Package ‘dabestr’

April 20, 2020

Type       Package
Title      Data Analysis using Bootstrap-Coupled Estimation
Version    0.2.5
Maintainer Joses W. Ho <joseshowh@gmail.com>

Description Data Analysis using Bootstrap-Coupled ESTimation.

Estimation statistics is a simple framework that avoids the pitfalls of significance testing. It uses familiar statistical concepts: means, mean differences, and error bars. More importantly, it focuses on the effect size of one’s experiment/intervention, as opposed to a false dichotomy engendered by P values.

An estimation plot has two key features:
1. It presents all datapoints as a swarmplot, which orders each point to display the underlying distribution.
2. It presents the effect size as a bootstrap 95% confidence interval on a separate but aligned axes.

Estimation plots are introduced in Ho et al., Nature Methods 2019, 1548-7105. <doi:10.1038/s41592-019-0470-3>.

The free-to-view PDF is located at <https://rdcu.be/bHhJ4>.

License file LICENSE

URL https://github.com/ACCLAB/dabestr

BugReports https://github.com/ACCLAB/dabestr/issues

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0), boot, magrittr

Imports cowplot, dplyr, ellipsis, ggplot2 (>= 3.2), forcats, ggforce, ggbeeswarm, plyr, RColorBrewer, rlang, simpleboot, stringr, tibble, tidyr

RoxygenNote 7.1.0

Suggests knitr, rmarkdown, tufte, testthat, vdiffr

VignetteBuilder knitr

NeedsCompilation no
**Description**

*dabest* applies a summary function (*func*, default *mean*) to the groups listed in *idx*, which are factors/strings in the *x* column of *.data*. The first element of *idx* is the control group. The difference between *func*(group_n) and *func*(control) is computed, for every subsequent element of *idx*.

For each comparison, a bootstrap confidence interval is constructed for the difference, and bias correction and acceleration is applied to correct for any skew. *dabest* uses bootstrap resampling to compute non-parametric assumption-free confidence intervals, and visualizes them using estimation plots with a specialized *plot.dabest* function.

**Usage**

```r
dabest(
  .data,  
  x,  
  y,  
  idx,  
  paired = FALSE,  
  id.column = NULL,  
  ci = 95,  
  reps = 5000,  
  func = mean,  
  seed = 12345  
)
```
Arguments

.data A data.frame or tibble.

x, y Columns in .data.

idx Accepts a vector containing factors or strings in the x column, or a list containing
vectors as noted above. If a vector is supplied, the first element will be the
control group, so all differences will be computed for every other group and this
first group. If a list of vectors is supplied, a multi-plot will be generated, with
the first element of each vector being the control group.

paired boolean, default FALSE. If TRUE, the two groups are treated as paired samples.
The control_group group is treated as pre-intervention and the test_group
group is considered post-intervention.

id.column, default NULL. A column name indicating the identity of the datapoint if the data
is paired. This must be supplied if paired is TRUE.

ci float, default 95. The level of the confidence intervals produced. The default ci
= 95 produces 95% CIs.

reps integer, default 5000. The number of bootstrap resamples that will be generated.

func function, default mean. This function will be applied to control and test
individually, and the difference will be saved as a single bootstrap resample.
Any NaNs will be removed automatically with na.omit.

seed integer, default 12345. This specifies the seed used to set the random num-
ber generator. Setting a seed ensures that the bootstrap confidence intervals for
the same data will remain stable over separate runs/calls of this function. See
set.seed for more details.

Details

Estimation statistics is a statistical framework that focuses on effect sizes and confidence intervals
around them, rather than \( P \) values and associated dichotomous hypothesis testing.

Value

A list with 7 elements: data, x, y, idx, id.column, result, and summary.
data, x, y, id.column, and idx are the same keywords supplied to dabest as noted above.

x and y are quoted variables for tidy evaluation by plot.

summary is a tibble with func applied to every group specified in idx. These will be used by
plot() to generate the estimation plot.

result is a tibble with the following 15 columns:

control_group, test_group
The name of the control group and test group respectively.

control_size, test_size
The number of observations in the control group and test group respectively.

func The func passed to bootdiff.
paired: Is the difference paired (TRUE) or not (FALSE)?
difference: The difference between the two groups; effectively func(test_group) - func(control_group).
variable: The variable whose difference is being computed, i.e. the column supplied to y.

ci: The ci passed to the bootdiff.

`bc`ci_low, `bc`ci_high: The lower and upper limits of the Bias Corrected and Accelerated bootstrap confidence interval.

`pct`ci_low, `pct`ci_high: The lower and upper limits of the percentile bootstrap confidence interval.

bootstraps: The array of bootstrap resamples generated.

References


See Also

`plot.dabest`, which generates an estimation plot from the `dabest` object.

Run `vignette("Using dabestr", package = "dabestr")` in the console to read more about using parameters to control the plot features.

Examples

```r
# Performing unpaired (two independent groups) analysis.
unpaired_mean_diff <- dabest(iris, Species, Petal.Width,
    idx = c("setosa", "versicolor"),
    paired = FALSE)

# Display the results in a user-friendly format.
unpaired_mean_diff

# Produce an estimation plot.
plot(unpaired_mean_diff)

# Performing paired analysis.
# First, we munge the 'iris' dataset so we can perform a within-subject comparison of sepal length vs. sepal width.

new.iris <- iris
new.iris$ID <- 1: length(new.iris)
setosa.only <-
    new.iris %>%
    tidyr::gather(key = Metric, value = Value, -ID, -Species) %>%
    dplyr::filter(Species %in% c("setosa"))
```
```r
paired_mean_diff <- dabest(setosa.only, Metric, Value,  
idx = c("Sepal.Length", "Sepal.Width"),  
paired = TRUE, id.col = ID)

# Computing the median difference.
unpaired_median_diff <- dabest(iris, Species, Petal.Width,  
idx = c("setosa", "versicolor", "virginica"),  
paired = FALSE,  
func = median)

# Using pipes to munge your data and then passing to `dabest`.
# First, we generate some synthetic data.
set.seed(12345)
N <- 70
c <- rnorm(N, mean = 50, sd = 20)
t1 <- rnorm(N, mean = 200, sd = 20)
t2 <- rnorm(N, mean = 100, sd = 70)
long.data <- tibble::tibble(Control = c, Test1 = t1, Test2 = t2)

# Munge the data using `gather`, then pass it directly to `dabest`
meandiff <- long.data %>%
tidyr::gather(key = Group, value = Measurement) %>%
dabest(x = Group, y = Measurement,  
idx = c("Control", "Test1", "Test2"),  
paired = FALSE)
```

---

**dabestr**  

---

**dabestr: A package for producing estimation plots.**

---

**Description**

The dabestr package provides a function to construct bootstrap confidence intervals for differences between groups **dabest** and a function to generate estimation plots **plot.dabest**.
lsat_scores  

Description

This dataset is taken from Thomason et al. (2014) who studied a group of 12 students who undertook a course in critical thinking. Their scores on the Logical Reasoning section of the Law School Aptitude Test (LSAT) were assessed before and after training.

Usage

lsat_scores

Format

A list with two elements: pretrain and posttrain.

Details

It is also found in Chapter 8 "The Paired Design" (pp 195 - 199) of Introduction to the New Statistics (Routledge, 2017), by Geoff Cumming and Robert Calin-Jageman.

Source


plot.dabest  

Estimation Plot

Description

An estimation plot has two key features.

1. It presents all datapoints as a swarmplot or sinaplot, which orders each point to display the underlying distribution.

2. It presents the effect size as a bootstrap 95 percent confidence interval on a separate but aligned axes.

Estimation plots emerge from estimation statistics, an intuitive framework that avoids the pitfalls of significance testing. It uses familiar statistical concepts: means, mean differences, and error bars. More importantly, it focuses on the effect size of one’s experiment/intervention, as opposed to a false dichotomy engendered by $P$ values. This function takes the output of the dabest function and produces an estimation plot.
Usage

## S3 method for class 'dabest'
plot(
  x,
  ...,
  color.column = NULL,
  palette = "Set1",
  float.contrast = TRUE,
  slopegraph = TRUE,
  group.summaries = "mean_sd",
  rawplot.type = c("swarmplot", "sinaplot"),
  rawplot.ylim = NULL,
  rawplot.ylabel = NULL,
  rawplot.markersize = 2,
  rawplot.groupwidth = 0.3,
  effsize.ylim = NULL,
  effsize.ylabel = NULL,
  effsize.markersize = 4,
  theme = ggplot2::theme_classic(),
  tick.fontsize = 11,
  axes.title.fontsize = 14,
  swarmplot.params = NULL,
  sinaplot.params = NULL,
  slopegraph.params = NULL
)

Arguments

x A dabest object, generated by the function of the same name.

... Signature for S3 generic function.

color.column default NULL. This is a column in the data.frame passed to the dabest function. This column will be treated as a factor and used to color the datapoints in the rawdata swarmplot.

palette default "Set1". Accepts any one of the RColorBrewer palettes, or a vector of colors. Colors can be specified as RGB hexcode or as a named color. See the "Palettes" section in scale_color_brewer for more on palettes. To obtain all 657 named colors in R, enter colors() at the console.

float.contrast default TRUE. If idx in the dabest object contains only 2 groups, float.contrast = TRUE will plot the effect size and the bootstrap confidence interval in a horizontally-aligned axes (also known as a Gardner-Altman plot.)

slopegraph boolean, default TRUE. If the dabest object contains paired comparisons, slopegraph = TRUE will plot the rawdata as a Tufte slopegraph.

group.summaries "mean_sd", "median_quartiles", or NULL. Plots the summary statistics for each group. If `mean_sd`, then the mean and standard deviation of each group is plotted as a gapped line beside each group. If `median_quartiles`, then the median
and 25th & 75th percentiles of each group is plotted instead. If `group.summaries = NULL`, the summaries are not shown.

**rawplot.type**
default "beeswarm". Accepts either "beeswarm" or "sinaplot". See `geom_quasirandom` and `geom_sina` for more information.

**rawplot.ylim**
default NULL. Enter a custom y-limit for the rawdata plot. Accepts a vector of length 2 (e.g. c(-50, 50)) that will be passed along to `coord_cartesian`.

**rawplot.ylabel**
default NULL. Accepts a string that is used to label the rawdata y-axis. If NULL, the column name passed to `y` is used.

**rawplot.markersize**
default 2. This is the size (in points) of the dots used to plot the individual datapoints. There are 72 points in one inch. See this article for more info.

**rawplot.groupwidth**
default 0.3. This is the maximum amount of spread (in the x-direction) allowed, for each group.

**effsize.ylim**
default NULL. Enter a custom y-limit for the effect size plot. This parameter is ignored if `float.contrast = TRUE`. Accepts a vector of length 2 (e.g. c(-50, 50)) that will be passed along to `coord_cartesian`.

**effsize.ylabel**
default NULL. Accepts a string that is used to label the effect size y-axis. If NULL, this axes will be labeled "(un)paired func difference", where func is the function passed to dabest.

**effsize.markersize**
default 4. This is the size (in points) of the dots used to indicate the effect size.

**theme**
default `theme_classic`.

**tick.fontsize**
default 11. This controls the font size (in points) of all tick labels.

**axes.title.fontsize**
default 14. This determines the font size (in points) of the axes titles.

**swarmplot.params**
default NULL. Supply list of keyword = value pairs to `geom_quasirandom`.

**sinaplot.params**
default NULL. Supply list of keyword = value pairs to `ggforce::geom_sina()`.

**slopegraph.params**
default NULL. Supply list of keyword = value pairs to `ggplot2::geom_line()`. This controls the appearance of the lines plotted for a paired slopegraph.

**Value**

A `ggplot` object.

**References**

See Also

The `dabest` function.
Run `vignette("Using dabestr", package = "dabestr")` in the console to read more about using parameters to control the plot features.

Examples

```r
# Performing unpaired (two independent groups) analysis.
unpaired_mean_diff <- dabest(iris, Species, Petal.Width,
idx = c("setosa", "versicolor"),
paired = FALSE)

# Create a Gardner-Altman estimation plot.
plot(unpaired_mean_diff)

# Comparing versicolor and virginica petal width to setosa petal width.
shared_control_data <- dabest(iris, Species, Petal.Width,
idx = c("setosa", "versicolor", "virginica")
)

# Create a Cumming estimation plot.
plot(shared_control_data)
```

print.dabest

Print a `dabest` object

Description

Print a `dabest` object

Usage

```r
## S3 method for class 'dabest'
print(x, ..., signif_digits = 3)
```

Arguments

- **x**  
  A dabest object, generated by the function of the same name.

- **...**  
  Signature for S3 generic function.

- **signif_digits**  
  integer, default 3. All numerical values in the printout will be rounded to this many significant digits.

Value

A summary of all the relevant effect sizes computed.
**Examples**

```r
# Performing unpaired (two independent groups) analysis.
unpaired_mean_diff <- dabest(iris, Species, Petal.Width,
   idx = c("setosa", "versicolor"),
   paired = FALSE)

# Display the results in a user-friendly format.
print(unpaired_mean_diff)
```

---

**transcription_scores**  
*Transcription Scores.*

**Description**

This dataset is taken from Mueller and Oppenheimer (2014), comparing the percentage of notes that was verbatim transcribed during a lecture by two independent groups of students: one using pen and paper, and one using laptops.

**Usage**

```r
transcription_scores
```

**Format**

A list with two elements: `pen` and `laptop`.

**Details**

It is also found in Chapter 7 "The Independent Groups Design" (pp 160 - 166) of Introduction to the New Statistics (Routledge, 2017), by Geoff Cumming and Robert Calin-Jageman.

**Source**

**wellbeing_ind**

*Wellbeing Scores (2 independent groups).*

**Description**

This is a synthetic dataset from Geoff Cumming. 20 students were randomly assigned to spend the afternoon reading in the library—the Control condition—or reading in the local botanical gardens—the Test condition. At the end of the session, each student completed a measure of his or her perceived well-being.

**Usage**

```
wellbeing_ind
```

**Format**

A list with two elements: control and test.

**Details**

It is found in Chapter 11 "The Paired Design" (page 286) of Understanding the New Statistics (Routledge, 2012) by Geoff Cumming.

**Source**


---

**wellbeing_paired**

*Wellbeing Scores (Before and after design).*

**Description**

This is a synthetic dataset from Geoff Cumming. A single group of 10 students first completed a well-being questionnaire (before), spent the afternoon reading in the botanical gardens, then gave well-being scores once again (after).

**Usage**

```
wellbeing_paired
```

**Format**

A list with two elements: before and after.
Details

It is found in Chapter 11 "The Paired Design" (page 291) of Understanding the New Statistics (Routledge, 2012) by Geoff Cumming.

Source

Index

*Topic datasets
  lsat_scores, 6
  transcription_scores, 10
  wellbeing_ind, 11
  wellbeing_paired, 11

coord_cartesian, 8

dabest, 2, 5, 6, 9
dabestr, 5

factor, 7

geom_quasirandom, 8
geom_sina, 8

lsat_scores, 6

mean, 2

plot.dabest, 2, 4, 5, 6
print.dabest, 9

scale_color_brewer, 7
set.seed, 3

theme_classic, 8
tibble, 3
transcription_scores, 10

wellbeing_ind, 11
wellbeing_paired, 11