simulated dataset for teaching the paired-samples t-test.

See Also

[mice](#), the one-sample companion data set. Datanovia tutorial: [T-Test in R: One-Sample, Independent \(Student & Welch\) & Paired](#).

Examples

```
data(mice2)
head(mice2)

# Paired-samples t-test: did the treatment change the weight?
t.test(mice2$before, mice2$after, paired = TRUE)
```

performance

Performance Data for Three-Way Mixed ANOVA

Description

Contains the performance score measures of participants at two time points. The aim of this study is to evaluate the effect of gender and stress on performance score. The three-way mixed ANOVA test can be used to investigate this question.

The data include two between-subjects factors (gender and stress) and one within-subject factor (time, repeated measures).

It is the data set used in the Datanovia tutorial “Mixed ANOVA in R: One Between- and One Within-Subjects Factor” (<https://www.datanovia.com/learn/biostatistics/anova/mixed-anova-in-r>).

Usage

```
data("performance")
```

Format

A data frame with 60 rows and 5 columns (stored as a tibble).

id participant identifier (1 to 60).

gender the participant’s gender, "male" or "female".

stress the stress level, "low", "moderate" or "high".

t1 a simulated performance score at the first time point (an arbitrary scale with no real-world units).

t2 the performance score at the second time point (same scale as t1).

Source

A simulated dataset for teaching three-way mixed ANOVA.

See Also

Datanovia tutorial: [Mixed ANOVA in R: One Between- and One Within-Subjects Factor](#).

Examples

```
data(performance)
head(performance)

# Three-way mixed ANOVA: gender and stress (between) x time (within)
perf_long <- reshape(performance, varying = c("t1", "t2"),
                     v.names = "score", timevar = "time", direction = "long")
summary(aov(score ~ gender * stress * factor(time) + Error(factor(id)/factor(time)),
           data = perf_long))
```

properties

*Properties Data for Chi-square Test of Independence***Description**

Contains the type of properties and the buyer types. Buyer categories are: "single male", "single female", "married couple" and "family".

The type of property these buyers purchased were sorted into four categories: "flat", "bungalow" (i.e., a one-storey home), "detached house" and "terrace" (i.e., a block of adjoining houses).

Chi-square test of independence can be used to assess the association between the type of buyer who purchases a property and the type of property that is purchased.

For a worked tutorial on this analysis, see the Datanovia lesson “Chi-Square Test of Independence in R: Are Two Variables Associated?” (<https://www.datanovia.com/learn/biostatistics/categorical/chi-square-test-of-independence-in-r>).

Usage

```
data("properties")
```

Format

A data frame with 333 rows and 2 columns (stored as a tibble).

property_type the type of property purchased: "flat", "bungalow", "detached house" or "terrace".

buyer_type the type of buyer: "single male", "single female", "married couple" or "family".

Source

A simulated dataset for teaching the chi-square test of independence.

See Also

Datanovia tutorial: [Chi-Square Test of Independence in R: Are Two Variables Associated?](#).

Examples

```
data("properties")  
head(properties)
```

```
# Chi-square test of independence between buyer type and property type  
chisq.test(table(properties$property_type, properties$buyer_type))
```

renalstone*Risk of Renal Stone Data for Cochran-Armitage Trend Test*

Description

Presents the frequencies of individuals at high risk of renal calculi according to age and gender. This a demo dataset for Cochran-Armitage trend test for investigating whether there is a linear trend between the proportion of individual with renal stone and ages.

It is the data set used in the Datanovia tutorial “Cochran-Armitage Trend Test in R: Test a Trend in Proportions” (<https://www.datanovia.com/learn/biostatistics/categorical/cochran-armitage-trend-test-in-r>)

Usage

```
data("renalstone")
```

Format

A data frame with 3513 rows and 3 columns (stored as a tibble).

gender the individual's gender, "male" or "female".

stone whether the individual has a renal stone, "yes" or "no".

age the age group, an ordered factor: "30-39" < "40-49" < "50-59".

Source

Frequencies reported by Hazra and Gogtay (2016), expanded to one row per individual.

References

Hazra, Avijit, and Nithya Jaideep Gogtay. 2016. "Biostatistics Series Module 4: Comparing Groups - Categorical Variables." In Indian Journal of Dermatology.

See Also

Datanovia tutorial: [Cochran-Armitage Trend Test in R: Test a Trend in Proportions](#).

Examples

```
data(renalstone)
xtabs(~stone+age+gender, data = renalstone)
```

selfesteem*Self-Esteem Score Data for One-way Repeated Measures ANOVA*

Description

The dataset contains 10 individuals' self-esteem score on three time points during a specific diet to determine whether their self-esteem improved.

One-way repeated measures ANOVA can be performed in order to determine the effect of time on the self-esteem score.

It is the data set used in the Datanovia tutorial "Repeated Measures ANOVA in R: Compare Means Across Time" (<https://www.datanovia.com/learn/biostatistics/anova/repeated-measures-anova-in-r>).

Usage

```
data("selfesteem")
```

Format

A data frame with 10 rows and 4 columns (stored as a tibble).

id participant identifier (1 to 10).

t1 the self-esteem score at the first time point.

t2 the self-esteem score at the second time point.

t3 the self-esteem score at the third time point.

The self-esteem score is a simulated measure on an arbitrary scale with no real-world units.

Source

A simulated dataset created for teaching one-way repeated measures ANOVA.

See Also

[selfesteem2](#), the two-way (treatment x time) companion data set. Datanovia tutorial: [Repeated Measures ANOVA in R: Compare Means Across Time](#).

Examples

```
data(selfesteem)
head(selfesteem)

# One-way repeated measures ANOVA of the self-esteem score over time
selfesteem_long <- reshape(selfesteem, varying = c("t1", "t2", "t3"),
                           v.names = "score", timevar = "time",
                           direction = "long")
summary(aov(score ~ factor(time) + Error(factor(id)/factor(time)),
            data = selfesteem_long))
```

selfesteem2*Self Esteem Score Data for Two-way Repeated Measures ANOVA*

Description

Data are the self esteem score of 12 individuals enrolled in 2 successive short-term trials (4 weeks) - control (placebo) and special diet trials.

The self esteem score was recorded at three time points: at the beginning (t1), midway (t2) and at the end (t3) of the trials.

The same 12 participants are enrolled in the two different trials with enough time between trials.

Two-way repeated measures ANOVA can be performed in order to determine whether there is interaction between time and treatment on the self esteem score.

For a worked tutorial on this analysis, see the Datanovia lesson “Repeated Measures ANOVA in R: Compare Means Across Time” (<https://www.datanovia.com/learn/biostatistics/anova/repeated-measures-anova-in-r>).

Usage

```
data("selfesteem2")
```

Format

A data frame with 24 rows and 5 columns (stored as a tibble).

id participant identifier (1 to 12); each participant takes part in both trials.

treatment the trial, "ctr" (control/placebo) or "Diet".

t1 the self esteem score at the beginning of the trial.

t2 the self esteem score midway through the trial.

t3 the self esteem score at the end of the trial.

The self esteem score is a simulated measure on an arbitrary scale with no real-world units.

Source

A simulated dataset created for teaching two-way repeated measures ANOVA.

See Also

[selfesteem](#), the one-way companion data set. Datanovia tutorial: [Repeated Measures ANOVA in R: Compare Means Across Time](#).

Examples

```
data(selfesteem2)
head(selfesteem2)

# Two-way repeated measures ANOVA (treatment x time), both within-subjects
selfesteem2_long <- reshape(selfesteem2, varying = c("t1", "t2", "t3"),
                             v.names = "score", timevar = "time",
                             idvar = c("id", "treatment"), direction = "long")
summary(aov(score ~ treatment * factor(time) +
             Error(factor(id)/(treatment * factor(time))),
         data = selfesteem2_long))
```

stress

Stress Data for Two-Way ANCOVA

Description

Researchers want to evaluate the effect of a new "treatment" and "exercise" on the stress score reduction after adjusting for "age".

Two-way ANCOVA can be performed in order to determine whether there is interaction between exercise and treatment on the stress score.

It is the data set used in the Datanovia tutorial “ANCOVA in R: Compare Group Means Adjusted for a Covariate” (<https://www.datanovia.com/learn/biostatistics/anova/ancova-in-r>).

Usage

```
data("stress")
```

Format

A data frame with 60 rows and 5 columns (stored as a tibble).

id participant identifier (1 to 60).

score the stress score; a simulated measure on an arbitrary scale with no real-world units.

treatment whether the participant received the treatment, "yes" or "no".

exercise the exercise level, "low", "moderate" or "high".

age the participant's age, in years.

Source

A simulated dataset created for teaching two-way ANCOVA.

See Also

Datanovia tutorial: [ANCOVA in R: Compare Group Means Adjusted for a Covariate](#).

Examples

```
data(stress)
head(stress)

# Two-way ANCOVA of the stress score, adjusting for age
summary(aov(score ~ age + treatment * exercise, data = stress))
```

taskachievement	<i>Task Achievement Data for Cochran's Q Test</i>
-----------------	---

Description

Repeated measures nominal designs where 73 subjects are asked to perform 3 tasks. The outcome of each task is a dichotomous value, success or failure. Each row correspond to a participant (called "block" in the jargon). This a demo dataset for Cochran's Q test.

It is the data set used in the Datanovia tutorial "Cochran's Q Test in R: Compare Three or More Paired Proportions" (<https://www.datanovia.com/learn/biostatistics/categorical/cochran-q-test-in-r>).

Usage

```
data("taskachievement")
```

Format

A data frame with 73 rows and 4 columns (stored as a tibble).

participant participant identifier (1 to 73).

Task1 the outcome of task 1: 0 (failure) or 1 (success).

Task2 the outcome of task 2: 0 (failure) or 1 (success).

Task3 the outcome of task 3: 0 (failure) or 1 (success).

Source

A simulated dataset for teaching Cochran's Q test.

See Also

Datanovia tutorial: [Cochran's Q Test in R: Compare Three or More Paired Proportions](#).

Examples

```
data(taskachievement)
head(taskachievement)

# Proportion of success for each task
colMeans(taskachievement[, c("Task1", "Task2", "Task3")])
```

titanic.raw*Survival of Passengers on the Titanic*

Description

Survival of passengers on the Titanic. This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic". Columns are economic status (Class), Sex, Age and Survived.

It is the data set used in the Datanovia tutorial “Descriptive Statistics for Categorical Data in R: Contingency Tables” (<https://www.datanovia.com/learn/biostatistics/categorical/categorical-descriptive-statistics-in-r>).

Usage

```
data("titanic.raw")
```

Format

A data frame with 2201 rows and 4 columns.

Class the passenger’s economic status, "1st", "2nd", "3rd" or "Crew".

Sex the passenger’s sex, "Male" or "Female".

Age the age category, "Child" or "Adult".

Survived whether the passenger survived, "No" or "Yes".

Source

Derived from the Titanic data set in base R, expanded to one row per passenger.

See Also

Datanovia tutorial: [Descriptive Statistics for Categorical Data in R: Contingency Tables](#).

Examples

```
data(titanic.raw)
with(titanic.raw, table(Class, Survived))
```

weightloss*Weight Loss Score Data for Three-way Repeated Measures ANOVA*

Description

A researcher wanted to assess the effects of Diet and Exercises on weight loss in 12 sedentary males.

The participants were enrolled in four trials: (1) no diet and no exercises; (2) diet only; (3) exercises only; and (4) diet and exercises combined.

Each participant performed all four trials. The order of the trials was counterbalanced and sufficient time was allowed between trials to allow any effects of previous trials to have dissipated (i.e., a "wash out" period).

Each trial lasted nine weeks and the weight loss score was measured at the beginning of each trial (t1), at the midpoint of each trial (t2) and at the end of each trial (t3).

Three-way repeated measures ANOVA can be performed in order to determine whether there is interaction between diet, exercises and time on the weight loss score.

For a worked tutorial on this analysis, see the Datanovia lesson "Repeated Measures ANOVA in R: Compare Means Across Time" (<https://www.datanovia.com/learn/biostatistics/anova/repeated-measures-anova-in-r>).

Usage

```
data("weightloss")
```

Format

A data frame with 48 rows and 6 columns (stored as a tibble).

id participant identifier (1 to 12); each participant performs all four diet/exercises trials.

diet whether the trial included a diet, "no" or "yes".

exercises whether the trial included exercises, "no" or "yes".

t1 the weight loss score at the beginning of the trial.

t2 the weight loss score at the midpoint of the trial.

t3 the weight loss score at the end of the trial.

The weight loss score is a simulated measure on an arbitrary scale with no real-world units.

Source

A simulated dataset created for teaching three-way repeated measures ANOVA.

See Also

Datanovia tutorial: [Repeated Measures ANOVA in R: Compare Means Across Time](#).

Examples

```
data(weightloss)
head(weightloss)

# Three-way repeated measures ANOVA (diet x exercises x time), all within
weightloss_long <- reshape(weightloss, varying = c("t1", "t2", "t3"),
                           v.names = "score", timevar = "time",
                           idvar = c("id", "diet", "exercises"),
                           direction = "long")
summary(aov(score ~ diet * exercises * factor(time) +
            Error(factor(id)/(diet * exercises * factor(time))),
        data = weightloss_long))
```

Index

AirPassengersDf, 2
antismoking, 3
anxiety, 4

depression, 5

genderweight, 6

headache, 7, 8
heartattack, 7, 8
heartdisease, 9
housetasks.raw, 11

infections, 12

jobsatisfaction, 13

marketing, 14
mice, 15, 16
mice2, 15, 16

performance, 17
properties, 18

renalstone, 19

selfesteem, 20, 21
selfesteem2, 20, 21
stress, 22

taskachievement, 23
titanic.raw, 24

weightloss, 25