

# Package ‘bayesrules’

October 12, 2022

**Type** Package

**Title** Datasets and Supplemental Functions from Bayes Rules! Book

**Version** 0.0.2

**Description** Provides datasets and functions used for analysis and visualizations in the Bayes Rules! book (<<https://www.bayesrulesbook.com>>). The package contains a set of functions that summarize and plot Bayesian models from some conjugate families and another set of functions for evaluation of some Bayesian models.

**License** GPL (>= 3)

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.1.2

**Suggests** knitr, rmarkdown

**Imports** ggplot2, janitor, magrittr, dplyr, stats, purrr, rstanarm, e1071, groupdata2

**Depends** R (>= 2.10)

**URL** <https://bayes-rules.github.io/bayesrules/docs/>,  
<https://github.com/bayes-rules/bayesrules/>

**BugReports** <https://github.com/bayes-rules/bayesrules/issues>

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Mine Dogucu [aut, cre] (<<https://orcid.org/0000-0002-8007-934X>>),  
Alicia Johnson [aut],  
Miles Ott [aut] (<<https://orcid.org/0000-0003-4457-6565>>)

**Maintainer** Mine Dogucu <[mdogucu@gmail.com](mailto:mdogucu@gmail.com)>

**Repository** CRAN

**Date/Publication** 2021-09-25 04:30:07 UTC

**R topics documented:**

|   |    |
|---|----|
| airbnb . . . . .                          | 3  |
| airbnb_small . . . . .                    | 4  |
| bald_eagles . . . . .                     | 5  |
| basketball . . . . .                      | 5  |
| bechdel . . . . .                         | 7  |
| big_word_club . . . . .                   | 7  |
| bikes . . . . .                           | 9  |
| bike_users . . . . .                      | 10 |
| bird_counts . . . . .                     | 11 |
| book_banning . . . . .                    | 11 |
| cherry_blossom_sample . . . . .           | 13 |
| classification_summary . . . . .          | 13 |
| classification_summary_cv . . . . .       | 14 |
| climbers_sub . . . . .                    | 15 |
| coffee_ratings . . . . .                  | 16 |
| coffee_ratings_small . . . . .            | 17 |
| equality_index . . . . .                  | 18 |
| fake_news . . . . .                       | 19 |
| football . . . . .                        | 20 |
| hotel_bookings . . . . .                  | 21 |
| loons . . . . .                           | 22 |
| moma . . . . .                            | 23 |
| moma_sample . . . . .                     | 24 |
| naive_classification_summary . . . . .    | 24 |
| naive_classification_summary_cv . . . . . | 25 |
| penguins_bayes . . . . .                  | 26 |
| plot_beta . . . . .                       | 27 |
| plot_beta_binomial . . . . .              | 27 |
| plot_beta_ci . . . . .                    | 28 |
| plot_binomial_likelihood . . . . .        | 29 |
| plot_gamma . . . . .                      | 29 |
| plot_gamma_poisson . . . . .              | 30 |
| plot_normal . . . . .                     | 31 |
| plot_normal_likelihood . . . . .          | 31 |
| plot_normal_normal . . . . .              | 32 |
| plot_poisson_likelihood . . . . .         | 33 |
| pop_vs_soda . . . . .                     | 33 |
| prediction_summary . . . . .              | 34 |
| prediction_summary_cv . . . . .           | 35 |
| pulse_of_the_nation . . . . .             | 36 |
| sample_mode . . . . .                     | 37 |
| spotify . . . . .                         | 37 |
| summarize_beta . . . . .                  | 38 |
| summarize_beta_binomial . . . . .         | 39 |
| summarize_gamma . . . . .                 | 40 |
| summarize_gamma_poisson . . . . .         | 40 |

|                                   |           |
|-----------------------------------|-----------|
| <i>airbnb</i>                     | 3         |
| summarize_normal_normal . . . . . | 41        |
| voices . . . . .                  | 42        |
| weather_australia . . . . .       | 42        |
| weather_perth . . . . .           | 44        |
| weather_WU . . . . .              | 45        |
| <b>Index</b>                      | <b>46</b> |

---

|        |                            |
|--------|----------------------------|
| airbnb | <i>Chicago AirBnB Data</i> |
|--------|----------------------------|

---

### Description

The AirBnB data was collated by Trinh and Ameri as part of a course project at St Olaf College, and distributed with "Broadening Your Statistical Horizons" by Legler and Roback. This data set includes the prices and features for 1561 AirBnB listings in Chicago, collected in 2016.

### Usage

airbnb

### Format

A data frame with 1561 rows and 12 variables. Each row represents a single AirBnB listing.

- price** the nightly price of the listing (in USD)
- rating** the listing's average rating, on a scale from 1 to 5
- reviews** number of user reviews the listing has
- room\_type** the type of listing (eg: Shared room)
- accommodates** number of guests the listing accommodates
- bedrooms** the number of bedrooms the listing has
- minimum\_stay** the minimum number of nights to stay in the listing
- neighborhood** the neighborhood in which the listing is located
- district** the broader district in which the listing is located
- walk\_score** the neighborhood's rating for walkability (0 - 100)
- transit\_score** the neighborhood's rating for access to public transit (0 - 100)
- bike\_score** the neighborhood's rating for bikeability (0 - 100)

### Source

Ly Trinh and Pony Ameri (2018). Airbnb Price Determinants: A Multilevel Modeling Approach. Project for Statistics 316-Advanced Statistical Modeling, St. Olaf College. Julie Legler and Paul Roback (2019). Broadening Your Statistical Horizons: Generalized Linear Models and Multilevel Models. <https://bookdown.org/roback/bookdown-bysh/>. <https://github.com/proback/BeyondMLR/blob/master/data/airbnb.csv/>

---

`airbnb_small`*Chicago AirBnB Data*

---

### Description

The AirBnB data was collated by Trinh and Ameri as part of a course project at St Olaf College, and distributed with "Broadening Your Statistical Horizons" by Legler and Roback. This data set, a subset of the airbnb data in the bayesrules package, includes the prices and features for 869 AirBnB listings in Chicago, collected in 2016.

### Usage

`airbnb_small`

### Format

A data frame with 869 rows and 12 variables. Each row represents a single AirBnB listing.

**price** the nightly price of the listing (in USD)

**rating** the listing's average rating, on a scale from 1 to 5

**reviews** number of user reviews the listing has

**room\_type** the type of listing (eg: Shared room)

**accommodates** number of guests the listing accommodates

**bedrooms** the number of bedrooms the listing has

**minimum\_stay** the minimum number of nights to stay in the listing

**neighborhood** the neighborhood in which the listing is located

**district** the broader district in which the listing is located

**walk\_score** the neighborhood's rating for walkability (0 - 100)

**transit\_score** the neighborhood's rating for access to public transit (0 - 100)

**bike\_score** the neighborhood's rating for bikeability (0 - 100)

### Source

Ly Trinh and Pony Ameri (2018). Airbnb Price Determinants: A Multilevel Modeling Approach. Project for Statistics 316-Advanced Statistical Modeling, St. Olaf College. Julie Legler and Paul Roback (2019). Broadening Your Statistical Horizons: Generalized Linear Models and Multilevel Models. <https://bookdown.org/roback/bookdown-bysh/>. <https://github.com/proback/BeyondMLR/blob/master/data/airbnb.csv/>

---

`bald_eagles`*Bald Eagle Count Data*

---

**Description**

Bald Eagle count data collected from the year 1981 to 2017, in late December, by birdwatchers in the Ontario, Canada area. The data was made available by the Bird Studies Canada website and distributed through the R for Data Science TidyTuesday project. A more complete data set with a larger selection of birds can be found in the `bird_counts` data in the `bayesrules` package.

**Usage**`bald_eagles`**Format**

A data frame with 37 rows and 5 variables. Each row represents Bald Eagle observations in the given year.

**year** year of data collection

**count** number of birds observed

**hours** total person-hours of observation period

**count\_per\_hour** count divided by hours

**count\_per\_week** `count_per_hour` multiplied by 168 hours per week

**Source**

[https://raw.githubusercontent.com/rfordatascience/tidyuesday/master/data/2019/2019-06-18/bird\\_counts.csv](https://raw.githubusercontent.com/rfordatascience/tidyuesday/master/data/2019/2019-06-18/bird_counts.csv).

---

`basketball`*WNBA Basketball Data*

---

**Description**

The WNBA Basketball Data was scraped from <https://www.basketball-reference.com/wnba/players/> and contains information on basketball players from the 2019 season.

**Usage**`basketball`

**Format**

A data frame with 146 rows and 30 variables. Each row represents a single WNBA basketball player. The variables on each player are as follows.

**player\_name** first and last name  
**height** height in inches  
**weight** weight in pounds  
**year** year of the WNBA season  
**team** team that the WNBA player is a member of  
**age** age in years  
**games\_played** number of games played by the player in that season  
**games\_started** number of games the player started in that season  
**avg\_minutes\_played** average number of minutes played per game  
**avg\_field\_goals** average number of field goals per game played  
**avg\_field\_goal\_attempts** average number of field goals attempted per game played  
**field\_goal\_pct** percent of field goals made throughout the season  
**avg\_three\_pointers** average number of three pointers per game played  
**avg\_three\_pointer\_attempts** average number of three pointers attempted per game played  
**three\_pointer\_pct** percent of three pointers made throughout the season  
**avg\_two\_pointers** average number of two pointers made per game played  
**avg\_two\_pointer\_attempts** average number of two pointers attempted per game played  
**two\_pointer\_pct** percent of two pointers made throughout the season  
**avg\_free\_throws** average number of free throws made per game played  
**avg\_free\_throw\_attempts** average number of free throws attempted per game played  
**free\_throw\_pct** percent of free throws made throughout the season  
**avg\_offensive\_rb** average number of offensive rebounds per game played  
**avg\_defensive\_rb** average number of defensive rebounds per game played  
**avg\_rb** average number of rebounds (both offensive and defensive) per game played  
**avg\_assists** average number of assists per game played  
**avg\_steals** average number of steals per game played  
**avg\_blocks** average number of blocks per game played  
**avg\_turnovers** average number of turnovers per game played  
**avg\_personal\_fouls** average number of personal fouls per game played. Note: after 5 fouls the player is not allowed to play in that game anymore  
**avg\_points** average number of points made per game played  
**total\_minutes** total number of minutes played throughout the season  
**starter** whether or not the player started in more than half of the games they played

**Source**

<https://www.basketball-reference.com/>

---

|         |  |
|---------|--|
| bechdel | <i>Bechdel Test for over 1500 movies</i> |
|---------|--|

---

**Description**

A dataset containing data behind the story "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" <https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-e>

**Usage**

```
bechdel
```

**Format**

A data frame with 1794 rows and 3 variables:

**year** The release year of the movie

**title** The title of the movie

**binary** Bechdel test result (PASS, FAIL)

**Source**

```
<https://github.com/fivethirtyeight/data/tree/master/bechdel/>
```

---

|               |                            |
|---------------|----------------------------|
| big_word_club | <i>Big Word Club (BWC)</i> |
|---------------|----------------------------|

---

**Description**

Data on the effectiveness of a digital learning program designed by the Abdul Latif Jameel Poverty Action Lab (J-PAL) to address disparities in vocabulary levels among children from households with different income levels.

**Usage**

```
big_word_club
```

**Format**

A data frame with 818 student-level observations and 31 variables:

**participant\_id** unique student id

**treat** control group (0) or treatment group (1)

**age\_months** age in months

**female** whether student identifies as female

**kindergarten** grade level, pre-school (0) or kindergarten (1)  
**teacher\_id** unique teacher id  
**school\_id** unique school id  
**private\_school** whether school is private  
**title1** whether school has Title 1 status  
**free\_reduced\_lunch** percent of school that receive free / reduced lunch  
**state** school location  
**esl\_observed** whether student has ESL status  
**special\_ed\_observed** whether student has special education status  
**new\_student** whether student enrolled after program began  
**distracted\_a1** student's distraction level during assessment 1 (0 = not distracted; 1 = mildly distracted; 2 = moderately distracted; 3 = extremely distracted)  
**distracted\_a2** same as distracted\_a1 but during assessment 2  
**distracted\_ppvt** same as distracted\_a1 but during standardized assessment  
**score\_a1** student score on BWC assessment 1  
**invalid\_a1** whether student's score on assessment 1 was invalid  
**score\_a2** student score on BWC assessment 2  
**invalid\_a2** whether student's score on assessment 2 was invalid  
**score\_ppvt** student score on standardized assessment  
**score\_ppvt\_age** score\_ppvt adjusted for age  
**invalid\_ppvt** whether student's score on standardized assessment was invalid  
**t\_logins\_april** number of teacher logins onto BWC system in April  
**t\_logins\_total** number of teacher logins onto BWC system during entire study  
**t\_weeks\_used** number of weeks of the BWC program that the classroom has completed  
**t\_words\_learned** teacher response to the number of words students had learned through BWC (0 = almost none; 1 = 1 to 5; 2 = 6 to 10)  
**t\_financial\_struggle** teacher response to the number of their students that have families that experience financial struggle  
**t\_misbehavior** teacher response to frequency that student misbehavior interferes with teaching (0 = never; 1 = rarely; 2 = occasionally; 3 = frequently)  
**t\_years\_experience** teacher's number of years of teaching experience  
**score\_pct\_change** percent change in scores before and after the program

### Source

These data correspond to the following study: Ariel Kalil, Susan Mayer, Philip Oreopoulos (2020). Closing the word gap with Big Word Club: Evaluating the Impact of a Tech-Based Early Childhood Vocabulary Program. Data was obtained through the was obtained through the Inter-university Consortium for Political and Social Research (ICPSR) <https://www.openicpsr.org/openicpsr/project/117330/version/V1/view/>.



---

bikes

*Capital Bikeshare Bike Ridership*

---

### Description

Data on ridership among registered members of the Capital Bikeshare service in Washington, D.C..

### Usage

bikes

### Format

A data frame with 500 daily observations and 13 variables:

**date** date of observation

**season** fall, spring, summer, or winter

**year** the year of the date

**month** the month of the date

**day\_of\_week** the day of the week

**weekend** whether or not the date falls on a weekend (TRUE or FALSE)

**holiday** whether or not the date falls on a holiday (yes or no)

**temp\_actual** raw temperature (degrees Fahrenheit)

**temp\_feel** what the temperature feels like (degrees Fahrenheit)

**humidity** humidity level (percentage)

**windspeed** wind speed (miles per hour)

**weather\_cat** weather category (categ1 = pleasant, categ2 = moderate, categ3 = severe)

**rides** number of bikeshare rides

### Source

Fanaee-T, Hadi and Gama, Joao (2013). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence. <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

---

bike\_users

*Capital Bikeshare Bike Ridership (Registered and Casual Riders)*

---

### Description

Data on ridership among registered members and casual users of the Capital Bikeshare service in Washington, D.C..

### Usage

bike\_users

### Format

A data frame with 534 daily observations, 267 each for registered riders and casual riders, and 13 variables:

**date** date of observation

**season** fall, spring, summer, or winter

**year** the year of the date

**month** the month of the date

**day\_of\_week** the day of the week

**weekend** whether or not the date falls on a weekend (TRUE or FALSE)

**holiday** whether or not the date falls on a holiday (yes or no)

**temp\_actual** raw temperature (degrees Fahrenheit)

**temp\_feel** what the temperature feels like (degrees Fahrenheit)

**humidity** humidity level (percentage)

**windspeed** wind speed (miles per hour)

**weather\_cat** weather category (categ1 = pleasant, categ2 = moderate, categ3 = severe)

**user** rider type (casual or registered)

**rides** number of bikeshare rides

### Source

Fanaee-T, Hadi and Gama, Joao (2013). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence. <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset/>

---

`bird_counts`*Bird Counts Data*

---

**Description**

Bird count data collected between the years 1921 and 2017, in late December, by birdwatchers in the Ontario, Canada area. The data was made available by the Bird Studies Canada website and distributed through the R for Data Science TidyTuesday project.

**Usage**`bird_counts`**Format**

A data frame with 18706 rows and 7 variables. Each row represents observations for the given bird species in the given year.

**year** year of data collection

**species** scientific name of observed bird species

**species\_latin** latin name of observed bird species

**count** number of birds observed

**hours** total person-hours of observation period

**count\_per\_hour** count divided by hours

**count\_per\_week** count\_per\_hour multiplied by 168 hours per week

**Source**

[https://github.com/rfordatascience/tidytuesday/blob/master/data/2019/2019-06-18/bird\\_counts.csv/](https://github.com/rfordatascience/tidytuesday/blob/master/data/2019/2019-06-18/bird_counts.csv/).

---

`book_banning`*Book Banning Data*

---

**Description**

The book banning data was collected by Fast and Hegland as part of a course project at St Olaf College, and distributed with "Broadening Your Statistical Horizons" by Legler and Roback. This data set includes the features and outcomes for 931 book challenges (ie. requests to ban a book) made in the US between 2000 and 2010. Information on the books being challenged and the characteristics of these books were collected from the American Library Society. State-level demographic information and political leanings were obtained from the US Census Bureau and Cook Political Report, respectively. Due to an outlying large number of challenges, book challenges made in the state of Texas were omitted.

**Usage**

book\_banning

**Format**

A data frame with 931 rows and 17 variables. Each row represents a single book challenge within the given state and date.

**title** title of book being challenged

**book\_id** identifier for the book

**author** author of the book

**date** date of the challenge

**year** year of the challenge

**removed** whether or not the challenge was successful (the book was removed)

**explicit** whether the book was challenged for sexually explicit material

**antifamily** whether the book was challenged for anti-family material

**occult** whether the book was challenged for occult material

**language** whether the book was challenged for inappropriate language

**lgbtq** whether the book was challenged for LGBTQ material

**violent** whether the book was challenged for violent material

**state** US state in which the challenge was made

**political\_value\_index** Political Value Index of the state (negative = leans Republican, 0 = neutral, positive = leans Democrat)

**median\_income** median income in the state, relative to the average state median income

**hs\_grad\_rate** high school graduation rate, in percent, relative to the average state high school graduation rate

**college\_grad\_rate** college graduation rate, in percent, relative to the average state college graduation rate

**Source**

Shannon Fast and Thomas Hegland (2011). Book Challenges: A Statistical Examination. Project for Statistics 316-Advanced Statistical Modeling, St. Olaf College. Julie Legler and Paul Roback (2019). Broadening Your Statistical Horizons: Generalized Linear Models and Multilevel Models. <https://bookdown.org/roback/bookdown-bysh/>. <https://github.com/proback/BeyondMLR/blob/master/data/bookbanningNoTex.csv>

---

cherry\_blossom\_sample *Cherry Blossom Running Race*

---

**Description**

A sub-sample of outcomes for the annual Cherry Blossom Ten Mile race in Washington, D.C.. This sub-sample was taken from the complete Cherry data in the mdsr package.

**Usage**

```
cherry_blossom_sample
```

**Format**

A data frame with 252 Cherry Blossom outcomes and 7 variables:

**runner** a unique identifier for the runner

**age** age of the runner

**net** time to complete the race, from starting line to finish line (minutes)

**gun** time between the official start of the of race and the finish line (minutes)

**year** year of the race

**previous** the number of previous years in which the subject ran in the race

**Source**

Data in the original Cherry data set were obtained from <https://www.cherryblossom.org/post-race/race-results/>.

---

classification\_summary

*Posterior Classification Summaries*

---

**Description**

Given a set of observed data including a binary response variable  $y$  and an rstanreg model of  $y$ , this function returns summaries of the model's posterior classification quality. These summaries include a confusion matrix as well as estimates of the model's sensitivity, specificity, and overall accuracy.

**Usage**

```
classification_summary(model, data, cutoff = 0.5)
```

**Arguments**

|        |   |
|--------|---|
| model  | an rstanreg model object with binary y  |
| data   | data frame including the variables in the model, both response y and predictors x |
| cutoff | probability cutoff to classify a new case as positive (0.5 is the default)        |

**Value**

a list

**Examples**

```
x <- rnorm(20)
z <- 3*x
prob <- 1/(1+exp(-z))
y <- rbinom(20, 1, prob)
example_data <- data.frame(x = x, y = y)
example_model <- rstanarm::stan_glm(y ~ x, data = example_data, family = binomial)
classification_summary(model = example_model, data = example_data, cutoff = 0.5)
```

---

classification\_summary\_cv

*Cross-Validated Posterior Classification Summaries*

---

**Description**

Given a set of observed data including a binary response variable y and an rstanreg model of y, this function returns cross validated estimates of the model's posterior classification quality: sensitivity, specificity, and overall accuracy. For hierarchical models of class lmerMod, the folds are comprised by collections of groups, not individual observations.

**Usage**

```
classification_summary_cv(model, data, group, k, cutoff = 0.5)
```

**Arguments**

|        |   |
|--------|---|
| model  | an rstanreg model object with binary y  |
| data   | data frame including the variables in the model, both response y (0 or 1) and predictors x                                      |
| group  | a character string representing the name of the factor grouping variable, ie. random effect (only used for hierarchical models) |
| k      | the number of folds to use for cross validation   |
| cutoff | probability cutoff to classify a new case as positive   |

**Value**

a list

**Examples**

```
x <- rnorm(20)
z <- 3*x
prob <- 1/(1+exp(-z))
y <- rbinom(20, 1, prob)
example_data <- data.frame(x = x, y = y)
example_model <- rstanarm::stan_glm(y ~ x, data = example_data, family = binomial)
classification_summary_cv(model = example_model, data = example_data, k = 2, cutoff = 0.5)
```

---

climbers\_sub

*Himalayan Climber Data*

---

**Description**

A sub-sample of the Himalayan Database distributed through the R for Data Science TidyTuesday project. This dataset includes information on the results and conditions for various Himalayan climbing expeditions. Each row corresponds to a single member of a climbing expedition team.

**Usage**

```
climbers_sub
```

**Format**

A data frame with 2076 observations (1 per climber) and 22 variables:

**expedition\_id** unique expedition identifier

**member\_id** unique climber identifier

**peak\_id** unique identifier of the expedition's destination peak

**peak\_name** name of the expedition's destination peak

**year** year of expedition

**season** season of expedition (Autumn, Spring, Summer, Winter)

**sex** climber gender identity which the database oversimplifies to a binary category

**age** climber age

**citizenship** climber citizenship

**expedition\_role** climber's role in the expedition (eg: Co-Leader)

**hired** whether the climber was a hired member of the expedition

**highpoint\_metres** the destination peak's highpoint (metres)

**success** whether the climber successfully reached the destination

**solo** whether the climber was on a solo expedition

**oxygen\_used** whether the climber utilized supplemental oxygen  
**died** whether the climber died during the expedition  
**death\_cause**  
**death\_height\_metres**  
**injured** whether the climber was injured on the expedition  
**injury\_type**  
**injury\_height\_metres**  
**count** number of climbers in the expedition  
**height\_metres** height of the peak in meters  
**first\_ascent\_year** the year of the first recorded summit of the peak (though not necessarily the actual first summit!)

### Source

Original source: <https://www.himalayandatabase.com/>. Complete dataset distributed by: <https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-09-22/>.

---

coffee\_ratings

*Coffee Ratings Data*

---

### Description

A sub-set of data on coffee bean ratings / quality originally collected by James LeDoux (jnzledoux) and distributed through the R for Data Science TidyTuesday project.

### Usage

coffee\_ratings

### Format

A data frame with 1339 batches of coffee beans and 27 variables on each batch.

**owner** farm owner  
**farm\_name** farm where beans were grown  
**country\_of\_origin** country where farm is  
**mill** where beans were processed  
**in\_country\_partner** country of coffee partner  
**altitude\_low\_meters** lowest altitude of the farm  
**altitude\_high\_meters** highest altitude of the farm  
**altitude\_mean\_meters** average altitude of the farm  
**number\_of\_bags** number of bags tested



**bag\_weight** weight of each tested bag  
**species** bean species  
**variety** bean variety  
**processing\_method** how beans were processed  
**aroma** bean aroma grade  
**flavor** bean flavor grade  
**aftertaste** bean aftertaste grade  
**acidity** bean acidity grade  
**body** bean body grade  
**balance** bean balance grade  
**uniformity** bean uniformity grade  
**clean\_cup** bean clean cup grade  
**sweetness** bean sweetness grade  
**moisture** bean moisture grade  
**category\_one\_defects** count of category one defects  
**category\_two\_defects** count of category two defects  
**color** bean color  
**total\_cup\_points** total bean rating (0 – 100)

### Source

[https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-07/coffee\\_ratings.csv](https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-07/coffee_ratings.csv).

---

coffee\_ratings\_small *Coffee Ratings Data*

---

### Description

A sub-set of data on coffee bean ratings / quality originally collected by James LeDoux (jnzledoux) and distributed through the R for Data Science TidyTuesday project. This is a simplified version of the coffee\_ratings data.

### Usage

```
coffee_ratings_small
```

**Format**

A data frame with 636 batches of coffee beans and 11 variables on each batch.

**farm\_name** farm where beans were grown  
**total\_cup\_points** total bean rating (0 – 100)  
**aroma** bean aroma grade  
**flavor** bean flavor grade  
**aftertaste** bean aftertaste grade  
**acidity** bean acidity grade  
**body** bean body grade  
**balance** bean balance grade  
**uniformity** bean uniformity grade  
**sweetness** bean sweetness grade  
**moisture** bean moisture grade

**Source**

[https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-07/coffee\\_ratings.csv](https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-07/coffee_ratings.csv).

---

|                |                                    |
|----------------|------------------------------------|
| equality_index | <i>LGBTQ+ Rights Laws by State</i> |
|----------------|------------------------------------|

---

**Description**

Data on the number of LGBTQ+ equality laws (as of 2019) and demographics in each U.S. state.

**Usage**

equality\_index

**Format**

A data frame with 50 observations, one per state, and 6 variables:

**state** state name  
**region** region in which the state falls  
**gop\_2016** percent of the 2016 presidential election vote earned by the Republican ("GOP") candidate  
**laws** number of LGBTQ+ rights laws (as of 2019)  
**historical** political leaning of the state over time (gop = Republican, dem = Democrat, swing = swing state)  
**percent\_urban** percent of state's residents that live in urban areas (by the 2010 census)

**Source**

Data on LGBTQ+ laws were obtained from Warbelow, Sarah, Courtney Avant, and Colin Kutney (2020). 2019 State Equality Index. Washington, DC. Human Rights Campaign Foundation. [https://assets2.hrc.org/files/assets/resources/HRC-SEI-2019-Report.pdf?\\_ga=2.148925686.1325740687.1594310864-1928808113.1594310864&\\_gac=1.213124768.1594312278.EAIaIQobChMI9dP2hMzA6gIVkcD BwE/](https://assets2.hrc.org/files/assets/resources/HRC-SEI-2019-Report.pdf?_ga=2.148925686.1325740687.1594310864-1928808113.1594310864&_gac=1.213124768.1594312278.EAIaIQobChMI9dP2hMzA6gIVkcD BwE/). Data on urban residency obtained from <https://www.icip.iastate.edu/tables/population/urban-pct-states/>.

fake\_news

*A collection of 150 news articles***Description**

A dataset containing data behind the study "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media" <https://arxiv.org/abs/1809.01286>. The news articles in this dataset were posted to Facebook in September 2016, in the run-up to the U.S. presidential election.

**Usage**

fake\_news

**Format**

A data frame with 150 rows and 6 variables:

**title** The title of the news article

**text** Text of the article

**url** Hyperlink for the article

**authors** Authors of the article

**type** Binary variable indicating whether the article presents fake or real news(fake, real)

**title\_words** Number of words in the title

**text\_words** Number of words in the text

**title\_char** Number of characters in the title

**text\_char** Number of characters in the text

**title\_caps** Number of words that are all capital letters in the title

**text\_caps** Number of words that are all capital letters in the text

**title\_caps\_percent** Percent of words that are all capital letters in the title

**text\_caps\_percent** Percent of words that are all capital letters in the text

**title\_excl** Number of characters that are exclamation marks in the title

**text\_excl** Number of characters that are exclamation marks in the text

**title\_excl\_percent** Percent of characters that are exclamation marks in the title

**text\_excl\_percent** Percent of characters that are exclamation marks in the text

**title\_has\_excl** Binary variable indicating whether the title of the article includes an exclamation point or not(TRUE, FALSE)

**anger** Percent of words that are associated with anger

**anticipation** Percent of words that are associated with anticipation

**disgust** Percent of words that are associated with disgust

**fear** Percent of words that are associated with fear

**joy** Percent of words that are associated with joy

**sadness** Percent of words that are associated with sadness

**surprise** Percent of words that are associated with surprise

**trust** Percent of words that are associated with trust

**negative** Percent of words that have negative sentiment

**positive** Percent of words that have positive sentiment

**text\_syllables** Number of syllables in text

**text\_syllables\_per\_word** Number of syllables per word in text

#### Source

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. (2018) FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media

---

football

*Football Brain Measurements*

---

#### Description

Brain measurements for football and non-football players as provided in the Lock5 package

#### Usage

football

#### Format

A data frame with 75 observations and 5 variables:

**group** control = no football, fb\_no\_concuss = football player but no concussions, fb\_concuss = football player with concussion history

**years** Number of years a person played football

**volume** Total hippocampus volume, in cubic centimeters

#### Source

Singh R, Meier T, Kuplicki R, Savitz J, et al., "Relationship of Collegiate Football Experience and Concussion With Hippocampal Volume and Cognitive Outcome," JAMA, 311(18), 2014

---

|                |                            |
|----------------|----------------------------|
| hotel_bookings | <i>Hotel Bookings Data</i> |
|----------------|----------------------------|

---

### Description

A random subset of the data on hotel bookings originally collected by Antonio, Almeida and Nunes (2019) and distributed through the R for Data Science TidyTuesday project.

### Usage

```
hotel_bookings
```

### Format

A data frame with 1000 hotel bookings and 32 variables on each booking.

**hotel** "Resort Hotel" or "City Hotel"

**is\_canceled** whether the booking was cancelled

**lead\_time** number of days between booking and arrival

**arrival\_date\_year** year of scheduled arrival

**arrival\_date\_month** month of scheduled arrival

**arrival\_date\_week\_number** week of scheduled arrival

**arrival\_date\_day\_of\_month** day of month of scheduled arrival

**stays\_in\_weekend\_nights** number of reserved weekend nights

**stays\_in\_week\_nights** number of reserved week nights

**adults** number of adults in booking

**children** number of children

**babies** number of babies

**meal** whether the booking includes breakfast (BB = bed & breakfast), breakfast and dinner (HB = half board), or breakfast, lunch, and dinner (FB = full board)

**country** guest's country of origin

**market\_segment** market segment designation (eg: TA = travel agent, TO = tour operator)

**distribution\_channel** booking distribution channel (eg: TA = travel agent, TO = tour operator)

**is\_repeated\_guest** whether or not booking was made by a repeated guest

**previous\_cancellations** guest's number of previous booking cancellations

**previous\_bookings\_not\_canceled** guest's number of previous bookings that weren't cancelled

**reserved\_room\_type** code for type of room reserved by guest

**assigned\_room\_type** code for type of room assigned by hotel

**booking\_changes** number of changes made to the booking

**deposit\_type** No Deposit, Non Refund, Refundable

**agent** booking travel agency  
**company** booking company  
**days\_in\_waiting\_list** number of days the guest waited for booking confirmation  
**customer\_type** Contract, Group, Transient, Transient-party (a transient booking tied to another transient booking)  
**average\_daily\_rate** average hotel cost per day  
**required\_car\_parking\_spaces** number of parking spaces the guest needed  
**total\_of\_special\_requests** number of guest special requests  
**reservation\_status** Canceled, Check-Out, No-Show  
**reservation\_status\_date** when the guest cancelled or checked out

### Source

Nuno Antonio, Ana de Almeida, and Luis Nunes (2019). "Hotel booking demand datasets." Data in Brief (22): 41-49. <https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/hotels.csv/>.

---

loons

*Loon Count Data*

---

### Description

Loon count data collected from the year 2000 to 2017, in late December, by birdwatchers in the Ontario, Canada area. The data was made available by the Bird Studies Canada website and distributed through the R for Data Science TidyTuesday project. A more complete data set with a larger selection of birds can be found in the `bird_counts` data in the `bayesrules` package.

### Usage

loons

### Format

A data frame with 18 rows and 5 variables. Each row represents loon observations in the given year.

**year** year of data collection  
**count** number of loons observed  
**hours** total person-hours of observation period  
**count\_per\_hour** count divided by hours  
**count\_per\_100** count\_per\_hour multiplied by 100 hours

### Source

[https://github.com/rfordatascience/tidytuesday/blob/master/data/2019/2019-06-18/bird\\_counts.csv](https://github.com/rfordatascience/tidytuesday/blob/master/data/2019/2019-06-18/bird_counts.csv).

---

moma

*Museum of Modern Art (MoMA) data*

---

### Description

The Museum of Modern Art data includes information about the individual artists included in the collection of the Museum of Modern Art in New York City. It does not include information about works for artist collectives or companies. The data was made available by MoMA itself and downloaded in December 2020.

### Usage

moma

### Format

A data frame with 10964 rows and 11 variables. Each row represents an individual artist in the MoMA collection.

**artist** name

**country** country of origin

**birth** year of birth

**death** year of death

**alive** whether or not the artist was living at the time of data collection (December 2020)

**genx** whether or not the artist is Gen X or younger, ie. born during 1965 or after

**gender** gender identity (as perceived by MoMA employees)

**department** MoMA department in which the artist's works most frequently appear

**count** number of the artist's works in the MoMA collection

**year\_acquired\_min** first year MoMA acquired one of the artist's works

**year\_acquired\_max** most recent year MoMA acquired one of the artist's works

### Source

<https://github.com/MuseumofModernArt/collection/blob/master/Artworks.csv/>.

moma\_sample

*Museum of Modern Art (MoMA) data sample*

---

**Description**

A random sample of 100 artists represented in the Museum of Modern Art in New York City. The data was made available by MoMA itself and downloaded in December 2020. It does not include information about artist collectives or companies.

**Usage**

moma\_sample

**Format**

A data frame with 100 rows and 10 variables. Each row represents an individual artist in the MoMA collection.

**artist** name**country** country of origin**birth** year of birth**death** year of death**alive** whether or not the artist was living at the time of data collection (December 2020)**genx** whether or not the artist is Gen X or younger, ie. born during 1965 or after**gender** gender identity (as perceived by MoMA employees)**count** number of the artist's works in the MoMA collection**year\_acquired\_min** first year MoMA acquired one of the artist's works**year\_acquired\_max** most recent year MoMA acquired one of the artist's works**Source**

<https://github.com/MuseumofModernArt/collection/blob/master/Artworks.csv/>.

---

naive\_classification\_summary

*Posterior Classification Summaries for a Naive Bayes model*

---

**Description**

Given a set of observed data including a categorical response variable  $y$  and a naiveBayes model of  $y$ , this function returns summaries of the model's posterior classification quality. These summaries include a confusion matrix as well as an estimate of the model's overall accuracy.



**Usage**

```
naive_classification_summary(model, data, y)
```

**Arguments**

|       |  |
|-------|--|
| model | a naiveBayes model object with categorical y         |
| data  | data frame including the variables in the model      |
| y     | a character string indicating the y variable in data |

**Value**

a list

**Examples**

```
data(penguins_bayes, package = "bayesrules")
example_model <- e1071::naiveBayes(species ~ bill_length_mm, data = penguins_bayes)
naive_classification_summary(model = example_model, data = penguins_bayes, y = "species")
```

---

naive\_classification\_summary\_cv

*Cross-Validated Posterior Classification Summaries for a Naive Bayes model*

---

**Description**

Given a set of observed data including a categorical response variable y and a naiveBayes model of y, this function returns a cross validated confusion matrix by which to assess the model's posterior classification quality.

**Usage**

```
naive_classification_summary_cv(model, data, y, k = 10)
```

**Arguments**

|       |  |
|-------|--|
| model | a naiveBayes model object with categorical y         |
| data  | data frame including the variables in the model      |
| y     | a character string indicating the y variable in data |
| k     | the number of folds to use for cross validation      |

**Value**

a list

**Examples**

```
data(penguins_bayes, package = "bayesrules")
example_model <- e1071::naiveBayes(species ~ bill_length_mm, data = penguins_bayes)
naive_classification_summary_cv(model = example_model, data = penguins_bayes, y = "species", k = 2)
```

---

penguins\_bayes                      *Penguins Data*

---

**Description**

Data on penguins in the Palmer Archipelago, originally collected by Gordan et al and distributed through the penguins data in the palmerpenguins package. In addition to the original penguins data is a variable `above_average_weight`.

**Usage**

```
penguins_bayes
```

**Format**

A data frame with 344 penguins and 9 variables on each.

**species** species (Adelie, Chinstrap, Gentoo)

**island** home island (Biscoe, Dream, Torgersen)

**year** year of observation

**bill\_length\_mm** length of bill (mm)

**bill\_depth\_mm** depth of bill (mm)

**flipper\_length\_mm** length of flipper (mm)

**body\_mass\_g** body mass (g)

**above\_average\_weight** whether or not the body mass exceeds 4200g (TRUE or FALSE)

**sex** male or female

**Source**

Gorman KB, Williams TD, and Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (Genus *Pygoscelis*). PLoS ONE, 9(3).

---

|           |   |
|-----------|---|
| plot_beta | <i>Plot a Beta Model for <math>\pi</math></i> |
|-----------|---|

---

**Description**

Plots the probability density function (pdf) for a Beta(alpha, beta) model of variable  $\pi$ .

**Usage**

```
plot_beta(alpha, beta, mean = FALSE, mode = FALSE)
```

**Arguments**

|             |   |
|-------------|---|
| alpha, beta | positive shape parameters of the Beta model                           |
| mean, mode  | a logical value indicating whether to display the model mean and mode |

**Value**

A density plot for the Beta model.

**Examples**

```
plot_beta(alpha = 1, beta = 12, mean = TRUE, mode = TRUE)
```

---

|                    |  |
|--------------------|--|
| plot_beta_binomial | <i>Plot a Beta-Binomial Bayesian Model</i> |
|--------------------|--|

---

**Description**

Consider a Beta-Binomial Bayesian model for parameter  $\pi$  with a Beta(alpha, beta) prior on  $\pi$  and Binomial likelihood with n trials and y successes. Given information on the prior (alpha and data) and data (y and n), this function produces a plot of any combination of the corresponding prior pdf, scaled likelihood function, and posterior pdf. All three are included by default.

**Usage**

```
plot_beta_binomial(  
  alpha,  
  beta,  
  y = NULL,  
  n = NULL,  
  prior = TRUE,  
  likelihood = TRUE,  
  posterior = TRUE  
)
```

**Arguments**

|             |  |
|-------------|--|
| alpha, beta | positive shape parameters of the prior Beta model                          |
| y           | observed number of successes   |
| n           | observed number of trials  |
| prior       | a logical value indicating whether the prior model should be plotted       |
| likelihood  | a logical value indicating whether the scaled likelihood should be plotted |
| posterior   | a logical value indicating whether posterior model should be plotted       |

**Value**

a ggplot

**Examples**

```
plot_beta_binomial(alpha = 1, beta = 13, y = 25, n = 50)
plot_beta_binomial(alpha = 1, beta = 13, y = 25, n = 50, posterior = FALSE)
```

---

plot\_beta\_ci

*Plot a Beta Model with Credible Interval*

---

**Description**

Plots the probability density function (pdf) for a Beta(alpha, beta) model of variable  $\pi$  with markings indicating a credible interval for  $\pi$ .

**Usage**

```
plot_beta_ci(alpha, beta, ci_level = 0.95)
```

**Arguments**

|             |   |
|-------------|---|
| alpha, beta | positive shape parameters of the Beta model |
| ci_level    | credible interval level                     |

**Value**

A density plot for the Beta model

**Examples**

```
plot_beta_ci(alpha = 7, beta = 12, ci_level = 0.80)
```

---

`plot_binomial_likelihood`*Plot a Binomial Likelihood Function*

---

**Description**

Plots the Binomial likelihood function for variable  $\pi$  given  $y$  observed successes in a series of  $n$  Binomial trials.

**Usage**

```
plot_binomial_likelihood(y, n, mle = FALSE)
```

**Arguments**

|                  |   |
|------------------|---|
| <code>y</code>   | number of successes   |
| <code>n</code>   | number of trials  |
| <code>mle</code> | a logical value indicating whether maximum likelihood estimate of $\pi$ , $y/n$ , should be plotted |

**Value**

a ggplot

**Examples**

```
plot_binomial_likelihood(y = 3, n = 10, mle = TRUE)
```

---

`plot_gamma`*Plot a Gamma Model for  $\lambda$* 

---

**Description**

Plots the probability density function (pdf) for a Gamma(shape, rate) model of variable  $\lambda$ .

**Usage**

```
plot_gamma(shape, rate, mean = FALSE, mode = FALSE)
```

**Arguments**

|                         |   |
|-------------------------|---|
| <code>shape</code>      | non-negative shape parameter of the Gamma model                       |
| <code>rate</code>       | non-negative rate parameter of the Gamma model                        |
| <code>mean, mode</code> | a logical value indicating whether to display the model mean and mode |

**Value**

A density plot for the Gamma model.

**Examples**

```
plot_gamma(shape = 2, rate = 11, mean = TRUE, mode = TRUE)
```

---

plot\_gamma\_poisson      *Plot a Gamma-Poisson Bayesian Model*

---

**Description**

Consider a Gamma-Poisson Bayesian model for rate parameter  $\lambda$  with a Gamma(shape, rate) prior on  $\lambda$  and a Poisson likelihood for the data. Given information on the prior (shape and rate) and data (the sample size  $n$  and  $\text{sum}_y$ ), this function produces a plot of any combination of the corresponding prior pdf, scaled likelihood function, and posterior pdf. All three are included by default.

**Usage**

```
plot_gamma_poisson(
  shape,
  rate,
  sum_y = NULL,
  n = NULL,
  prior = TRUE,
  likelihood = TRUE,
  posterior = TRUE
)
```

**Arguments**

|            |   |
|------------|---|
| shape      | non-negative shape parameter of the Gamma prior                             |
| rate       | non-negative rate parameter of the Gamma prior                              |
| sum_y      | sum of observed data values for the Poisson likelihood                      |
| n          | number of observations for the Poisson likelihood                           |
| prior      | a logical value indicating whether the prior model should be plotted.       |
| likelihood | a logical value indicating whether the scaled likelihood should be plotted. |
| posterior  | a logical value indicating whether posterior model should be plotted.       |

**Value**

a ggplot

**Examples**

```
plot_gamma_poisson(shape = 100, rate = 20, sum_y = 39, n = 6)
plot_gamma_poisson(shape = 100, rate = 20, sum_y = 39, n = 6, posterior = FALSE)
```

---

|             |   |
|-------------|---|
| plot_normal | <i>Plot a Normal Model for <math>\mu</math></i> |
|-------------|---|

---

**Description**

Plots the probability density function (pdf) for a Normal(mean, sd<sup>2</sup>) model of variable  $\mu$ .

**Usage**

```
plot_normal(mean, sd)
```

**Arguments**

|      |  |
|------|--|
| mean | mean parameter of the Normal model               |
| sd   | standard deviation parameter of the Normal model |

**Value**

a ggplot

**Examples**

```
plot_normal(mean = 3.5, sd = 0.5)
```

---

|                        |  |
|------------------------|--|
| plot_normal_likelihood | <i>Plot a Normal Likelihood Function</i> |
|------------------------|--|

---

**Description**

Plots the Normal likelihood function for variable  $\mu$  given a vector of Normal data  $y$ .

**Usage**

```
plot_normal_likelihood(y, sigma = NULL)
```

**Arguments**

|       |   |
|-------|---|
| y     | vector of observed data   |
| sigma | optional value for assumed standard deviation of $y$ . by default, this is calculated by the sample standard deviation of $y$ . |

**Value**

a ggplot of Normal likelihood

**Examples**

```
plot_normal_likelihood(y = rnorm(50, mean = 10, sd = 2), sigma = 1.5)
```

---

```
plot_normal_normal      Plot a Normal-Normal Bayesian model
```

---

**Description**

Consider a Normal-Normal Bayesian model for mean parameter  $\mu$  with a  $N(\text{mean}, \text{sd}^2)$  prior on  $\mu$  and a Normal likelihood for the data. Given information on the prior (mean and sd) and data (the sample size  $n$ , mean  $y_{\text{bar}}$ , and standard deviation  $\text{sigma}$ ), this function produces a plot of any combination of the corresponding prior pdf, scaled likelihood function, and posterior pdf. All three are included by default.

**Usage**

```
plot_normal_normal(  
  mean,  
  sd,  
  sigma = NULL,  
  y_bar = NULL,  
  n = NULL,  
  prior = TRUE,  
  likelihood = TRUE,  
  posterior = TRUE  
)
```

**Arguments**

|            |  |
|------------|--|
| mean       | mean of the Normal prior   |
| sd         | standard deviation of the Normal prior                                     |
| sigma      | standard deviation of the data, or likelihood standard deviation           |
| y_bar      | sample mean of the data  |
| n          | sample size of the data  |
| prior      | a logical value indicating whether the prior model should be plotted       |
| likelihood | a logical value indicating whether the scaled likelihood should be plotted |
| posterior  | a logical value indicating whether posterior model should be plotted       |

**Value**

a ggplot

**Examples**

```
plot_normal_normal(mean = 0, sd = 3, sigma= 4, y_bar = 5, n = 3)  
plot_normal_normal(mean = 0, sd = 3, sigma= 4, y_bar = 5, n = 3, posterior = FALSE)
```



---

`plot_poisson_likelihood`*Plot a Poisson Likelihood Function*

---

**Description**

Plots the Poisson likelihood function for variable  $\lambda$  given a vector of Poisson counts  $y$ .

**Usage**

```
plot_poisson_likelihood(y, lambda_upper_bound = 10)
```

**Arguments**

`y` vector of observed Poisson counts  
`lambda_upper_bound` upper bound for lambda values to display on x-axis

**Value**

a ggplot of Poisson likelihood

**Examples**

```
plot_poisson_likelihood(y = c(4, 2, 7), lambda_upper_bound = 10)
```

---

`pop_vs_soda`*Pop vs Soda vs Coke*

---

**Description**

Results of a volunteer survey on how people around the U.S. refer to fizzy cola drinks. The options are "pop", "soda", "coke", or "other".

**Usage**

```
pop_vs_soda
```

**Format**

A data frame with 374250 observations, one per survey respondent, and 4 variables:

**state** the U.S. state in which the respondent resides

**region** region in which the state falls (as defined by the U.S. Census)

**word\_for\_cola** how the respondent refers to fizzy cola drinks

**pop** whether or not the respondent refers to fizzy cola drinks as "pop"

**Source**

The survey responses were obtained at <https://popvssoda.com/> which is maintained by Alan McConchie.

---

prediction\_summary      *Posterior Predictive Summaries*

---

**Description**

Given a set of observed data including a quantitative response variable  $y$  and an `rstanreg` model of  $y$ , this function returns 4 measures of the posterior prediction quality. Median absolute prediction error (`mae`) measures the typical difference between the observed  $y$  values and their posterior predictive medians (`stable = TRUE`) or means (`stable = FALSE`). Scaled `mae` (`mae_scaled`) measures the typical number of absolute deviations (`stable = TRUE`) or standard deviations (`stable = FALSE`) that observed  $y$  values fall from their predictive medians (`stable = TRUE`) or means (`stable = FALSE`). `within_50` and `within_90` report the proportion of observed  $y$  values that fall within their posterior prediction intervals, the probability levels of which are set by the user.

**Usage**

```
prediction_summary(
  model,
  data,
  prob_inner = 0.5,
  prob_outer = 0.95,
  stable = FALSE
)
```

**Arguments**

|                         |  |
|-------------------------|--|
| <code>model</code>      | an <code>rstanreg</code> model object with quantitative $y$  |
| <code>data</code>       | data frame including the variables in the model, both response $y$ and predictors $x$  |
| <code>prob_inner</code> | posterior predictive interval probability (a value between 0 and 1)  |
| <code>prob_outer</code> | posterior predictive interval probability (a value between 0 and 1)  |
| <code>stable</code>     | <code>TRUE</code> returns the number of absolute deviations and <code>FALSE</code> returns the standard deviations that observed $y$ values fall from their predictive medians |

**Value**

a tibble

**Examples**

```
example_data <- data.frame(x = sample(1:100, 20))
example_data$y <- example_data$x*3 + rnorm(20, 0, 5)
example_model <- rstanarm::stan_glm(y ~ x, data = example_data)
prediction_summary(example_model, example_data, prob_inner = 0.6, prob_outer = 0.80, stable = TRUE)
```

## Description

Given a set of observed data including a quantitative response variable  $y$  and an `rstanreg` model of  $y$ , this function returns 4 cross-validated measures of the model's posterior prediction quality: Median absolute prediction error (`mae`) measures the typical difference between the observed  $y$  values and their posterior predictive medians (`stable = TRUE`) or means (`stable = FALSE`). Scaled `mae` (`mae_scaled`) measures the typical number of absolute deviations (`stable = TRUE`) or standard deviations (`stable = FALSE`) that observed  $y$  values fall from their predictive medians (`stable = TRUE`) or means (`stable = FALSE`). `within_50` and `within_90` report the proportion of observed  $y$  values that fall within their posterior prediction intervals, the probability levels of which are set by the user. For hierarchical models of class `lmerMod`, the folds are comprised by collections of groups, not individual observations.

## Usage

```
prediction_summary_cv(  
  data,  
  group,  
  model,  
  k,  
  prob_inner = 0.5,  
  prob_outer = 0.95  
)
```

## Arguments

|                         |   |
|-------------------------|---|
| <code>data</code>       | data frame including the variables in the model, both response $y$ and predictors $x$   |
| <code>group</code>      | a character string representing the name of the factor grouping variable, ie. random effect (only used for hierarchical models) |
| <code>model</code>      | an <code>rstanreg</code> model object with quantitative $y$   |
| <code>k</code>          | the number of folds to use for cross validation   |
| <code>prob_inner</code> | posterior predictive interval probability (a value between 0 and 1)   |
| <code>prob_outer</code> | posterior predictive interval probability (a value between 0 and 1)   |

## Value

list

**Examples**

```
example_data <- data.frame(x = sample(1:100, 20))
example_data$y <- example_data$x*3 + rnorm(20, 0, 5)
example_model <- rstanarm::stan_glm(y ~ x, data = example_data)
prediction_summary_cv(model = example_model, data = example_data, k = 2)
```

---

pulse\_of\_the\_nation     *Cards Against Humanity's Pulse of the Nation Survey*

---

**Description**

Cards Against Humanity's "Pulse of the Nation" project (<https://thepulseofthenation.com/>) conducted monthly polls into people's social and political views, as well as some silly things. This data includes responses to a subset of questions included in the poll conducted in September 2017.

**Usage**

```
pulse_of_the_nation
```

**Format**

A data frame with observations on 1000 survey respondents with 15 variables:

**income** income in \ \$1000s

**age** age in years

**party** political party affiliation

**trump\_approval** approval level of Donald Trump's job performance

**education** maximum education level completed

**robots** opinion of how likely their job is to be replaced by robots within 10 years

**climate\_change** belief in climate change

**transformers** the number of Transformers film the respondent has seen

**science\_is\_honest** opinion of whether scientists are generally honest and serve the public good

**vaccines\_are\_safe** opinion of whether vaccines are safe and protect children from disease

**books** number of books read in the past year

**ghosts** whether or not they believe in ghosts

**fed\_sci\_budget** respondent's estimate of the percentage of the federal budget that is spent on scientific research

**earth\_sun** belief about whether the earth is always farther away from the sun in winter than in summer (TRUE or FALSE)

**wise\_unwise** whether the respondent would rather be wise but unhappy, or unwise but happy

**Source**

[https://thepulseofthenation.com/downloads/201709-CAH\\_PulseOfTheNation\\_Raw.csv](https://thepulseofthenation.com/downloads/201709-CAH_PulseOfTheNation_Raw.csv)

---

|             |                    |
|-------------|--------------------|
| sample_mode | <i>Sample Mode</i> |
|-------------|--------------------|

---

**Description**

Calculate the sample mode of vector x.

**Usage**

```
sample_mode(x)
```

**Arguments**

x                    vector of sample data

**Value**

sample mode

**Examples**

```
sample_mode(rbeta(100, 2, 7))
```

---

|         |                          |
|---------|--------------------------|
| spotify | <i>Spotify Song Data</i> |
|---------|--------------------------|

---

**Description**

A sub-sample of the Spotify song data originally collected by Kaylin Pavlik (kaylinquest) and distributed through the R for Data Science TidyTuesday project.

**Usage**

```
spotify
```

**Format**

A data frame with 350 songs (or tracks) and 23 variables:

**track\_id** unique song identifier

**title** song name

**artist** song artist

**popularity** song popularity from 0 (low) to 100 (high)

**album\_id** id of the album on which the song appears

**album\_name** name of the album on which the song appears

**album\_release\_date** when the album was released

**playlist\_name** Spotify playlist on which the song appears

**playlist\_id** unique playlist identifier

**genre** genre of the playlist

**subgenre** subgenre of the playlist

**danceability** a score from 0 (not danceable) to 100 (danceable) based on features such as tempo, rhythm, etc.

**energy** a score from 0 (low energy) to 100 (high energy) based on features such as loudness, timbre, entropy, etc.

**key** song key

**loudness** song loudness (dB)

**mode** 0 (minor key) or 1 (major key)

**speechiness** a score from 0 (non-speechy tracks) to 100 (speechy tracks)

**acousticness** a score from 0 (not acoustic) to 100 (very acoustic)

**instrumentalness** a score from 0 (not instrumental) to 100 (very instrumental)

**liveness** a score from 0 (no live audience presence on the song) to 100 (strong live audience presence on the song)

**valence** a score from 0 (the song is more negative, sad, angry) to 100 (the song is more positive, happy, euphoric)

**tempo** song tempo (beats per minute)

**duration\_ms** song duration (ms)

### Source

[https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/spotify\\_songs.csv/](https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/spotify_songs.csv/).

---

summarize\_beta

*Summarize a Beta Model for  $\pi$*

---

### Description

Summarizes the expected value, variance, and mode of a Beta(alpha, beta) model for variable  $\pi$ .

### Usage

```
summarize_beta(alpha, beta)
```

### Arguments

alpha, beta      positive shape parameters of the Beta model

**Value**

a summary table

**Examples**

```
summarize_beta(alpha = 1, beta = 15)
```

---

summarize\_beta\_binomial

*Summarize a Beta-Binomial Bayesian model*

---

**Description**

Consider a Beta-Binomial Bayesian model for parameter  $\pi$  with a Beta(alpha, beta) prior on  $\pi$  and Binomial likelihood with n trials and y successes. Given information on the prior (alpha and data) and data (y and n), this function summarizes the mean, mode, and variance of the prior and posterior Beta models of  $\pi$ .

**Usage**

```
summarize_beta_binomial(alpha, beta, y = NULL, n = NULL)
```

**Arguments**

|             |   |
|-------------|---|
| alpha, beta | positive shape parameters of the prior Beta model |
| y           | number of successes                               |
| n           | number of trials                                  |

**Value**

a summary table

**Examples**

```
summarize_beta_binomial(alpha = 1, beta = 15, y = 25, n = 50)
```

---

|                 |   |
|-----------------|---|
| summarize_gamma | <i>Summarize a Gamma Model for <math>\lambda</math></i> |
|-----------------|---|

---

**Description**

Summarizes the expected value, variance, and mode of a Gamma(shape, rate) model for variable  $\lambda$ .

**Usage**

```
summarize_gamma(shape, rate)
```

**Arguments**

|       |   |
|-------|---|
| shape | positive shape parameter of the Gamma model |
| rate  | positive rate parameter of the Gamma model  |

**Value**

a summary table

**Examples**

```
summarize_gamma(shape = 1, rate = 15)
```

---

|                         |  |
|-------------------------|--|
| summarize_gamma_poisson | <i>Summarize the Gamma-Poisson Model</i> |
|-------------------------|--|

---

**Description**

Consider a Gamma-Poisson Bayesian model for rate parameter  $\lambda$  with a Gamma(shape, rate) prior on  $\lambda$  and a Poisson likelihood for the data. Given information on the prior (shape and rate) and data (the sample size  $n$  and  $\text{sum}_y$ ), this function summarizes the mean, mode, and variance of the prior and posterior Gamma models of  $\lambda$ .

**Usage**

```
summarize_gamma_poisson(shape, rate, sum_y = NULL, n = NULL)
```

**Arguments**

|       |  |
|-------|--|
| shape | positive shape parameter of the Gamma prior            |
| rate  | positive rate parameter of the Gamma prior             |
| sum_y | sum of observed data values for the Poisson likelihood |
| n     | number of observations for the Poisson likelihood      |



**Value**

data frame

**Examples**

```
summarize_gamma_poisson(shape = 3, rate = 4, sum_y = 7, n = 12)
```

---

summarize\_normal\_normal

*Summarize a Normal-Normal Bayesian model*

---

**Description**

Consider a Normal-Normal Bayesian model for mean parameter  $\mu$  with a  $N(\text{mean}, \text{sd}^2)$  prior on  $\mu$  and a Normal likelihood for the data. Given information on the prior (mean and sd) and data (the sample size  $n$ , mean  $y_{\text{bar}}$ , and standard deviation  $\sigma$ ), this function summarizes the mean, mode, and variance of the prior and posterior Normal models of  $\mu$ .

**Usage**

```
summarize_normal_normal(mean, sd, sigma = NULL, y_bar = NULL, n = NULL)
```

**Arguments**

|       |  |
|-------|--|
| mean  | mean of the Normal prior   |
| sd    | standard deviation of the Normal prior                           |
| sigma | standard deviation of the data, or likelihood standard deviation |
| y_bar | sample mean of the data  |
| n     | sample size of the data  |

**Value**

data frame

**Examples**

```
summarize_normal_normal(mean = 2.3, sd = 0.3, sigma = 5.1, y_bar = 128.5, n = 20)
```

---

voices

*Voice Pitch Data*

---

### Description

Voice pitch data collected by Winter and Grawunder (2012). In an experiment, subjects participated in role-playing dialog under various conditions, while researchers monitored voice pitch (Hz). The conditions spanned different scenarios (eg: making an appointment, asking for a favor) and different attitudes to use in the scenario (polite or informal).

### Usage

voices

### Format

A data frame with 84 rows and 4 variables. Each row represents a single observation for the given subject.

**subject** subject identifier

**scenario** context of the dialog (encoded as A, B, ..., G)

**attitude** whether the attitude to use in dialog was polite or informal

**pitch** average voice pitch (Hz)

### Source

Winter, B., & Grawunder, S. (2012). The Phonetic Profile of Korean Formal and Informal Speech Registers. *Journal of Phonetics*, 40, 808-815. [https://bodo-winter.net/data\\_and\\_scripts/POP.csv](https://bodo-winter.net/data_and_scripts/POP.csv). [https://bodo-winter.net/tutorial/bw\\_LME\\_tutorial2.pdf](https://bodo-winter.net/tutorial/bw_LME_tutorial2.pdf).

---

weather\_australia

*Weather Data for 3 Australian Cities*

---

### Description

A sub-sample of daily weather information from the weatherAUS data in the rattle package for three Australian cities: Wollongong, Hobart, and Uluru.

### Usage

weather\_australia

**Format**

A data frame with 300 daily observations and 22 variables from 3 Australian weather stations:

**location** one of three weather stations  
**mintemp** minimum temperature (degrees Celsius)  
**maxtemp** maximum temperature (degrees Celsius)  
**rainfall** rainfall (mm)  
**windgustdir** direction of strongest wind gust  
**windgustspeed** speed of strongest wind gust (km/h)  
**winddir9am** direction of wind gust at 9am  
**winddir3pm** direction of wind gust at 3pm  
**windspeed9am** wind speed at 9am (km/h)  
**windspeed3pm** wind speed at 3pm (km/h)  
**humidity9am** humidity level at 9am (percent)  
**humidity3pm** humidity level at 3pm (percent)  
**pressure9am** atmospheric pressure at 9am (hpa)  
**pressure3pm** atmospheric pressure at 3pm (hpa)  
**temp9am** temperature at 9am (degrees Celsius)  
**temp3pm** temperature at 3pm (degrees Celsius)  
**raintoday** whether or not it rained today (Yes or No)  
**risk\_mm** the amount of rain today (mm)  
**raintomorrow** whether or not it rained the next day (Yes or No)  
**year** the year of the date  
**month** the month of the date  
**day\_of\_year** the day of the year

**Source**

Data in the original weatherAUS data set were obtained from <https://www.bom.gov.au/climate/data/>. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

---

weather\_perth

*Weather Data for Perth, Australia*

---

### Description

A sub-sample of daily weather information on Perth, Australia from the weatherAUS data in the rattle package.

### Usage

weather\_perth

### Format

A data frame with 1000 daily observations and 21 variables:

**mintemp** minimum temperature (degrees Celsius)

**maxtemp** maximum temperature (degrees Celsius)

**rainfall** rainfall (mm)

**windgustdir** direction of strongest wind gust

**windgustspeed** speed of strongest wind gust (km/h)

**winddir9am** direction of wind gust at 9am

**winddir3pm** direction of wind gust at 3pm

**windspeed9am** wind speed at 9am (km/h)

**windspeed3pm** wind speed at 3pm (km/h)

**humidity9am** humidity level at 9am (percent)

**humidity3pm** humidity level at 3pm (percent)

**pressure9am** atmospheric pressure at 9am (hpa)

**pressure3pm** atmospheric pressure at 3pm (hpa)

**temp9am** temperature at 9am (degrees Celsius)

**temp3pm** temperature at 3pm (degrees Celsius)

**raintoday** whether or not it rained today (Yes or No)

**risk\_mm** the amount of rain today (mm)

**raintomorrow** whether or not it rained the next day (Yes or No)

**year** the year of the date

**month** the month of the date

**day\_of\_year** the day of the year

### Source

Data in the original weatherAUS data set were obtained from <https://www.bom.gov.au/climate/data/>. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

---

 weather\_WU

*Weather Data for 2 Australian Cities*


---

### Description

A sub-sample of daily weather information from the weatherAUS data in the rattle package for two Australian cities, Wollongong and Uluru. The weather\_australia data in the bayesrules package combines this data with a third city

### Usage

```
weather_WU
```

### Format

A data frame with 200 daily observations and 22 variables from 2 Australian weather stations:

**location** one of two weather stations  
**mintemp** minimum temperature (degrees Celsius)  
**maxtemp** maximum temperature (degrees Celsius)  
**rainfall** rainfall (mm)  
**windgustdir** direction of strongest wind gust  
**windgustspeed** speed of strongest wind gust (km/h)  
**winddir9am** direction of wind gust at 9am  
**winddir3pm** direction of wind gust at 3pm  
**windspeed9am** wind speed at 9am (km/h)  
**windspeed3pm** wind speed at 3pm (km/h)  
**humidity9am** humidity level at 9am (percent)  
**humidity3pm** humidity level at 3pm (percent)  
**pressure9am** atmospheric pressure at 9am (hpa)  
**pressure3pm** atmospheric pressure at 3pm (hpa)  
**temp9am** temperature at 9am (degrees Celsius)  
**temp3pm** temperature at 3pm (degrees Celsius)  
**raintoday** whether or not it rained today (Yes or No)  
**risk\_mm** the amount of rain today (mm)  
**raintomorrow** whether or not it rained the next day (Yes or No)  
**year** the year of the date  
**month** the month of the date  
**day\_of\_year** the day of the year

### Source

Data in the original weatherAUS data set were obtained from <https://www.bom.gov.au/climate/data>. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

# Index

## \* datasets

- airbnb, 3
  - airbnb\_small, 4
  - bald\_eagles, 5
  - basketball, 5
  - bechdel, 7
  - big\_word\_club, 7
  - bike\_users, 10
  - bikes, 9
  - bird\_counts, 11
  - book\_banning, 11
  - cherry\_blossom\_sample, 13
  - climbers\_sub, 15
  - coffee\_ratings, 16
  - coffee\_ratings\_small, 17
  - equality\_index, 18
  - fake\_news, 19
  - football, 20
  - hotel\_bookings, 21
  - loons, 22
  - moma, 23
  - moma\_sample, 24
  - penguins\_bayes, 26
  - pop\_vs\_soda, 33
  - pulse\_of\_the\_nation, 36
  - spotify, 37
  - voices, 42
  - weather\_australia, 42
  - weather\_perth, 44
  - weather\_WU, 45
- 
- airbnb, 3
  - airbnb\_small, 4
- 
- bald\_eagles, 5
  - basketball, 5
  - bechdel, 7
  - big\_word\_club, 7
  - bike\_users, 10
  - bikes, 9
  - bird\_counts, 11
  - book\_banning, 11
  - cherry\_blossom\_sample, 13
  - classification\_summary, 13
  - classification\_summary\_cv, 14
  - climbers\_sub, 15
  - coffee\_ratings, 16
  - coffee\_ratings\_small, 17
  - equality\_index, 18
  - fake\_news, 19
  - football, 20
  - hotel\_bookings, 21
  - loons, 22
  - moma, 23
  - moma\_sample, 24
  - naive\_classification\_summary, 24
  - naive\_classification\_summary\_cv, 25
  - penguins\_bayes, 26
  - plot\_beta, 27
  - plot\_beta\_binomial, 27
  - plot\_beta\_ci, 28
  - plot\_binomial\_likelihood, 29
  - plot\_gamma, 29
  - plot\_gamma\_poisson, 30
  - plot\_normal, 31
  - plot\_normal\_likelihood, 31
  - plot\_normal\_normal, 32
  - plot\_poisson\_likelihood, 33
  - pop\_vs\_soda, 33
  - prediction\_summary, 34
  - prediction\_summary\_cv, 35
  - pulse\_of\_the\_nation, 36
  - sample\_mode, 37

spotify, [37](#)  
summarize\_beta, [38](#)  
summarize\_beta\_binomial, [39](#)  
summarize\_gamma, [40](#)  
summarize\_gamma\_poisson, [40](#)  
summarize\_normal\_normal, [41](#)

voices, [42](#)

weather\_australia, [42](#)  
weather\_perth, [44](#)  
weather\_WU, [45](#)