Package ‘statsr’

May 5, 2020

Type Package

Title Companion Software for the Coursera Statistics with R Specialization

Version 0.2.0

Date 2020-05-05

Maintainer Merlise Clyde <clyde@duke.edu>

Description Data and functions to support Bayesian and frequentist inference and decision making for the Coursera Specialization "Statistics with R". See <https://github.com/StatsWithR/statsr> for more information.

LazyData true

License MIT + file LICENSE

RoxygenNote 7.1.0

Encoding UTF-8

Depends R (>= 3.2.0)

Imports dplyr, rmarkdown, knitr, ggplot2, broom, gridExtra, shiny, cubature, tidyr, tibble, utils

Suggests HistData

URL https://github.com/StatsWithR/statsr

BugReports https://github.com/StatsWithR/statsr/issues

NeedsCompilation no

Author Colin Rundel [aut], Mine Cetinkaya-Rundel [aut], Merlise Clyde [aut, cre], David Banks [aut]

Repository CRAN

Date/Publication 2020-05-05 14:10:02 UTC
allow_shiny

Simple check to determine if code is being run in RStudio with the shiny runtime

Description

Simple check to determine if code is being run in RStudio with the shiny runtime

Usage

allow_shiny()
**Description**


**Usage**

ames

**Format**

A tbl_df with 2930 rows and 82 variables:

- **Order** Observation number.
- **PID** Parcel identification number - can be used with city web site for parcel review.
- **area** Above grade (ground) living area square feet.
- **price** Sale price in USD.
- **MS.SubClass** Identifies the type of dwelling involved in the sale.
- **MS.Zoning** Identifies the general zoning classification of the sale.
- **Lot.Frontage** Linear feet of street connected to property.
- **Lot.Area** Lot size in square feet.
- **Street** Type of road access to property.
- **Alley** Type of alley access to property.
- **Lot.Shape** General shape of property.
- **Land.Contour** Flatness of the property.
- **Utilities** Type of utilities available.
- **Lot.Config** Lot configuration.
- **Land.Slope** Slope of property.
- **Neighborhood** Physical locations within Ames city limits (map available).
- **Condition.1** Proximity to various conditions.
- **Condition.2** Proximity to various conditions (if more than one is present).
- **Bldg.Type** Type of dwelling.
- **House.Style** Style of dwelling.
- **Overall.Qual** Rates the overall material and finish of the house.
- **Overall.Cond** Rates the overall condition of the house.
- **Year.Built** Original construction date.
Year.Remod.Add  Remodel date (same as construction date if no remodeling or additions).
Roof.Style  Type of roof.
Roof.Matl  Roof material.
Exterior.1st  Exterior covering on house.
Exterior.2nd  Exterior covering on house (if more than one material).
Mas.Vnr.Type  Masonry veneer type.
Mas.Vnr.Area  Masonry veneer area in square feet.
Exter.Qual  Evaluates the quality of the material on the exterior.
Exter.Cond  Evaluates the present condition of the material on the exterior.
Foundation  Type of foundation.
Bsmt.Qual  Evaluates the height of the basement.
Bsmt.Cond  Evaluates the general condition of the basement.
Bsmt.Exposure  Refers to walkout or garden level walls.
BsmtFin.Type.1  Rating of basement finished area.
BsmtFin.SF.1  Type 1 finished square feet.
BsmtFin.Type.2  Rating of basement finished area (if multiple types).
BsmtFin.SF.2  Type 2 finished square feet.
Bsmt.Unf.SF  Unfinished square feet of basement area.
Total Bsmt.SF  Total square feet of basement area.
Heating  Type of heating.
Heating.QC  Heating quality and condition.
Central.Air  Central air conditioning.
Electrical  Electrical system.
X1st.Flr.SF  First Floor square feet.
X2nd.Flr.SF  Second floor square feet.
Low.Qual.Fin.SF  Low quality finished square feet (all floors).
Full.Bath  Full bathrooms above grade.
Half.Bath  Half baths above grade.
Bedroom.AbvGr  Bedrooms above grade (does NOT include basement bedrooms).
Kitchen.AbvGr  Kitchens above grade.
TotRms.AbvGrd  Total rooms above grade (does not include bathrooms).
Functional  Home functionality (Assume typical unless deductions are warranted).
Fireplaces  Number of fireplaces.
Fireplace.Qu  Fireplace quality.
Garage.Type  Garage location.
Garage.Yr.Blt  Year garage was built.
Garage.Finish  Interior finish of the garage.
Garage.Cars  Size of garage in car capacity.
Garage.Area  Size of garage in square feet.
Garage.Qual  Garage quality.
Garage.Cond  Garage condition.
Paved.Drive  Paved driveway.
Wood.Deck.SF  Wood deck area in square feet.
Open.Porch.SF  Open porch area in square feet.
Enclosed.Porch  Enclosed porch area in square feet.
X3Ssn.Porch  Three season porch area in square feet.
Screen.Porch  Screen porch area in square feet.
Pool.Area  Pool area in square feet.
Pool.QC  Pool quality.
Fence  Fence quality.
Misc.Feature  Miscellaneous feature not covered in other categories.
Misc.Val  Dollar value of miscellaneous feature.
Mo.Sold  Month Sold (MM).
Yr.Sold  Year Sold (YYYY).
Sale.Type  Type of sale.
Sale.Condition  Condition of sale.

Source


ames_sampling_dist  Run the ames sampling distribution shiny app

Description

Run the ames sampling distribution shiny app

Usage

ames_sampling_dist()
**Description**  
Arbuthnot’s data describes male and female christenings (births) for London from 1629-1710.

**Usage**  
arbuthnot

**Format**  
A tbl_df with with 82 rows and 3 variables:
- **year**: year, ranging from 1629 to 1710
- **boys**: number of male christenings (births)
- **girls**: number of female christenings (births)

**Details**  
John Arbuthnot (1710) used these time series data to carry out the first known significance test. During every one of the 82 years, there were more male christenings than female christenings. As Arbuthnot wondered, we might also wonder if this could be due to chance, or whether it meant the birth ratio was not actually 1:1.

**Source**  
These data are excerpted from the Arbuthnot data set in the HistData package.

---

**Description**  
Survey results on atheism across several countries and years. Each row represents a single respondent.

**Usage**  
atheism
Format
A tbl_df with 88032 rows and 3 variables:

- **nationality** Country of the individual surveyed.
- **response** A categorical variable with two levels: atheist and non-atheist.
- **year** Year in which the person was surveyed.

Source
WIN-Gallup International Press Release

---

Description
Utility function for calculating the posterior probability of each machine being "good" in a two-armed bandit problem. Calculated result is based on observed win loss data, prior belief about which machine is good and the probability of the good and bad machine paying out.

Usage
```r
bandit_posterior(
  data,
  prior = c(m1_good = 0.5, m2_good = 0.5),
  win_probs = c(good = 1/2, bad = 1/3)
)
```

Arguments
- **data** data frame containing win loss data
- **prior** prior vector containing the probabilities of Machine 1 and Machine 2 being good, defaults to 0.5 and 0.5 respectively.
- **win_probs** vector containing the probabilities of winning on the good and bad machine respectively.

Value
A vector containing the posterior probability of Machine 1 and Machine 2 being the good machine.

Examples
```r
data = data.frame(machine = c(1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L),
bandit_posterior(data)
```
bandit_sim

Run the Bandit Simulation shiny app

Description

Run the Bandit Simulation shiny app

Usage

bandit_sim()

bayes_inference

Bayesian hypothesis tests and credible intervals

Description

Bayesian hypothesis tests and credible intervals

Usage

bayes_inference(
  y,
  x = NULL,
  data,
  type = c("ci", "ht"),
  statistic = c("mean", "proportion"),
  success = NULL,
  null = NULL,
  cred_level = 0.95,
  alternative = c("twosided", "less", "greater"),
  hypothesis_prior = c(H1 = 0.5, H2 = 0.5),
  n_0 = 1,
  beta_prior = NULL,
  beta_prior1 = NULL,
  beta_prior2 = NULL,
  verbose = TRUE,
  show_summ = verbose,
  show_res = verbose,
  show_plot = verbose
)
Arguments

y  Response variable, can be numerical or categorical
x  Explanatory variable, categorical (optional)
data  Name of data frame that y and x are in
type  of inference; "ci" (credible interval) or "ht" (hypothesis test)
statistic  population parameter to estimate: mean or proportion
success  which level of the categorical variable to call "success", i.e. do inference on
null  null value for the hypothesis test
cred_level  confidence level, value between 0 and 1
alternative  direction of the alternative hypothesis; "less","greater", or "twosided"
hypothesis_prior  discrete prior for H1 and H2, default is the uniform prior: c(H1=0.5,H2=0.5)
n_0  Prior sample size for calculating the Bayes factor of the twosided test of one mean
beta_prior, beta_prior1, beta_prior2  beta priors for p (or p_1 and p_2) for one or two proportion inference
verbose  whether output should be verbose or not, default is TRUE
show_summ  print summary stats, set to verbose by default
show_res  print results, set to verbose by default
show_plot  print inference plot, set to verbose by default

Value

Results of inference task performed

brfss

Behavioral Risk Factor Surveillance System 2013 (Subset)

Description

This data set is a small subset of BRFSS results from the 2013 survey, each row represents an individual respondent.

Usage

brfss
Format

A tbl_df with with 5000 rows and 6 variables:

- **weight**  Weight in pounds.
- **height**  Height in inches.
- **sex**  Sex
- **exercise**  Any exercise in the last 30 days
- **fruit_per_day**  Number of servings of fruit consumed per day.
- **vege_per_day**  Number of servings of dark green vegetables consumed per day.

Source


---

calc_streak  
*Calculate hit streaks.*

Description

Calculate hit streaks.

Usage

calc_streak(x)

Arguments

- **x**  A data frame or character vector of hits ("H") and misses ("M").

Value

A data frame with one column, `length`, containing the length of each hit streak.

Examples

data(kobe_basket)
calc_streak(kobe_basket$shot)
credible_interval_app

**Run the Credible Interval shiny app**

**Description**
Run the Credible Interval shiny app

**Usage**
credible_interval_app()

**evals**
*Teachers evaluations at the University of Texas at Austin*

**Description**
The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin (variables beginning with cls). In addition, six students rated the professors’ physical appearance (variables beginning with bty). (This is a slightly modified version of the original data set that was released as part of the replication data for Data Analysis Using Regression and Multilevel/Hierarchical Models (Gelman and Hill, 2007).

**Usage**
evals

**Format**
A data frame with 463 rows and 21 variables:

- **score**: Average professor evaluation score: (1) very unsatisfactory - (5) excellent
- **rank**: Rank of professor: teaching, tenure track, tenure
- **ethnicity**: Ethnicity of professor: not minority, minority
- **gender**: Gender of professor: female, male
- **language**: Language of school where professor received education: english or non-english
- **age**: Age of professor
- **cls_perc_eval**: Percent of students in class who completed evaluation
- **cls_did_eval**: Number of students in class who completed evaluation
- **cls_students**: Total number of students in class
- **cls_level**: Class level: lower, upper
- **cls_profs**: Number of professors teaching sections in course in sample: single, multiple
- **cls_credits**: Number of credits of class: one credit (lab, PE, etc.), multi credit
**bty_f1lower**  Beauty rating of professor from lower level female: (1) lowest - (10) highest

**bty_f1upper**  Beauty rating of professor from upper level female: (1) lowest - (10) highest

**bty_m1lower**  Beauty rating of professor from lower level male: (1) lowest - (10) highest

**bty_m1upper**  Beauty rating of professor from upper level male: (1) lowest - (10) highest

**bty_m2upper**  Beauty rating of professor from second upper level male: (1) lowest - (10) highest

**bty_avg**  Average beauty rating of professor

**pic_outfit**  Outfit of professor in picture: not formal, formal

**pic_color**  Color of professor's picture: color, black & white

**Source**


---

### inference

**Hypothesis tests and confidence intervals**

#### Description

Hypothesis tests and confidence intervals

#### Usage

```r
inference(
  y,
  x = NULL,
  data,
  type = c("ci", "ht"),
  statistic = c("mean", "median", "proportion"),
  success = NULL,
  order = NULL,
  method = c("theoretical", "simulation"),
  null = NULL,
  alternative = c("less", "greater", "twosided"),
  sig_level = 0.05,
  conf_level = 0.95,
  boot_method = c("perc", "se"),
  nsim = 15000,
  seed = NULL,
  verbose = TRUE,
  show_var_types = verbose,
  show_summ_stats = verbose,
  show_eda_plot = verbose,
  show_inf_plot = verbose,
  show_res = verbose
)
```
**Arguments**

- **y**: Response variable, can be numerical or categorical
- **x**: Explanatory variable, categorical (optional)
- **data**: Name of data frame that y and x are in
- **type**: Method of inference; "ci" (confidence interval) or "ht" (hypothesis test)
- **statistic**: Parameter to estimate: mean, median, or proportion
- **success**: Which level of the categorical variable to call "success", i.e. do inference on
- **order**: When x is given, order of levels of x in which to subtract parameters
- **method**: Method of inference; "theoretical" (CLT based) or "simulation" (randomization/bootstrap)
- **null**: Null value for a hypothesis test
- **alternative**: Direction of the alternative hypothesis; "less", "greater", or "twosided"
- **sig_level**: Significance level, value between 0 and 1 (used only for ANOVA to determine if posttests are necessary)
- **conf_level**: Confidence level, value between 0 and 1
- **boot_method**: Bootstrap method; "perc" (percentile) or "se" (standard error)
- **nsim**: Number of simulations
- **seed**: Seed to be set, default is NULL
- **verbose**: Whether output should be verbose or not, default is TRUE
- **show_var_types**: Print variable types, set to verbose by default
- **show_summ_stats**: Print summary stats, set to verbose by default
- **show_eda_plot**: Print EDA plot, set to verbose by default
- **show_inf_plot**: Print inference plot, set to verbose by default
- **show_res**: Print results, set to verbose by default

**Value**

Results of inference task performed

---

**kobe_basket**  
*Kobe Bryant basketball performance*

**Description**

Data from the five games the Los Angeles Lakers played against the Orlando Magic in the 2009 NBA finals.

**Usage**

`kobe_basket`
Format

A data frame with 133 rows and 6 variables:

vs A categorical vector, ORL if the Los Angeles Lakers played against Orlando
game A numerical vector, game in the 2009 NBA finals
quarter A categorical vector, quarter in the game, OT stands for overtime
time A character vector, time at which Kobe took a shot
description A character vector, description of the shot
shot A categorical vector, H if the shot was a hit, M if the shot was a miss

Details

Each row represents a shot Kobe Bryant took during the five games of the 2009 NBA finals. Kobe Bryant’s performance earned him the title of Most Valuable Player and many spectators commented on how he appeared to show a hot hand.

mlb11

Major League Baseball team data

Description

Data from all 30 Major League Baseball teams from the 2011 season.

Usage

mlb11

Format

A data frame with 30 rows and 12 variables:

team Team name.
runs Number of runs.
at_bats Number of at bats.
hits Number of hits.
homeruns Number of home runs.
bat_avg Batting average.
strikeouts Number of strikeouts.
stolen_bases Number of stolen bases.
wins Number of wins.
new_onbase Newer variable: on-base percentage, a measure of how often a batter reaches base for any reason other than a fielding error, fielder’s choice, dropped/uncaught third strike, fielder’s obstruction, or catcher’s interference.
new_slug Newer variable: slugging percentage, popular measure of the power of a hitter calculated as the total bases divided by at bats.
new_obs Newer variable: on-base plus slugging, calculated as the sum of the on-base and slugging percentages.
Description

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

Usage

nc

Format

A tbl_df with 1000 rows and 13 variables:

- fage: father’s age in years
- mage: mother’s age in years
- mature: maturity status of mother
- weeks: length of pregnancy in weeks
- premie: whether the birth was classified as premature (premie) or full-term
- visits: number of hospital visits during pregnancy
- marital: whether mother is ‘married’ or ‘not married’ at birth
- gained: weight gained by mother during pregnancy in pounds
- weight: weight of the baby at birth in pounds
- lowbirthweight: whether baby was classified as low birthweight (‘low’) or not (‘not low’)
- gender: gender of the baby, ‘female’ or ‘male’
- habit: status of the mother as a ‘nonsmoker’ or a ‘smoker’
- whitemom: whether mom is ‘white’ or ‘not white’

Source

State of North Carolina.
Description

On-time data for a random sample of flights that departed NYC (i.e. JFK, LGA or EWR) in 2013.

Usage

nycflights

Format

A tbl_df with 32,735 rows and 16 variables:

- `year,month,day` Date of departure
- `dep_time,arr_time` Departure and arrival times, local tz.
- `dep_delay,arr_delay` Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
- `hour,minute` Time of departure broken in to hour and minutes
- `carrier` Two letter carrier abbreviation. See airlines in the nycflights13 package for more information
- `tailnum` Plane tail number
- `flight` Flight number
- `origin,dest` Origin and destination. See airports in the nycflights13 package for more information, or google airport the code.
- `air_time` Amount of time spent in the air
- `distance` Distance flown

Source

plot_bandit_posterior

Description
Generates a plot that shows the bandit posterior values as they are sequentially updated by the provided win / loss data.

Usage
plot_bandit_posterior(
data,
prior = c(m1_good = 0.5, m2_good = 0.5),
win_probs = c(good = 1/2, bad = 1/3)
)

Arguments
- data: data frame containing win loss data
- prior: prior vector containing the probabilities of Machine 1 and Machine 2 being good, defaults to 50-50.
- win_probs: vector containing the probabilities of winning on the good and bad machine respectively.

Examples

```r
data = data.frame(machine = c(1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L),
plot_bandit_posterior(data)
```

plot_ss

Description
An interactive function that will generate a scatterplot of two variables, then allow the user to click the plot in two locations to draw a best fit line. Residuals are drawn by default; boxes representing the squared residuals are optional.

Usage
plot_ss(x, y, data, showSquares = FALSE, leastSquares = FALSE)
Arguments

- **x**  the name of numerical vector 1
- **y**  the name of numerical vector 2
- **data**  the dataframe in which x and y can be found
- **showSquares**  logical option to show boxes representing the squared residuals
- **leastSquares**  logical option to bypass point entry and automatically draw the least squares line

---

**Male and female births in the US**

Description

Counts of the total number of male and female births in the United States from 1940 to 2013.

Usage

```r
present
```

Format

A tbl_df with 74 rows and 3 variables:

- **year**  year, ranging from 1940 to 2013
- **boys**  number of male births
- **girls**  number of female births

Source

Description

Repeating sampling.

Usage

rep_sample_n(tbl, size, replace = FALSE, reps = 1)

Arguments

tbl       tbl of data.
size      The number of rows to select.
replace   Sample with or without replacement?
reps      The number of samples to collect.

Value

A tbl_df that aggregates all created samples, with the addition of a `replicate` column that the tbl_df is also grouped by

---

statsr

`statsr`: A companion package for the Coursera Statistics with R specialization

Description

See https://github.com/StatsWithR/statsr for more information.
Description

The data were gathered as part of a random sample of 935 respondents throughout the United States.

Usage

wage

Format

A tbl_df with 935 rows and 17 variables:

- wage  weekly earnings (dollars)
- hours  average hours worked per week
- iq     IQ score
- kww    Knowledge of world work score
- educ   years of education
- exper  years of work experience
- tenure years with current employer
- age    age in years
- married =1 if married
- black  =1 if black
- south  =1 if live in south
- urban  =1 if live in a Standard Metropolitan Statistical Area
- sibs   number of siblings
- brthord birth order
- meduc  mother’s education (years)
- feduc  father’s education (years)
- lwage  natural log of wage

Source

Index

*Topic datasets
  ames, 3
  arbuthnot, 6
  atheism, 6
  brfss, 9
  evals, 11
  kobe_basket, 13
  mlb11, 14
  nc, 15
  nycflights, 16
  present, 18
  wage, 20

allow_shiny, 2
ames, 3
ames_sampling_dist, 5
Arbuthnot, 6
arbuthnot, 6
atheism, 6

bandit_posterior, 7
bandit_sim, 8
bayes_inference, 8
brfss, 9

calc_streak, 10
credible_interval_app, 11

evals, 11
inference, 12
kobe_basket, 13
mlb11, 14
nc, 15
nycflights, 16

plot_bandit_posterior, 17
plot_ss, 17

present, 18
rep_sample_n, 19
statsr, 19
wage, 20