Package: posthocbounds (via r-universe)

November 27, 2026

Type Package

Title Post Hoc Bounds on True Discoveries for Gaussian Tests

Version 0.1.0

Description Computes lower confidence bounds on the number of true discoveries in any selection of hypotheses, chosen after seeing the data, from independent one-sided Gaussian test statistics. The bounds hold simultaneously over all selections, as in the post hoc inference of Goeman and Solari (2011) <doi:10.1214/11-STS356>. The bound on all hypotheses is the smallest of the Sidak bound, closed testing with the largest p-value as local test, and bounds from counts of statistics above thresholds and from a centered statistic that gathers many weak signals. For a selection, the bound also uses a reference family of p-value levels as in Blanchard, Neuvial and Roquain (2020) <doi:10.1214/19-AOS1847>. The package reports bounds for the most significant hypotheses and a bound on the false discovery proportion for each hypothesis, and it also covers statistics that are dependent within independent blocks of bounded size.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.5.0)

Imports graphics, stats, utils

LazyData true

NeedsCompilation no

Author Satoshi Nakamura [aut, cre]

Maintainer Satoshi Nakamura <s-nakamura@med.mie-u.ac.jp>

Repository https://cran.r-universe.dev

Date/Publication 2026-11-27 22:08:39 UTC

RemoteUrl https://github.com/cran/posthocbounds

RemoteRef HEAD

RemoteSha 45f7f430644b80deffa31c7077b881ad918fffbc

Contents

posthocbounds-package . . . 2
combined_bound . . . 3
dialysis . . . 5
hypothesis_table . . . 6
plot.combined_bound . . . 7
ranked_bounds . . . 8
select_by_fdp . . . 9
true_discoveries . . . 10

Index **12**

`posthocbounds-package` *Post hoc bounds on true discoveries for Gaussian tests*

Description

The package computes lower confidence bounds on the number of true discoveries in selections of hypotheses from one-sided Gaussian test statistics. The bounds hold simultaneously over all selections, so a selection may be chosen after looking at the data.

Details

Start with `combined_bound`, which takes the statistics or the one-sided p-values and prints the main results. The other functions read the result:

- `true_discoveries`: bounds for any selections.
- `ranked_bounds`: bounds for the k most significant hypotheses, for every k.
- `hypothesis_table`: one row per hypothesis, with a bound on the false discovery proportion to report next to its p-value.
- `select_by_fdp`: the largest list of most significant hypotheses with a given bound on the false discovery proportion.
- `plot`: the confidence envelope.

The data set `dialysis` gives an example with 6641 dialysis facilities.

Author(s)

Satoshi Nakamura <s-nakamura@med.mie-u.ac.jp>

combined_bound *Post hoc bounds on true discoveries*

Description

Computes, from one-sided Gaussian test statistics, lower confidence bounds on the number of true discoveries that hold simultaneously for every selection of hypotheses, including selections made after looking at the data.

Usage

```
combined_bound(z = NULL, p = NULL, alpha = 0.05, block_size = 1,
               equal_blocks = FALSE)

## S3 method for class 'combined_bound'
print(x, ...)

## S3 method for class 'combined_bound'
summary(object, ...)
```

Arguments

z	numeric vector of test statistics, one per hypothesis. Large values are evidence against the null hypothesis. Names, if present, are used as hypothesis labels.
p	numeric vector of one-sided p-values, as an alternative to z. They are converted by z = qnorm(1 - p), so they must come from Gaussian statistics as described in Details.
alpha	error level; the bounds hold simultaneously with probability at least 1 - alpha.
block_size	largest number of statistics in a block of dependent statistics. The default 1 means that all statistics are independent. See Details.
equal_blocks	logical; set to TRUE only when every block has exactly block_size statistics. It allows one more component bound (threshold counts with Chernoff tolerances) when block_size > 1.
x, object	a result of combined_bound.
...	not used.

Details

Model: Statistic z_i follows a Gaussian distribution with mean θ_i and variance 1, and hypothesis i has the null hypothesis $H_i : \theta_i \leq 0$. A selected hypothesis is a true discovery when $\theta_i > 0$ and a false discovery otherwise. The statistics must be independent, or, when block_size = b > 1, fall into independent blocks of at most b statistics with any dependence inside a block. The procedure needs only b, not the blocks themselves.

Guarantee: With probability at least `1 - alpha`, every selection S of the hypotheses contains at least the reported number of true discoveries, for all selections at once. Selections may therefore be chosen after looking at the data, and several selections may be reported together.

Method: The bound N on the number of false discoveries among all hypotheses is the smallest of the following component bounds, each computed at level `alpha/4`:

1. the number of statistics below the Sidak cutoff $\Phi^{-1}\{(1-\alpha/4)^{1/m}\}$ (the Bonferroni cutoff $\Phi^{-1}\{1-\alpha/(4m)\}$ when `block_size > 1`);
2. closed testing with the largest p-value as local test;
3. two bounds from the numbers of statistics above a grid of thresholds, with Bernstein and with Chernoff tolerances, and a bound from a centered statistic that gathers many weak signals; these three share one union bound at level `alpha/4`.

For a selection S, the bound on the false discoveries in S is the smallest of $|S|$, N rounded down, the number of hypotheses in S below the cutoff, and a reference family bound at level `alpha/4` that counts the p-values of S above a grid of levels $q = 2^{-1}, 2^{-2}, \ldots$. The bound on true discoveries is $|S|$ minus this number.

When many hypotheses carry weak signals, the threshold counts and the centered statistic can make the bound on the number of false discoveries much smaller than the Sidak bound and the bound of closed testing with Simes local tests.

Value

An object of class `"combined_bound"`, a list with components

`m`	number of hypotheses.
`alpha`, `block_size`, `equal_blocks`	the arguments.
`z`, `p_value`	statistics and one-sided p-values.
`labels`	hypothesis labels, or `NULL`.
`components`	named vector of the component bounds on the number of false discoveries among all hypotheses (`NA` for a component that is not used).
`bound`	the combined bound N, the smallest component bound.
`false_all`	the bound on false discoveries among all hypotheses, N rounded down.
`z_cutoff`	the cutoff for individual significance.
`significant`	logical vector; `TRUE` for hypotheses that are true discoveries on their own. They control the familywise error rate at level `alpha`.
`call`	the call.

Use `true_discoveries`, `ranked_bounds`, `hypothesis_table` and `select_by_fdp` to read the bounds, and `plot` to draw them.

References

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26, 584–597. doi:10.1214/11STS356

Blanchard, G., Neuvial, P. and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48, 1281–1303. doi:10.1214/19AOS1847

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626–633. doi:10.1080/01621459.1967.10482935

See Also

`true_discoveries`, `ranked_bounds`, `hypothesis_table`, `select_by_fdp`, `plot.combined_bound`

Examples

```
## 2000 hypotheses, 300 of them with a weak signal
set.seed(1)
z <- rnorm(2000, mean = c(rep(2, 300), rep(0, 1700)))
fit <- combined_bound(z)
fit
summary(fit)

## the same data given as one-sided p-values
fit_p <- combined_bound(p = pnorm(z, lower.tail = FALSE))
fit_p$false_all == fit$false_all

## dialysis facilities: mortality below 1.2 times the national rate
se <- (log(dialysis$upper) - log(dialysis$lower)) / (2 * qnorm(0.975))
z <- (log(1.2 * 22.2) - log(dialysis$mortality)) / se
fit <- combined_bound(z)
fit
```

`dialysis` *Mortality of United States dialysis facilities*

Description

Risk adjusted mortality rates of 6641 dialysis facilities in the United States, with their 95 percent confidence limits, from the Dialysis Facility Compare data of the Centers for Medicare and Medicaid Services.

Usage

```
dialysis
```

Format

A data frame with 6641 rows and 7 columns:

state two letter state code.

ownership factor, `"for-profit"` or `"non-profit"`.

chain logical; `TRUE` if the facility belongs to a chain.

patients number of patients included in the survival summary.

mortality risk adjusted mortality rate, in percent.

lower, upper limits of the published 95 percent confidence interval for the mortality rate, in percent.

Details

The rows are the facilities with survival data (availability code 001), at least 100 patients in the survival summary and positive mortality rate and lower limit. The national mortality rate of the same release is 22.2 percent.

Treating each published interval as a Gaussian interval on the log scale gives the standard error `(log(upper) - log(lower)) / (2 * qnorm(0.975))` of the log rate. For a margin c, the statistic `(log(c * 22.2) - log(mortality)) / se` tests the null hypothesis that the mortality rate of the facility is at least c times the national rate; the examples use $c = 1.2$.

Source

Centers for Medicare and Medicaid Services, Provider Data Catalog, "Dialysis Facility - Listing by Facility" (dataset 23ew-n7w9, release of 16 June 2026, file 'DFC_FACILITY.csv'), https://data.cms.gov/provider-data/dataset/23ew-n7w9. The data are works of the United States government and are in the public domain.

Examples

```
se <- (log(dialysis$upper) - log(dialysis$lower)) / (2 * qnorm(0.975))
z <- (log(1.2 * 22.2) - log(dialysis$mortality)) / se
summary(z)
```

---

hypothesis_table *Bounds for each hypothesis*

Description

Lists each hypothesis with its statistic, its p-value, its rank, whether it is significant individually, and a bound on the false discovery proportion (FDP) that can be reported next to its p-value.

Usage

```
hypothesis_table(fit, sort = FALSE)
```

Arguments

`fit`	a result of `combined_bound`.
`sort`	logical; if `TRUE`, rows are ordered from the most significant hypothesis, otherwise they follow the input.

Details

For the hypothesis of rank s, `fdp_at_most` is the smallest FDP bound among the selections of the k most significant hypotheses with $k \geq s$, that is, among the lists of most significant hypotheses that contain it. It is built in the same way as the q-value of Storey (2003) from the FDP bounds of `ranked_bounds`, and it is nondecreasing in the rank. With probability at least `1 - alpha`, for every level q at once, the hypotheses with `fdp_at_most <= q` have a false discovery proportion of at most q.

Value

A data frame with the columns

`hypothesis`	label, or index if the statistics had no names.
`z`	statistic.
`p_value`	one-sided p-value, not adjusted for multiplicity.
`rank`	rank from the largest statistic.
`significant`	`TRUE` if the hypothesis is a true discovery on its own; these hypotheses control the familywise error rate at level `alpha`.
`fdp_at_most`	bound on the false discovery proportion, as in Details.

References

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31, 2013–2035. doi:10.1214/aos/1074290335

See Also

`select_by_fdp`, `ranked_bounds`

Examples

```
se <- (log(dialysis$upper) - log(dialysis$lower)) / (2 * qnorm(0.975))
z <- (log(1.2 * 22.2) - log(dialysis$mortality)) / se
fit <- combined_bound(z)
head(hypothesis_table(fit, sort = TRUE))
```

`plot.combined_bound` *Plot the confidence envelope*

Description

Draws the lower bound on the number of true discoveries among the k most significant hypotheses against k, with the line of slope 1 for reference.

Usage

```
## S3 method for class 'combined_bound'
plot(x, log = "",
     xlab = "Number of most significant hypotheses (k)",
     ylab = "True discoveries, at least", main = NULL, ...)
```

Arguments

x	a result of combined_bound.
log	passed to plot.default; for example "xy" for logarithmic axes, in which case values of k with a zero bound are left out.
xlab, ylab, main	axis labels and title.
...	further arguments passed to plot.default.

Value

x, invisibly.

See Also

ranked_bounds

Examples

```
set.seed(1)
z <- rnorm(2000, mean = c(rep(2, 300), rep(0, 1700)))
plot(combined_bound(z))
```

---

ranked_bounds *Bounds for the most significant hypotheses*

Description

Reports, for each k, the bounds for the selection of the k hypotheses with the largest statistics. As a function of k, the lower bound on true discoveries is a confidence envelope.

Usage

```
ranked_bounds(fit, k = NULL)
```

Arguments

fit	a result of combined_bound.
k	numbers of most significant hypotheses; the default NULL gives every k from 1 to the number of hypotheses.

Details

Ties among the statistics are broken by the order of the input. The bounds hold simultaneously for all k, and for all other selections, with probability at least `1 - alpha`.

Value

A data frame with the columns

`k`	number of most significant hypotheses.
`z_k`	the k-th largest statistic.
`true_at_least`	lower bound on the number of true discoveries among them.
`false_at_most`	upper bound on the number of false discoveries among them.
`fdp_at_most`	upper bound on their false discovery proportion.

See Also

`combined_bound`, `plot.combined_bound`, `hypothesis_table`

Examples

```
set.seed(1)
z <- rnorm(2000, mean = c(rep(2, 300), rep(0, 1700)))
fit <- combined_bound(z)
ranked_bounds(fit, k = c(10, 50, 100, 200, 300, 500))
```

`select_by_fdp` *Largest list of most significant hypotheses with a given FDP bound*

Description

Returns the largest selection of most significant hypotheses whose bound on the false discovery proportion (FDP) is at most `fdp`.

Usage

```
select_by_fdp(fit, fdp = 0.1)
```

Arguments

`fit`	a result of `combined_bound`.
`fdp`	the largest FDP bound allowed, between 0 and 1.

Details

The selection consists of the hypotheses whose `fdp_at_most` in `hypothesis_table` is at most `fdp`. With probability at least `1 - alpha`, the FDP of the returned selection is at most `fdp` for every value of `fdp` at once.

Value

An integer vector of hypothesis indices, from the most significant, named by the hypothesis labels when they exist. It is empty when no selection meets the bound.

See Also

hypothesis_table, true_discoveries

Examples

```
se <- (log(dialysis$upper) - log(dialysis$lower)) / (2 * qnorm(0.975))
z <- (log(1.2 * 22.2) - log(dialysis$mortality)) / se
fit <- combined_bound(z)
top <- select_by_fdp(fit, fdp = 0.1)
length(top)
true_discoveries(fit, top)
```

---

true_discoveries *Bounds on true discoveries in selections of hypotheses*

Description

Reports, for one or more selections of hypotheses, the lower bound on the number of true discoveries, the upper bound on the number of false discoveries and the upper bound on the false discovery proportion (FDP).

Usage

```
true_discoveries(fit, select = NULL, by = NULL, details = FALSE)
```

Arguments

fit	a result of combined_bound.
select	a selection, or a list of selections. A selection is a vector of indices, a logical vector with one value per hypothesis, or a vector of hypothesis labels. The default NULL selects all hypotheses.
by	alternatively to select, a vector with one value per hypothesis; each distinct value defines one selection.
details	logical; if TRUE, adds the three bounds on false discoveries whose minimum is reported (see Value).

Details

The bounds hold simultaneously for all selections with probability at least 1 - alpha, so the selections may be chosen after looking at the data and reported together.

Value

A data frame with one row per selection and the columns

`selection`	name of the selection.
`selected`	number of hypotheses in the selection.
`true_at_least`	lower bound on the number of true discoveries.
`false_at_most`	upper bound on the number of false discoveries.
`fdp_at_most`	upper bound on the false discovery proportion, `false_at_most / selected`.

With `details = TRUE`, also

`from_all`	the bound on false discoveries among all hypotheses, capped at the selection size.
`from_cutoff`	the number of selected hypotheses below the cutoff for individual significance.
`from_pvalues`	the reference family bound from the selected p-values.

See Also

`combined_bound`, `ranked_bounds`

Examples

```
se <- (log(dialysis$upper) - log(dialysis$lower)) / (2 * qnorm(0.975))
z <- (log(1.2 * 22.2) - log(dialysis$mortality)) / se
fit <- combined_bound(z)

## all facilities, and the facilities of two states
true_discoveries(fit)
true_discoveries(fit, list(California = dialysis$state == "CA",
                           `New York` = dialysis$state == "NY"))

## one selection per ownership type, with the parts of the bound
true_discoveries(fit, by = dialysis$ownership, details = TRUE)
```

Index

∗ **datasets**
    dialysis, 5
∗ **package**
    posthocbounds-package, 2

combined_bound, *2*, 3, *6*, *8–11*

dialysis, *2*, 5

hypothesis_table, *2*, *4*, *5*, 6, *9*, *10*

plot, *2*, *4*
plot.combined_bound, *5*, 7, *9*
plot.default, *8*
posthocbounds *(*posthocbounds-package*)*, 2
posthocbounds-package, 2
print.combined_bound *(*combined_bound*)*, 3
print.summary.combined_bound
        *(*combined_bound*)*, 3

ranked_bounds, *2*, *4*, *5*, *7*, 8, *8*, *11*

select_by_fdp, *2*, *4*, *5*, *7*, 9
summary.combined_bound
        *(*combined_bound*)*, 3

true_discoveries, *2*, *4*, *5*, 10, *10*