# A Bootstrap-Based Heterogeneity Test for Between-study Heterogeneity in Meta-Analysis

This R package boot.heterogeneity provides functions for testing the between-study heterogeneity in meta-analysis of standardized mean differences (d), Fisher-transformed Pearson's correlations (r), and log odds ratio (OR).

In the following three examples, we describe how to use our package boot.heterogeneity to test the between-study heterogeneity for each of the three effect sizes (d, r, OR). *Datasets*, *R codes*, and *Output* are provided so that applied researchers can easily replicate each example or modify the codes for their own datasets.

- The three example *datasets* are internal in our package, and researchers can load the datasets using boot.heterogeneity:::[dataset_name]. In each of the example datasets, the rows correspond to studies in meta-analysis, and the columns correspond to required input for that study, which includes, but is not limited to effect size, sample size(s), and moderators.

- The example *R codes* adopt the default value for some of the arguments (e.g., default nominal alpha level is 0.05). To change the defaults, use help() or ? to access the documentation page of each function (e.g., help(boot.run.cor)).

- The *output* is formatted to have the same layout across the examples.

Inclusion of *moderators* is an option for researchers who are interested in measuring the size of the between-study heterogeneity and using factors to explain the systematic between-study heterogeneity. To see how we include moderators, please go to section 1.2.

*Heterogeneity magnitude test* is a test in which the researchers can compare the magnitude of the between-study heterogeneity against a specific level. This specific level is denoted as lambda in our alternative hypothesis. To see how we test a specific lambda in the alternative hypothesis, please go to section 2.2.

*Parallel implementation* of the bootstrapping process can save us considerable amount of computing time, especially when the number of bootstrap replications is large. To see how we accelerate the bootstrapping process with parallel implementation and computing nodes, please go to section 3.2.

In the main text of the article, an "Empirical Illustration" section is included to discuss the three examples in more detail.

## Outline

- 0. Installation of the package
- 1. Standardized Mean Differences (d)
    - 1.1 Without moderators
    - 1.2 With moderators
- 2. Fisher-transformed Pearson's correlations (r)
    - 2.1 Heterogeneity magnitude test: lambda=0
    - 2.2 Heterogeneity magnitude test: lambda=0.08
- 3. Log odds ratio (OR)
    - 3.1 Without parallel implementation
    - 3.2 With parallel implementation

## 0. Installation of the package

For most recent updates, researchers are highly recommended to install the development version of this package from GitHub using the following syntax:

```r
# install.packages("devtools")
library(devtools)
devtools::install_github("gabriellajg/boot.heterogeneity",
                         force = TRUE,
                         build_vignettes = TRUE,
                         dependencies = TRUE)
library(boot.heterogeneity)
```

The newest version of this package will also be available from CRAN shortly.

Note that you'll need the following packages to install this package successfully:

```r
library(metafor) # for Q-test
#> Loading required package: Matrix
#> Loading 'metafor' package (version 2.4-0). For an overview
#> and introduction to the package please type: help(metafor).
library(pbmcapply) # optional - for parallel implementation of bootstrapping
#> Loading required package: parallel
library(HSAUR3) # for an example dataset in the tutorial
#> Loading required package: tools
library(knitr) # for knitting the tutorial
library(rmarkdown) # for knitting the tutorial
```

## 1. Standardized Mean Differences (d)

`boot.d()` is the function to test the between-study heterogeneity in meta-analysis of standardized mean differences (d).

### 1.1 Without moderators

Load the example dataset `selfconcept` first:

```r
selfconcept <- boot.heterogeneity:::selfconcept
```

`selfconcept` consists of 18 studies in which the effect of open versus traditional education on students' self-concept was studied (Hedges et al., 1981). The columns of `selfconcept` are: sample sizes of the two groups (`n1` and `n2`), Hedges's `g`, Cohen's `d`, and a moderator `X` (`X` not used in the current example).

```r
head(selfconcept, 3)
#>    n1  n2      g           d        X
#> 1 100 180  0.100   0.09972997  0.100
#> 2 131 138 -0.162  -0.16154452 -0.162
#> 3  40  40 -0.090  -0.08913183 -0.091
```

Extract the required arguments from `selfconcept`:

```r
# n1 and n2 are lists of samples sizes in two groups
n1 <- selfconcept$n1
n2 <- selfconcept$n2
# g is a list of effect sizes
g <- selfconcept$g
```

If `g` is a list of biased estimates of standardized mean differences in the meta-analytical study, a small-sample adjustment must be applied:

```
cm <- (1-3/(4*(n1+n2-2)-1)) #correct factor to compensate for small sample bias (Hedges, 1981)
d <- cm*g
```

Run the heterogeneity test using function `boot.d()` and adjusted effect size `d`:

```
boot.run <- boot.d(n1, n2, est = d, model = 'random', p_cut = 0.05)
```

Alternatively, such an adjustment can be performed on unadjusted effect size `g` by specifying `adjust = TRUE`:

```
boot.run2 <- boot.d(n1, n2, est = g, model = 'random', adjust = TRUE, p_cut = 0.05)
```

`boot.run` and `boot.run2` will return the same results:

```
boot.run
#>                 stat   p_value Heterogeneity
#> Qtest      23.391659 0.136929           n.s
#> boot.REML   2.037578 0.053100           n.s
```

```
boot.run2
#>                 stat   p_value Heterogeneity
#> Qtest      23.391659 0.136929           n.s
#> boot.REML   2.037578 0.053100           n.s
```

- The first line presents the results of Q-test of a random-effects model. The Q-statistic is $Q(\mathrm{df} = 17) = 23.39$ and the associated p-value is 0.137. Using a cutoff alpha level (i.e., nominal alpha level) of either 0.05 or 0.1, this statistic is n.s (not significant). The homogeneity assumption is not rejected.
- The second line presents the results of B-REML-LR. The B-REML-LRT statistic is 2.04 and the bootstrap-based p-value is 0.053. The assumption of homogeneity is not rejected with an alpha level of 0.05 but will be rejected at an alpha level of 0.1.

## 1.2 With moderators

Load an hypothetical dataset `hypo_moder` first:

```
hypo_moder <- boot.heterogeneity:::hypo_moder
```

Three moderators (cov.z1, cov.z2, cov.z3) are included:

```
head(hypo_moder)
#>     n1  n2          d       cov.z1      cov.z2     cov.z3
#> 1   59  65  0.8131324 -0.005767173  0.80418951  1.2383041
#> 2  166 165  1.0243732  2.404653389 -0.05710677 -0.2793463
#> 3   68  68  1.5954236  0.763593461  0.50360797  1.7579031
#> 4   44  31  0.6809888 -0.799009249  1.08576936  0.5607461
#> 5   98  95 -1.3017946 -1.147657009 -0.69095384 -0.4527840
#> 6   44  31 -1.9398508 -0.289461574 -1.28459935 -0.8320433
```

Again, run the heterogeneity test using `boot.d()` with all moderators included in a matrix `mods` and model type specified as `model = 'mixed'`:

```
boot.run3 <- boot.d(n1 = hypo_moder$n1,
                    n2 = hypo_moder$n2,
                    est = hypo_moder$d,
                    model = 'mixed',
                    mods = cbind(hypo_moder$cov.z1, hypo_moder$cov.z2, hypo_moder$cov.z3),
                    p_cut = 0.05)
```

The results in `boot.run3` will in the same format as `boot.run` and `boot.run2`:

```
boot.run3
#>                  stat      p_value   Heterogeneity
#> Qtest       31.849952   0.000806            sig
#> boot.REML    9.283428   0.000400            sig
```

In the presence of moderators, the function above tests whether the variability in the true standardized mean differences after accounting for the moderators included in the model is larger than sampling variability alone (Viechtbauer, 2010).

- In the first line, the Q-statistic is $Q(df = 11) = 31.85$ and the associated p-value is 0.0008. This statistic is significant (sig) at an alpha level of 0.05, meaning that the true effect sizes after accounting for the moderators are heterogeneous.

- In the second line, the B-REML-LR statistic is 9.28 and the bootstrap-based p-value is 0.0004. This means that the true effect sizes after accounting for the moderators are heterogeneous at an alpha level of 0.05.

**For the following two examples (Fisher-transformed Pearson's correlations r; log odds ratio OR), no moderators are included, but one can simply include moderators as in section 1.2.**

**2. Fisher-transformed Pearson's correlations (r)**

`boot.fcor()` is the function to test the between-study heterogeneity in meta-analysis of Fisher-transformed Pearson's correlations (r).

**2.1 Heterogeneity magnitude test: lambda=0**

Load the example dataset `sensation` first:

```
sensation <- boot.heterogeneity:::sensation
```

Extract the required arguments from `sensation`:

```
# n is a list of samples sizes
n <- sensation$n
# Pearson's correlation
r <- sensation$r
# Fisher's Transformation
z <- 1/2*log((1+r)/(1-r))
```

Run the heterogeneity test using `boot.fcor()`:

```
boot.run.cor <- boot.fcor(n, z, model = 'random', p_cut = 0.05)
```

The test of between-study heterogeneity has the following results:

```
boot.run.cor
#>                  stat       p_value     Heterogeneity
#> Qtest       29.060970   0.00385868              sig
#> boot.REML    6.133111   0.00400882              sig
```

- In the first line, the Q-statistic is $Q(df = 12) = 29.06$ and the associated p-value is 0.004. This statistic is significant (sig) at an alpha level of 0.05, meaning that the true effect sizes are heterogeneous.

- In the second line, the B-REML-LR statistic is 6.13 and the bootstrap-based p-value is 0.004. This means that the true effect sizes are heterogeneous at an alpha level of 0.05.

## 2.2 Heterogeneity magnitude test: lambda=0.08

Run the heterogeneity test using `boot.fcor()`:

```
boot.run.cor2 <- boot.fcor(n, z, lambda=0.08, model = 'random', p_cut = 0.05)
```

The test of between-study heterogeneity has the following results:

```
boot.run.cor2
#>                  stat       p_value     Heterogeneity
#> boot.REML     2.42325    0.04607372              sig
```

- When lambda=0.08, the alternative hypothesis is that the magnitude of the between-study heterogeneity is larger than 0.08. Here the B-REML-LR statistic is 2.42 and the bootstrap-based p-value is 0.046. The null hypothesis is rejected in favor of the alternative hypothesis. This means that the true effect sizes are heterogeneous and the magnitude of the between-study heterogeneity is significantly larger than 0.08 at an alpha level of 0.05.

## 3. Log odds ratio (OR)

### 3.1 Without parallel implementation

`boot.lnOR()` is the function to test the between-study heterogeneity in meta-analysis of Natural-logarithm-transformed odds ratio (OR).

Load the example dataset `smoking` from R package `HSAUR3`:

```
library(HSAUR3)
data(smoking)
```

Extract the required arguments from `smoking`:

```
# Y1: receive treatment; Y2: stop smoking
n_00 <- smoking$tc - smoking$qc  # not receive treatement yet not stop smoking
n_01 <- smoking$qc # not receive treatement but stop smoking
n_10 <- smoking$tt - smoking$qt # receive treatement but not stop smoking
n_11 <- smoking$qt # receive treatement and stop smoking
```

The log odds ratios can be computed, but they are not needed by `boot.lnOR()`:

```
lnOR <- log(n_11*n_00/n_01/n_10)
lnOR
#>  [1]  0.6151856 -0.0235305  0.5658078  0.4274440  1.0814445  0.9109288
#>  [7]  0.9647431  0.7103890  1.0375520 -0.1407277  0.7747272  1.7924180
#> [13]  1.2021192  0.3607987  0.2876821  0.2110139  1.2591392  0.1549774
#> [19]  1.3411739  0.2963470  0.6116721  0.3786539  0.5389965  0.7532417
#> [25]  0.5653138  0.3786539
```

Run the heterogeneity test using `boot.lnOR()`:

```
boot.run.lnOR <- boot.lnOR(n_00, n_01, n_10, n_11, model = 'random', p_cut = 0.05)
```

The test of between-study heterogeneity has the following results:

```
boot.run.lnOR
#>                stat     p_value    Heterogeneity
#> Qtest       34.873957  0.09050857             n.s
#> boot.REML    3.071329  0.03706729             sig
```

- In the first line, the Q-statistic is $Q(df = 25) = 34.87$ and the associated p-value is 0.091. This statistic is not significant (n.s) at an alpha level of 0.05, meaning that the assumption of homogeneity cannot be rejected.

- In the second line, the B-REML-LR statistic is 3.07 and the bootstrap-based p-value is 0.037. This means that the assumption of homogeneity is rejected and the true effect sizes are heterogeneous at an alpha level of 0.05.

### 3.2 With parallel implementation

Run the heterogeneity test using `boot.lnOR()` with parallel computing and 4 cores:

```
boot.run.lnOR2 <- boot.lnOR(n_00, n_01, n_10, n_11, model = 'random', p_cut = 0.05,
                            parallel = TRUE, cores = 4)
```

The test of between-study heterogeneity has the same results as those in 3.1:

```
boot.run.lnOR2
#/======================================================/ 100%, Elapsed 00:41
#>                stat     p_value    Heterogeneity
#> Qtest       34.873957  0.09050857             n.s
#> boot.REML    3.071329  0.03706729             sig
```

```
sessionInfo()
#> R version 4.0.3 (2020-10-10)
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS Big Sur 10.16
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] tools     parallel  stats     graphics  grDevices utils     datasets
#> [8] methods   base
#>
#> other attached packages:
#> [1] rmarkdown_2.6           knitr_1.30               HSAUR3_1.0-10
#> [4] pbmcapply_1.5.0         metafor_2.4-0            Matrix_1.3-0
#> [7] boot.heterogeneity_0.1.2
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.5         compiler_4.0.3     prettyunits_1.1.1 remotes_2.2.0
#>  [5] testthat_3.0.1     digest_0.6.27      pkgbuild_1.2.0     pkgload_1.1.0
#>  [9] evaluate_0.14      nlme_3.1-151       lattice_0.20-41    memoise_1.1.0
#> [13] lifecycle_0.2.0    rlang_0.4.9        cli_2.2.0          rstudioapi_0.13
#> [17] yaml_2.2.1         xfun_0.19          withr_2.3.0        stringr_1.4.0
```

```
#> [21] roxygen2_7.1.1    xml2_1.3.2        desc_1.2.0        fs_1.5.0
#> [25] devtools_2.3.2    rprojroot_2.0.2   grid_4.0.3        glue_1.4.2
#> [29] R6_2.5.0          processx_3.4.5    fansi_0.4.1       sessioninfo_1.1.1
#> [33] callr_3.5.1       purrr_0.3.4       magrittr_2.0.1    htmltools_0.5.0
#> [37] ps_1.5.0          ellipsis_0.3.1    usethis_2.0.0     assertthat_0.2.1
#> [41] stringi_1.5.3     crayon_1.3.4
```