Package ‘CrossScreening’

April 21, 2017

Type Package

Title Cross-Screening in Observational Studies that Test Many Hypotheses

Version 0.1.1

Date 2017-04-20


Imports stats, parallel, plyr, tables

Suggests knitr, ggplot2

License GPL-2

RoxygenNote 5.0.1

LazyData true

VignetteBuilder knitr

NeedsCompilation no

Author Paul Rosenbaum [aut], Qingyuan Zhao [aut, cre]

Maintainer Qingyuan Zhao <qyzhao@wharton.upenn.edu>

Repository CRAN

Date/Publication 2017-04-21 05:56:52 UTC

R topics documented:

CrossScreening-package .................................................. 2
bonferroni.fg ................................................................. 2
cross.screen ................................................................. 3
fallback.test ................................................................. 6
kappa2gamma ................................................................. 7
Description

Cross-screening is a new method that uses both random halves of the sample to screen and test many hypotheses. It generally improves statistical power in observational studies when many hypotheses are tested simultaneously.

Bonferroni’s correction with fixed $\Gamma$

Usage

bonferroni.fg(d, gamma = 1, mm = c(2, 2, 2), two.sided = TRUE)

Arguments

d  a matrix of treatment-minus-control differences.
gamma  sensitivity parameter (maximum odds different from a randomized experiment).
mm  test statistic, either a vector of length 3 or a matrix of three rows where each column corresponds to a U-statistic. Default is the (approximate) Wilcoxon’s signed rank test.
two.sided  whether a two-sided test should be used. If FALSE, test the one-sided alternative that the center of $d$ is positive.
**cross.screen**

**Details**

If \( \text{mm} \) is a matrix, this function computes a one-sided or two-sided p-value with each statistic (i.e. there is a p-value for every column of \( \text{d} \) and every column of \( \text{mm} \)), then does a Bonferroni correction over all the p-values.

**Value**

a vector of sensitivity values for each column of \( \text{d} \)

**Author(s)**

Qingyuan Zhao

---

**cross.screen**

Cross-screening

**Description**

Main functions that implements the cross-screening method in observational studies. `cross.screen` sorts the hypotheses by their sensitivity values and `cross.screen.fg` sorts by p-values at a fixed sensitivity \( \Gamma \).

**Usage**

```r
cross.screen(d1, d2, gamma = 1, mm = c(2, 2, 2), screen.value = c("sen", "p"), screen.method = c("threshold", "least.sensitive"), alpha.screen = 0.05, gamma.screen = gamma, least.sensitive = 2, two.sided = TRUE)

cross.screen.fg(d1, d2, gamma = 1, mm = c(2, 2, 2), screen.method = c("threshold", "least.sensitive"), alpha.screen = 0.05, gamma.screen = gamma, least.sensitive = 2, two.sided = TRUE)
```

**Arguments**

- `d1`: screen/test sample (treatment-minus-control differences), can be a matrix (rows are observations, columns are hypotheses)
- `d2`: test/screen sample, can be a matrix
- `gamma`: sensitivity parameter (maximum odds different from a randomized experiment)
- `mm`: a vector of matrix. If matrix, adaptively choose statistic. NULL means Wilcoxon's signed rank statistic.
- `screen.value`: either "sen" (using sensitivity value) or "p" (using p-value).
- `screen.method`: either keep all hypotheses significant at gamma.screen (option "threshold") or keep the least sensitive hypotheses (option "least.sensitive").
- `alpha.screen`: significance level used in screening.
gamma.screen screening threshold, default is 0, meaning no screening is used.

least.sensitive

the number of least sensitive hypotheses to keep

two.sided if TRUE, automatically select the sign to test; if FALSE, test the one-sided alternative that the center of d is positive.

Value
cross.screen returns a list

s1.kappa kappa values used to screen the hypotheses calculated using the first sample

s1.stat test statistics chosen using the first sample, if \( \mathbb{M} \) has more than 1 column

s1.side signs of alternative hypotheses chosen using the first sample

s1.order order of the hypotheses by s1.kappa if s1.kappa is above the threshold gamma.screen

p1 p-values computed using the first sample at sensitivity gamma

s2.kappa kappa values used to screen the hypotheses calculated using the second sample

s2.stat test statistics chosen using the second sample, if \( \mathbb{M} \) has more than 1 column

s2.side signs of alternative hypotheses chosen using the second sample

s2.order order of the hypotheses by s1.kappa if s1.kappa is above the threshold gamma.screen

p2 p-values computed using the second sample at sensitivity gamma

p Bonferroni adjusted p-values at sensitivity gamma computed using p1 and p2 (they can be directly used to control FWER)

cross.screen.fg returns a list

s1.p p-values used to screen the hypotheses calculated using the first sample

s1.stat test statistics chosen using the first sample, if \( \mathbb{M} \) has more than 1 column

s1.side signs of alternative hypotheses chosen using the first sample

s1.order order of the hypotheses by s1.p if s1.p is below the threshold alpha.screen

p1 p-values computed using the first sample at sensitivity gamma

s2.p p-values used to screen the hypotheses calculated using the second sample

s2.stat test statistics chosen using the second sample, if \( \mathbb{M} \) has more than 1 column

s2.side signs of alternative hypotheses chosen using the second sample

s2.order order of the hypotheses by s2.p if s2.p is above the threshold alpha.screen

p2 p-values computed using the second sample at sensitivity gamma

p Bonferroni adjusted p-values at sensitivity gamma computed using p1 and p2 (they can be directly used to control FWER)

Functions

• cross.screen.fg: Cross-screening with fixed \( \Gamma \)
Author(s)
Qingyuan Zhao

References

Examples

```r
n <- 100
p <- 20
d <- matrix(rnorm(n * p), n, p)
d[, 1] <- d[, 1] + 2
d1 <- d[(n/2):n]
d2 <- d[(n/2+1):n]
cross.screen(d1, d2,
    gamma = 9,
    gamma.screen = 1.25)
```

```
## One can run the hidden function CrossScreening:::table5(no.sims = 1)
## to generate Table 5 in the paper.

## The following code generates Table 1 in the paper.
require(CrossScreening)
data(nhanes.fish)
data(nhanes.fish.match)
data(nhanes.fish)
match <- nhanes.fish.match
outcomes <- grep("^o\", names(data))
log2diff <- function(y1, y2) {
    if (min(c(y1, y2)) == 0) {
        y1 <- y1 + 1
        y2 <- y2 + 1
    }
    log2(y1) - log2(y2)
}
d <- sapply(outcomes, function(j) log2diff(data[match$treated, j], data[match$control, j]))
set.seed(11)
split <- sample(1:nrow(d), nrow(d) / 2, replace = FALSE)
d1 <- d[split,]
d2 <- d[-split,]
mm <- matrix(c(2, 2, 2, 8, 5, 8), ncol = 2)
data.frame(outcome = names(data)[outcomes],
    p.value =
cross.screen(d1, d2,
```

```r
```
fallback.test

fallback.test

Description

Fallback procedure for multiple testing

Usage

fallback.test(p, alpha = 0.05, spread = 1)

Arguments

- **p**: a vector of p-values
- **alpha**: significance level
- **spread**: the way to spread alpha, either a vector of the same length as p or a single number to indicate equal spread in the first spread hypotheses.

Value

the rejected hypotheses (TRUE means reject, FALSE means accept)

Author(s)

Qingyuan Zhao

References

**kappa2gamma**

Transform sensitivity parameter in different scales

**Description**
Transform sensitivity parameter in different scales

**Usage**
kappa2gamma(kappa)  
gamma2kappa(gamma)

**Arguments**
kappa \( \kappa = \gamma/(1 + \gamma) \)
gamma the odds of treatment of two matched units can differ at most by a factor of gamma

**Functions**
- gamma2kappa: Transform a sensitivity parameter from \( \gamma \) scale to \( \kappa \) scale

**lead**

*Lead in children*

**Description**
Morton et al. (1982) compared the levels of lead in the blood of 33 children whose fathers worked in a factory where lead was used in the manufacture of batteries to 33 lead levels in matched control children of the same age from the same neighborhood. The variables are as follows:

**Usage**
data(lead)

**Format**
A data.frame.

**Details**
control lead levels (ug/dl)  
level father’s potential exposure  
hyg hygiene of father employed in the lead industry
References


---

**methotrexate**

**Methotrexate workers**

### Description

Methotrexate is a drug used to treat cancer, but there is concern that it may itself be carcinogenic in healthy individuals who are exposed while either manufacturing or administering the drug. Deng et al. (2005) compared 21 workers engaged in the production of methotrexate to 21 unexposed controls matched for age, gender, and smoking. The variables are (prefix “w” means exposed and “c” means control)

- **Mftcr** mutant frequency of TCR gene
- **Mfhrpt** mutant frequency of hprt gene
- **mtl** mean tail length
- **mtm** mean tail moment
- **id** identifier
- **sex** sex
- **age** age
- **smoke** smoking
- **years** exposure years
- **protection** protection measures, G for gloves, M for mask, N for none
- **mask** if the protection includes mask

### Usage

data(methotrexate)

### Format

A data.frame.

### References

multrnks

Approximate scores for ranks.

Description

This function modifies the multrnks function in the sensitivitymw package by also providing the exact scores. The scores are also normalized so that the maximum is 1.

Usage

multrnks(rk, mm, score.method = c("approximate", "exact"))

Arguments

rk
a vector of ranks

mm
a vector (m, munder, mover) or a matrix, each column a vector (m, munder, mover) that indicates the U-statistic. NULL means Wilcoxon's signed rank test.

score.method
either approximate score or exact score

Details

Exact and approximate rank scores yield similar bounds on P-values when sample size is large. The exact rank scores involve very large combinatorial coefficients when the same size is very large, whereas the nearly equivalent approximate scores do not.

Author(s)

Paul Rosenbaum, Qingyuan Zhao

nhanes.fish

Health effects of fish

Description

Data from NHANES (2013-2014) with 1107 observations and 87 variables. Variables whose name start with "o." are lab measurements (such as blood mercury) that can be used as outcomes. The demographics and background variables include gender, age, income, indicator for missing income, race, education, indicator for smoked ever, number of cigararttes smoked in the last month. Fish intakes in the last month (in servings) are summed up in the "fish" variable, which is used to create the binary indicator "fish.level".

Usage

data(nhanes.fish)
nhanes.log2diff

Format

A data.frame.

nhanes.fish.match  Pair matching result

Description

Each row is a matched pair, the first/second entry is the id of low/high fish intake in the nhanes.fish data frame.

Usage

data(nhanes.fish.match)

Format

A data.frame.

nhanes.log2diff  Obtains treatment-minus-control differences in the nhanes.fish dataset

Description

Obtains treatment-minus-control differences in the nhanes.fish dataset

Usage

nhanes.log2diff()

Value

a 234 * 46 matrix of the log2 differences
**power.sen**

*Power of sensitivity analysis*

**Description**

Power of sensitivity analysis

**Usage**

```r
power.sen(mu.F = 1/2, sigma.F = sqrt(1/3), d = NULL, mm = c(2, 2, 2),
gamma = 1, alpha = 0.05, I = 100, approx.method = c("changing.alpha",
"fixed.alpha"), score.method = c("approximate", "exact"))
```

**Arguments**

- `mu.F`: mean of the signed rank statistic
- `sigma.F`: standard deviation of the signed rank statistic
- `d`: empirical data used to estimate `mu.F` and `sigma.F` by jackknife
- `mm`: test statistic, either a vector of length 3 or a matrix of three rows where each column corresponds to a U-statistic. Default is the (approximate) Wilcoxon's signed rank test.
- `gamma`: target sensitivity level
- `alpha`: target significance level
- `I`: sample size
- `approx.method`: which approximation method to use?
- `score.method`: either approximate score or exact score

**Details**

If `approx.method` is "fixed.alpha", then the significance level `alpha` is considered fixed and the corresponding quantile negligible. Otherwise we also use the `alpha`-quantile in the approximation formula. For more detail, see the reference.

**Value**

power of the sensitivity analysis, possibly a vector if `mm` has multiple columns.

**References**

Examples

```r
power.sen(d = rnorm(100) + 0.5, I = 200, gamma = 2)

## The following code reproduces an example of power analysis in Zhao (2017)
power.sen(0.76, sqrt(0.26), gamma = 2.5, I = 200)
power.sen(0.76, sqrt(0.26), gamma = 2.5, I = 200, approx.method = "fixed.alpha")
```

---

**recycle.test**

*Recycling procedure for multiple testing*

### Description

Recycling procedure for multiple testing

### Usage

```r
recycle.test(p, alpha = 0.05)
```

### Arguments

- `p` a vector of p-values
- `alpha` significance level

### Details

WARNING: only supports recycle the first two tests.

### Value

rejected hypotheses

### Author(s)

Qingyuan Zhao
Description

This function implements Rosenbaum’s sensitivity analysis for pair-matched observational study with general signed score test. It is faster and more flexible than the `psens` function in the package `rbounds`.

Usage

```r
sen(d, mm = NULL, gamma = 1, alternative = c("greater", "less"),
    approx.method = c("normal", "permutation"),
    score.method = c("approximate", "exact"), tau = 0, num.perms = 10000)
```

Arguments

- `d`: a vector of treatment-minus-control differences
- `mm`: a vector (m, munder, mover) or a matrix, each column a vector (m, munder, mover) that indicates the U-statistic. NULL means Wilcoxon’s signed rank test.
- `gamma`: a vector of sensitivity parameters (must be >= 1).
- `alternative`: report p-value corresponds to the maximum ("upper") or minimum ("lower") bound
- `approx.method`: how to compute the $p$-value upper bound? either "normal" approximation or random "permutations".
- `score.method`: either approximate score or exact score
- `tau`: a scalar, null hypothesis is the additive effect is tau (default 0)
- `num.perms`: number of Monte-Carlo simulations used to compute the sensitivity value, if `approx.method` is "permutations".

Value

A list

- `p.value`: p-values corresponding to each entry of `gamma`
- `p.value2`: two sided p-values
- `gamma.hat`: estimate of design sensitivity
- `T`: test statistic
- `E`: Means of the test statistic under sensitivity `gamma`
- `V`: Variances of the test statistic under sensitivity `gamma`
- `eff.size`: Effect size of T compared to E and V
- `E.gamma1`: Expectation of T under null at Gamma = 1
Author(s)
Paul Rosenbaum, Qingyuan Zhao

References

Examples
```r
require(CrossScreening)
data(lead)
d.lead <- lead$exposed[-21] - lead$control[-21]
sen(d.lead, gamma = c(1, 2, 3, 4, 5, 6))
```

---

### sen.ci

**Point estimate and confidence interval for sensitivity analysis**

**Description**
Point estimate and confidence interval for sensitivity analysis

**Usage**
```r
sen.ci(d, mm = c(2, 2, 2), gamma = 1, alpha = 0.05, alpha.up = alpha/2, alpha.low = alpha/2, score.method = c("approximate", "exact"))
```

**Arguments**
- `d` a vector of treatment-minus-control differences
- `mm` a vector (m, munder, mover) that indicates the U-statistic. Does not support matrix `mm` in this function.
- `gamma` a vector of sensitivity parameters (must be >= 1).
- `alpha` significance level for the outer confidence interval
- `alpha.up` upper-tail probability of the confidence interval
- `alpha.low` lower-tail probability of the confidence interval
- `score.method` either approximate score or exact score

**Details**
See the `senmwCI` function in the `sensitivitymw` package.
Value

a list

point.estimate An interval of point estimates allowing for a bias of gamma in treatment assignment.

ci An confidence interval allowing for a bias of gamma in treatment assignment.

Author(s)

Qingyuan Zhao

Examples

data(lead)
d.lead <- lead$exposed[-21] - lead$control[-21]

sen.ci(d.lead, gamma = c(1, 2), alpha.up = 0, alpha.low = 0.05)

sen.value

Description

Compute sensitivity value

Usage

sen.value(d, alpha = 0.05, mm = c(2, 2, 2), alternative = c("greater", "less", "two.sided"), score.method = c("approximate", "exact"))

Arguments

d a vector or matrix of treatment-minus-control differences (each column corresponds to a hypothesis)

alpha significance level

mm test statistic, either a vector of length 3 or a matrix of three rows where each column corresponds to a U-statistic. Default is the (approximate) Wilcoxon's signed rank test.

alternative report p-value corresponds to the maximum ("upper") or minimum ("lower") bound

score.method either approximate score or exact score

Details

The alternative direction is the the center of d is greater than 0.
Value

sensitivity value, i.e. the kappa value such that the p-value becomes just insignificant. If \( mm \) is a matrix, then return a vector of sensitivity values corresponding to each column of \( mm \).

Author(s)

Qingyuan Zhao

References


Examples

d <- rnorm(100) + 1
gamma.star <- kappa2gamma(sen.value(d, alpha = 0.05, mm = matrix(c(2, 2, 2, 8, 5, 8), ncol = 2)))
gamma.star
sen(d, mm = c(2, 2, 2), gamma = gamma.star[1])$p.value # should equal the significance level 0.05
Index

*Topic datasets
  lead, 7
  methotrexate, 8
  nhanes.fish, 9
  nhanes.fish.match, 10

bonferroni.fg, 2

cross.screen, 3
CrossScreening-package, 2

fallback.test, 6

gamma2kappa (kappa2gamma), 7

kappa2gamma, 7

lead, 7

methotrexate, 8
multrnks, 9

nhanes.fish, 9
nhanes.fish.match, 10
nhanes.log2diff, 10

power.sen, 11

recycle.test, 12

sen, 13

sen.ci, 14

sen.value, 15