

# Package ‘statsr’

July 23, 2025

**Type** Package

**Title** Companion Software for the Coursera Statistics with R  
Specialization

**Version** 0.3.0

**Date** 2021-01-21

**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** 7.1.1

**Encoding** UTF-8

**Depends** R (>= 3.3.0), BayesFactor

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

**Suggests** spelling, HistData, testthat (>= 3.0.0)

**URL** <https://github.com/StatsWithR/statsr>

**BugReports** <https://github.com/StatsWithR/statsr/issues>

**Language** en-US

**Config/testthat/edition** 3

**NeedsCompilation** no

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

**Repository** CRAN

**Date/Publication** 2021-01-22 20:40:03 UTC

Contents

|                                 |           |
|---------------------------------|-----------|
| ames . . . . .                  | 2         |
| ames_sampling_dist . . . . .    | 5         |
| arbuthnot . . . . .             | 6         |
| atheism . . . . .               | 6         |
| bandit_posterior . . . . .      | 7         |
| bandit_sim . . . . .            | 8         |
| bayes_inference . . . . .       | 9         |
| BF_app . . . . .                | 12        |
| brfss . . . . .                 | 12        |
| calc_streak . . . . .           | 13        |
| credible_interval_app . . . . . | 13        |
| evals . . . . .                 | 14        |
| inference . . . . .             | 15        |
| kobe_basket . . . . .           | 17        |
| mlb11 . . . . .                 | 17        |
| nc . . . . .                    | 18        |
| nycflights . . . . .            | 19        |
| plot_bandit_posterior . . . . . | 20        |
| plot_ss . . . . .               | 21        |
| present . . . . .               | 21        |
| rep_sample_n . . . . .          | 22        |
| statsr . . . . .                | 23        |
| tapwater . . . . .              | 23        |
| wage . . . . .                  | 24        |
| zinc . . . . .                  | 25        |
| <b>Index</b>                    | <b>26</b> |

---

|      |                                     |
|------|-------------------------------------|
| ames | <i>Housing prices in Ames, Iowa</i> |
|------|-------------------------------------|

---

Description

Data set contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. See <http://www.amstat.org/publications/jse/v19n3/d> for detailed variable descriptions.

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).

**Bsmt.Full.Bath** Basement full bathrooms.

**Bsmt.Half.Bath** Basement half bathrooms.

**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.

**Kitchen.Qual** Kitchen quality.

**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

De Cock, Dean. "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project." Journal of Statistics Education 19.3 (2011).

---

|                    |                                       |
|--------------------|---------------------------------------|
| ames_sampling_dist | <i>Simulate Sampling Distribution</i> |
|--------------------|---------------------------------------|

---

### Description

Run the interactive ames sampling distribution shiny app to illustrate sampling distributions using variables from the ‘ames’ dataset.

### Usage

```
ames_sampling_dist()
```

### Examples

```
if (interactive()) {  
  ames_sampling_dist()  
}
```

---

|           |                                         |
|-----------|-----------------------------------------|
| arbuthnot | <i>Male and female births in London</i> |
|-----------|-----------------------------------------|

---

### Description

Arbuthnot's data describes male and female christenings (births) for London from 1629-1710.

### Usage

```
arbuthnot
```

### Format

A `tbl_df` 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.

---

|         |                                  |
|---------|----------------------------------|
| atheism | <i>Atheism in the world data</i> |
|---------|----------------------------------|

---

### 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

---

|                  |                         |
|------------------|-------------------------|
| bandit_posterior | <i>bandit posterior</i> |
|------------------|-------------------------|

---

**Description**

Utility function for calculating the posterior probability of each machine being "good" in 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**

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

**Arguments**

|                        |                                                                                                                        |
|------------------------|------------------------------------------------------------------------------------------------------------------------|
| <code>data</code>      | data frame containing win loss data                                                                                    |
| <code>prior</code>     | prior vector containing the probabilities of Machine 1 and Machine 2 being good, defaults to 0.5 and 0.5 respectively. |
| <code>win_probs</code> | 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.

**See Also**

[bandit\\_sim](#) to generate data and [plot\\_bandit\\_posterior](#) to visualize.





```
bandit_posterior(data)
plot_bandit_posterior(data)
```

---

|                 |                                                         |
|-----------------|---------------------------------------------------------|
| bayes_inference | <i>Bayesian hypothesis tests and credible intervals</i> |
|-----------------|---------------------------------------------------------|

---

## Description

Bayesian hypothesis tests and credible intervals

## Usage

```
bayes_inference(
  y,
  x = NULL,
  data,
  type = c("ci", "ht"),
  statistic = c("mean", "proportion"),
  method = c("theoretical", "simulation"),
  success = NULL,
  null = NULL,
  cred_level = 0.95,
  alternative = c("twosided", "less", "greater"),
  hypothesis_prior = c(H1 = 0.5, H2 = 0.5),
  prior_family = "JZS",
  n_0 = 1,
  mu_0 = null,
  s_0 = 0,
  v_0 = -1,
  rscale = 1,
  beta_prior = NULL,
  beta_prior1 = NULL,
  beta_prior2 = NULL,
  nsim = 10000,
  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                                                                                                                                                                |
| method                               | of inference; "theoretical" (quantile based) or "simulation"                                                                                                                                                        |
| 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: $c(H1=0.5, H2=0.5)$                                                                                                                                     |
| prior_family                         | character string representing default priors for inference or testing ("JSZ", "JUI", "ref"). See notes for details.                                                                                                 |
| n_0                                  | $n_0$ is the prior sample size in the Normal prior for the mean                                                                                                                                                     |
| mu_0                                 | the prior mean in one sample mean problems or the prior difference in two sample problems. For hypothesis testing, this is all the null value if null is not supplied.                                              |
| s_0                                  | the prior standard deviation of the data for the conjugate Gamma prior on $1/\sigma^2$                                                                                                                              |
| v_0                                  | prior degrees of freedom for conjugate Gamma prior on $1/\sigma^2$                                                                                                                                                  |
| rscale                               | is the scaling parameter in the Cauchy prior: $1/n_0 \sim \text{Gamma}(1/2, \text{rscale}^2/2)$ leads to $\mu_0$ having a $\text{Cauchy}(0, \text{rscale}^2 * \sigma^2)$ prior distribution for prior_family="JZS". |
| beta_prior, beta_prior1, beta_prior2 | beta priors for p (or p_1 and p_2) for one or two proportion inference                                                                                                                                              |
| nsim                                 | number of Monte Carlo draws; default is 10,000                                                                                                                                                                      |
| 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.

**Note**

For inference and testing for normal means several default options are available. "JZS" corresponds to using the Jeffreys reference prior on  $\sigma^2$ ,  $p(\sigma^2) = 1/\sigma^2$ , and the Zellner-Siow Cauchy prior on the standardized effect size  $\mu/\sigma$  or  $(\mu_1 - \mu_2)/\sigma$  with a location of  $\mu_0$  and scale rscale. The "JUI" option also uses the Jeffreys reference prior on  $\sigma^2$ , but the Unit Information prior on the standardized effect,  $N(\mu_0, 1)$ . The option "ref" uses the improper uniform prior on the standardized effect and the Jeffreys reference prior on  $\sigma^2$ . The latter cannot be used for hypothesis testing due to the ill-determination of Bayes factors. Finally "NG" corresponds to the conjugate Normal-Gamma prior.

**References**

<https://statswithr.github.io/book/>

**Examples**

```

# inference for the mean from a single normal population using
# Jeffreys Reference prior,  $p(\mu, \sigma^2) = 1/\sigma^2$ 

library(BayesFactor)
data(tapwater)

# Calculate 95% CI using quantiles from Student t derived from ref prior
bayes_inference(tthm, data=tapwater,
  statistic="mean",
  type="ci", prior_family="ref",
  method="theoretical")

# Calculate 95% CI using simulation from Student t using an informative mean and ref
# prior for  $\sigma^2$ 

bayes_inference(tthm, data=tapwater,
  statistic="mean", mu_0=9.8,
  type="ci", prior_family="JUI",
  method="theo")

# Calculate 95% CI using simulation with the
# Cauchy prior on  $\mu$  and reference prior on  $\sigma^2$ 

bayes_inference(tthm, data=tapwater,
  statistic="mean", mu_0 = 9.8, rscale=sqrt(2)/2,
  type="ci", prior_family="JZS",
  method="simulation")

# Bayesian t-test  $\mu = 0$  with ZJS prior
bayes_inference(tthm, data=tapwater,
  statistic="mean",
  type="ht", alternative="twosided", null=80,
  prior_family="JZS",
  method="sim")

# Bayesian t-test for two means

data(chickwts)
chickwts = chickwts[chickwts$feed %in% c("horsebean","linseed"),]
# Drop unused factor levels
chickwts$feed = factor(chickwts$feed)
bayes_inference(y=weight, x=feed, data=chickwts,
  statistic="mean", mu_0 = 0, alt="twosided",
  type="ht", prior_family="JZS",
  method="simulation")

```

---

 BF\_app

---

*Run the interactive Bayes Factor shiny app*


---

### Description

This app illustrates how changing the Z score and prior precision affects the Bayes Factor for testing H1 that the mean is zero versus H2 that the mean is not zero for data arising from a normal population. Lindley's paradox occurs for large sample sizes when the Bayes factor favors H1 even though the Z score is large or the p-value is small enough to reach statistical significance and the values of the sample mean do not reflex practical significance based on the prior distribution. Bartlett's paradox may occur when the prior precision goes to zero, leading to Bayes factors that favor H1 regardless of the data. A prior precision of one corresponds to the unit information prior.

### Usage

```
BF_app()
```

### Examples

```
if (interactive()) {
  BF_app()
}
```

---

 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**

Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2013.

---

|             |                                  |
|-------------|----------------------------------|
| calc_streak | <i>Calculate hitting streaks</i> |
|-------------|----------------------------------|

---

**Description**

Calculate hitting 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 | <i>Credible Interval shiny app</i> |
|-----------------------|------------------------------------|

---

**Description**

Run the 'shiny' credible interval app to generate credible intervals under the prior or posterior distribution for Beta, Gamma and Gaussian families. Sliders are used to adjust the hyperparameters in the distribution so that one may see how the resulting credible intervals and plotted distributions change.

**Usage**

```
credible_interval_app()
```

**Examples**

```
if (interactive()) {
  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\_f2upper** Beauty rating of professor from second 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**

These data appear in Hamermesh DS, and Parker A. 2005. Beauty in the classroom: instructors pulchritude and putative pedagogical productivity. *Economics of Education Review* 24(4):369-376.

inference

*Hypothesis tests and confidence intervals***Description**

Hypothesis tests and confidence intervals

**Usage**

```
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      | 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          | 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

## Examples

```
data(tapwater)

# Calculate 95% CI using quantiles using a Student t distribution
inference(tthm, data=tapwater,
           statistic="mean",
           type="ci",
           method="theoretical")

inference(tthm, data=tapwater,
           statistic="mean",
           type="ci",
           boot_method = "perc",
           method="simulation")

# Inference for a proportion
# Calculate 95% confidence intervals for the proportion of atheists

data("atheism")
library("dplyr")
us12 <- atheism %>%
  filter(nationality == "United States" , atheism$year == "2012")
inference(y = response, data = us12, statistic = "proportion",
          type = "ci",
```



```
method = "theoretical",
success = "atheist")
```

---

|             |                                           |
|-------------|-------------------------------------------|
| kobe_basket | <i>Kobe Bryant basketball performance</i> |
|-------------|-------------------------------------------|

---

### 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 | <i>Major League Baseball team data</i> |
|-------|----------------------------------------|

---

### 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.

**Source**

[mlb.com](#)

---

|    |                              |
|----|------------------------------|
| nc | <i>North Carolina births</i> |
|----|------------------------------|

---

**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.

---

nycflights

*Flights data*

---

**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

Hadley Wickham (2014). nycflights13: Data about flights departing NYC in 2013. R package version 0.1. <https://CRAN.R-project.org/package=nycflights13>

---

plot\_bandit\_posterior *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.            |

## See Also

[bandit\\_sim](#) to generate data to use below

## Examples

```
# capture data from the `shiny` app `bandit_sim`.
data = data.frame(machine = c(1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L),
  outcome = c("W", "L", "W", "L", "L", "W", "L", "L", "L", "W"))
plot_bandit_posterior(data)
```

---

plot\_ss

plot\_ss

---

### Description

An interactive shiny app that will generate a scatterplot of two variables, then allow the user to click the plot in two locations to draw a best fitting 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 on x-axis                                           |
| y            | the name of numerical vector 2 on y-axis                                           |
| 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 |

### Examples

```
## Not run: plot_ss
```

---

present

*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

```
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**

Data up to 2002 appear in Mathews TJ, and Hamilton BE. 2005. Trend analysis of the sex ratio at birth in the United States. National Vital Statistics Reports 53(20):1-17. Data for 2003 - 2013 have been collected from annual National Vital Statistics Reports published by the US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

---

|              |                                         |
|--------------|-----------------------------------------|
| rep_sample_n | <i>Repeating Sampling from a Tibble</i> |
|--------------|-----------------------------------------|

---

**Description**

Repeating Sampling from a Tibble

**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

**Examples**

```
data(nc)
rep_sample_n(nc, size=10, replace=FALSE, reps=1)
```

statsr

*statsr: A companion package for Statistics with R*

## Description

R package to support the online open access book "An Introduction to Bayesian Thinking" available at <https://statswithr.github.io/book/> and videos for the Coursera "Statistics with R" Specialization. The package includes data sets, functions and Shiny Applications for learning frequentist and Bayesian statistics with R. The two main functions for inference and decision making are 'inference' and 'bayes\_inference' which support confidence/credible intervals and hypothesis testing with one sample or two samples from Gaussian and Bernoulli populations. Shiny apps are used to illustrate how prior hyperparameters or changes in the data may influence posterior distributions.

## Details

See <https://github.com/StatsWithR/statsr> for the development version and additional information or for additional background and illustrations of functions the online book <https://statswithr.github.io/book/>.

tapwater

*Total Trihalomethanes in Tapwater*

## Description

Trihalomethanes are formed as a by-product predominantly when chlorine is used to disinfect water for drinking. They result from the reaction of chlorine or bromine with organic matter present in the water being treated. THMs have been associated through epidemiological studies with some adverse health effects and many are considered carcinogenic. In the United States, the EPA limits the total concentration of the four chief constituents (chloroform, bromoform, bromodichloromethane, and dibromochloromethane), referred to as total trihalomethanes (TTHM), to 80 parts per billion in treated water.

## Usage

tapwater

## Format

A dataframe with 28 rows and 6 variables:

**date** Date of collection

**tthm** average total trihalomethanes in ppb

**samples** number of samples

**nondetects** number of samples where tthm not detected (0)

**min** min tthm in ppb in samples

**max** max tthm in ppb in samples

**Source**

National Drinking Water Database for Durham, NC. <https://www.ewg.org>

---

|      |                  |
|------|------------------|
| wage | <i>Wage data</i> |
|------|------------------|

---

**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**

Jeffrey M. Wooldridge (2000). Introductory Econometrics: A Modern Approach. South-Western College Publishing.



---

zinc

*Zinc Concentration in Water*

---

**Description**

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water.

**Usage**

zinc

**Format**

A data frame with 10 observations on the following 4 variables.

location sample number

bottom zinc concentration in bottom water

surface zinc concentration in surface water

difference difference between zinc concentration at the bottom and surface

**Source**

[PennState Eberly College of Science Online Courses](#)

**Examples**

```
data(zinc)
str(zinc)
plot(bottom ~ surface, data=zinc)
# use paired t-test to test if difference in means is zero
```

# Index

## \* datasets

- ames, [2](#)
- arbuthnot, [6](#)
- atheism, [6](#)
- brfss, [12](#)
- evals, [14](#)
- kobe\_basket, [17](#)
- mlb11, [17](#)
- nc, [18](#)
- nycflights, [19](#)
- present, [21](#)
- tapwater, [23](#)
- wage, [24](#)
- zinc, [25](#)

- ames, [2](#)
- ames\_sampling\_dist, [5](#)
- Arbuthnot, [6](#)
- arbuthnot, [6](#)
- atheism, [6](#)

- bandit\_posterior, [7](#), [8](#)
- bandit\_sim, [7](#), [8](#), [20](#)
- bayes\_inference, [9](#)
- BF\_app, [12](#)
- brfss, [12](#)

- calc\_streak, [13](#)
- credible\_interval\_app, [13](#)

- evals, [14](#)

- inference, [15](#)

- kobe\_basket, [17](#)

- mlb11, [17](#)

- nc, [18](#)
- nycflights, [19](#)

- plot\_bandit\_posterior, [7](#), [8](#), [20](#)

- plot\_ss, [21](#)
- present, [21](#)
- rep\_sample\_n, [22](#)
- statsr, [23](#)
- tapwater, [23](#)
- wage, [24](#)
- zinc, [25](#)