Package ‘math3150package’

September 22, 2024

Title Package for MATH 3150 - Applied Statistics at SUU

Version 0.0.0.9000

Description This package contains functions and data files needed for MATH 3150 - Applied Statistics taught at Southern Utah University.

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 4.4),
nortest,
tidyverse

LazyData true

Author Rick Brown [aut, cre]

Maintainer Rick Brown <richardbrown1@suu.edu>

Contents

baseball 2
bf_test 3
births 3
cars99 4
ci_capture 4
class_data_f2019 5
diseases 6
epilepsy 6
idealwt 7
influence_plots 7
logistic_plots 8
mlbsalaries 8
nitrogen 9
norm_test 9
test_levene 10

Index **11**

`baseball` *baseball Data Set*

Description

This data set contains information for 337 baseball players.

Usage

```
baseball
```

Format

A tibble with 337 rows and 28 variables:

salary The salary in $1000s.

average Batting average of the player.

obp On base percentage of the player

runs Number of runs scored.

hits Number of hits in total.

doubles Number of doubles hit.

triples Number of triples hit.

homeruns Number of homeruns hit.

rbis Number of runs batted in.

walks Number of times walked.

sos Number of strikeouts.

sbs Number of stolen bases.

errors Number of errors committed.

freeagent Factor indicating whether the player is a free agent or is eligible for free agency.

arbitration Factor indicating whether the player has arbitration or is eligible for arbitration.

runsperso Number of runs per strikeout (runs/sos).

hitsperso Number of hits per strikeout (hits/sos).

hrsperso Number of homeruns per strikeout (homeruns/sos).

rbisperso Number of rbis per strikeout (rbis/sos).

walksperso Number of walks per strikeout (walks/sos).

obppererror On base percentage per error (obp/errors).

runspererror Number of runs scored per error (runs/errors).

hitspererror Number of hits per error (hits/errors).

hrspererror Number of homeruns per error (homeruns/errors).

sospererror Number of strikeouts per error (sos/errors).

sbsobp Number of stolen bases times on base percentage (sbs*obp).* \itemsbsrunsNumber of stolen bases times number of runs scored (sbsruns).

sbshits Number of stolen bases times number of hits (sbs*hits).

`bf_test` *Perform the Brown-Forsythe Test for Equality of Variance*

Description

When given a model, this will break up into groups of equal size with a default of 2 perform a Brown Forsythe test. By default, the test will be performed with the jackknife residuals.

Usage

```
bf_test(model, num_groups = 2, resid_type = "jackknife", plot_graph = TRUE)
```

Arguments

`model`	A model of type lm or glm.
`num_groups`	The number of groups to be used. The default is 2.
`resid_type`	The type of residuals. The default is "jackknife", but supports "raw", "standard", and "pearson".
`plot_graph`	Whether to return a plot or not. The default is TRUE.

Value

This function returns an ANOVA table and, if requested a plot of the residuals

Examples

```
mod <- lm(mpg ~ disp, data = mtcars)
bf_test(mod)
```

`births` *births Data Set*

Description

Data from 1995-1997 for a study hat examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data set consists of 278 cases (rows) with two columns indicating the level of prenatal care and type of birth for each set of twins.

Usage

```
births
```

Format

A tibble with 278 rows and 2 variables:

prenatal A factor indicating the prenatal care the mother received: Adequate, Inadequate, or Intensive.

type A factor indicating the classification of the birth: "Preterm (induced or cesarean)", "Preterm (without procedures)", and "Term or post-term"

`cars99` *cars99 Data Set*

Description

This data set contains information on 109 vechicles from 1999.

Usage

```
cars99
```

Format

A tibble with 109 rows and 11 variables:

Model The vechicle model name.

CityMPG The miles per gallon (MPG) for the vehicle in the city.

HwyMPG The miles per gallon (MPG) for the vehicle on the highway

FuelCap The fuel capacity of the vehicle in gallons.

Weight The weight of the vehicle in lbs.

FrontWt The front weight of the vehicle.

Accel0_30 The time it takes, in seconds, for the vehicle to accelerate from 0 to 30 mph.

Accel0_60 The time it takes, in seconds, for the vehicle to accelerate from 0 to 60 mph.

QtrMile The time it takes, in seconds, for the vehicle to travel a quarter of a mile.

`ci_capture` *Computes many confidence intervals and returns how many capture the true mean*

Description

This function reads in a given sample size, mean mu, standard deviation sigma, confidence coefficient, confidence interval type ("z" or "t"), and the number of simulated samples desired, and returns a count of how many of the corresponding confidence intervals captured the true mean mu.

Usage

```
ci_capture(n, mu, sigma, conf_level = 95, ci_type, n_ints, plot_graph = TRUE)
```

Arguments

`n` Sample size of each sample.

`mu` The true population mean.

`conf_level` The confidence level. By default, this it 95 for 95% intervals.

`ci_type` The type of confidence interval to create. "z" for a z-interval and "t" for a t-interval.

`n_ints` The number of intervals to create.

`plot_graph` Whether to return a plot or not. The default is TRUE.

Value

This function returns a graph of all the intervals, if requested, and the number of intervals that contained the true mean.

Examples

```
ci_capture(10, 10, 5, 95, "t", 500)
```

`class_data_f2019` *class_data_f2019 Data Set*

Description

This data set contains information on 42 students from Fall of 2019.

Usage

```
class_data_f2019
```

Format

A tibble with 42 rows and 7 variables:

level A factor indicating the class level the student is: Freshman, Sophomore, Junior, Senior, or Graduate.

major A character indicating the major of the student.

sex A factor indicating the sex of the student: F for female or M for male.

ski A factor indicating downhill preference: Ski, Snowboard, or Neither.

penny A factor indicating preference regarding the penny: Abolish, Retain or No Answer.

speed An integer indicating the fastest speed the student had driven in a vehicle (in mph).

sleep A numeric variable indicating how long the student slept the night before.

`diseases` *diseases Data Set*

Description

This data set contains information for each state and Washington, D.C. about the number of reported cases of AIDS, syphilis, and tuberculosis.

Usage

```
diseases
```

Format

A tibble with 51 rows and 4 variables:

State A charcter vector indicating the state.

AIDS The number of reporeted AIDS cases.

Syphilis The number of reporeted syphilis cases.

Tuberculosis The number of reporeted tuberculosis cases.

`epilepsy` *epilepsy Data Set*

Description

This data set contains information on the number of seizures, which treatment, and the age of 59 patients with epilepsy.

Usage

```
epilepsy
```

Format

A tibble with 59 rows and 4 variables:

id The patient ID.

numseiz The number of seizures the patient had.

age The age of the patient

`idealwt` *idealwt Data Set*

Description

This data set contains weight information on 182 people (119 females and 63 males). Actual weights, ideal weights, and the difference between them are recorded.

Usage

```
idealwt
```

Format

A tibble with 182 rows and 4 variables:

sex A factor indicating the sex of the person: Female or Male.

actual The person's actual weight.

ideal The person's ideal weight.

diff The difference between the person's actual weight and their ideal weight. Negative values indicate that the person weighs less than what they consider ideal.

`influence_plots` *Creates many diagnostic plots*

Description

This function creates many diagnostic plots for a given model. These plots include residual plots, a leverage plot, a Cook's distance plot, a DfFits plot, and DfBetas plots for the intercept and all slopes.

Usage

```
influence_plots(model, missing_group = NULL)
```

Arguments

`model`	The model for which we would like these plots. This can be of class "lm" or "glm" with a binomial family.
`missing_group`	Used for multinomial regression to indicate which group is not being plotted.

Value

This function returns plots. The user must press "enter" or "return" to view subsequent plots.

Examples

```
mod <- lm(mpg ~ disp, data = mtcars)
influence_plots(mod)
```

`logistic_plots` *Creates many diagnostic plots for logistic or multinomial models*

Description

This function feeds into the `influence_plots()` function to create many diagnostic plots for a given logistic or multinomial model. These plots include residual plots, a leverage plot, a Cook's distance plot, a DfFits plot, and DfBetas plots for the intercept and all slopes.

Usage

```
logistic_plots(model)
```

Arguments

`model`	The model for which we would like these plots. This can be of class "glm" with a binomial family or of class "multinom" from the `nnet` library.
`missing_group`	Used for multinomial regression to indicate which group is not being plotted.

Value

This function returns plots. The user must press "enter" or "return" to view subsequent plots.

Examples

```
mod <- glm(vs ~ disp, data = mtcars, family = "binomial")
logistic_plots(mod)
```

`mlbsalaries` *mlbsalaries Data Set*

Description

This data set contains information on 877 MLB players from 2018.

Usage

```
mlbsalaries
```

Format

A tibble with 877 rows and 5 variables:

Rank The ranking of the player's salary with 1 being the most money earned.

Name A character vector containing the name of the player.

Team A character vector containing the team the player played for.

Position A factor indicating what position the player played: SP for starting pitcher, 1B for first base, 2B for second base, 3B for third base, SS for shortstop, OF for outfield, RP for relief pitcher, and DH for designated hitter.

Salary A numeric vector containing the salary the player made for the 2018 season.

nitrogen *nitrogen Data Set*

Description

Nitrogen content of trees in an orchard, the growing tips of 150 leaves are clipped from trees throughout the orchard. These leaves are ground to form one composite sample, which the researcher assays for percentage of nitrogen. Composite samples obtained from a random sample of 36 orchards throughout the state gave the nitrogen contents.

Usage

```
nitrogen
```

Format

A data frame with 36 rows 1 variable:

nitrogen The nitrogen content for the trees.

norm_test *Performs either the Shapiro-Francia or the Shapiro-Wilk normality test*

Description

When given a model, this will perfrom either the Shapiro-Francia or the Shapiro-Wilk normality test depending user request or based on sample size. Both of these tests have a null hypothesis that the residuals are normally distributed and an alternative that the residuals are not normally distributed.

Usage

```
norm_test(model, resid_type = "raw", test = "default", plot_graph = TRUE)
```

Arguments

`model`	A model of type lm or glm.
`resid_type`	The type of residuals. The default is "raw", but supports "jackknife", and "standard" as well.
`test`	The type of test that should be performed. Options are "default", "sf", and "sw". "sf" performs the Shapiro-Francia test, "sw" performs the Shapiro-Wilk, and "defualt" performs the Shapiro-Francia is n > 30 or performs the Shapiro-Wilk if n <= 30.
`plot_graph`	Whether to return a plot or not. The default is TRUE.

Value

This function returns a test statistic, and p-value and, if requested a plot of the residuals

Examples

```
mod <- lm(mpg ~ disp, data = mtcars)
bf_test(mod)
```

test_levene — *Performs Levene’s test many times to investigate standard deviations needed to reject that the true SDs are equal*

Description

This function performs Leven’s test for the equality of variance in the ANOVA setting many times for three randomly generated normal samples when given the sample size and records how often the null is rejected, what the standard deviations for each group is when the null isrejected, the p-values when the null was rejected, and the ratio of largest SD to smallest SD for each time the null was rejected.

Usage

```
test_levene(n, num_sims)
```

Arguments

n	Sample size of each group.
num_sims	The number of Levene’s tests that should be performed.

Value

This function returns a list that contain the standard deviations for each group when we reject the null hypothesis in Levene’s test, the p-values for each time we rejected the null hypothesis, the number of samples, the ratio of largest SD to smallest SD for each time we rejected the null, and the number of times we rejected the null.

Examples

```
test_levene(10, 1000)
```

Index

∗ **datasets**
baseball, 2
births, 3
cars99, 4
class_data_f2019, 5
diseases, 6
epilepsy, 6
idealwt, 7
mlbsalaries, 8
nitrogen, 9

baseball, 2
bf_test, 3
births, 3

cars99, 4
ci_capture, 4
class_data_f2019, 5

diseases, 6

epilepsy, 6

idealwt, 7
influence_plots, 7

logistic_plots, 8

mlbsalaries, 8

nitrogen, 9
norm_test, 9

test_levene, 10