Simulation protocol:

Comparison of confidence intervals summarizing the uncertainty of the combined estimate of a meta-analysis

Leonhard Held, Felix Hofmann

For the present protocol is inspired by Burton et al. (2006) and Morris et al. (2019).
The simulation is implemented in `simulate_all.R`.

Contents

1 Aims and objectives

The aim of this simulation study is the comparison of confidence intervals (CIs) summarizing the uncertainty of the combined estimate of a meta-analysis. Specifically, we focus on CIs constructed using p-value functions that implement the p-value combination methods from Edgington (1972) and Fisher (1932). The underlying data sets are simulated as described in Section 2. In Section 3 we describe which CI construction methods we compare in this simulation study and what criteria we use to evaluate them.

2 Simulation of the data sets

2.1 Allowance for failures

We expect no failures, i. e. , for all simulated data sets all type of CI methods should lead to a valid CI and all valid CIs should lead to valid CI criteria. If a failure occurs, we stop the simulation and investigate the reason for the failure.

2.2 Software to perform simulations

The simulation study is performed using the statistical software R (R Core Team, 2021). We save the output of `sessionInfo()` giving information on the used version of R, packages, and platform with the simulation results.

2.3 Random number generator

We use the package *doRNG* with its default random number generator to ensure that random numbers generated inside parallel for loops are independent and reproducible.

2.4 Scenarios to be investigated

The 720 simulated scenarios consist of all combinations of the following parameters:

- Higgin's I^2 heterogeneity measure $\in \{0, 0.3, 0.6, 0.9\}$.
- Number of studies summarized by the meta-analysis $k \in \{3, 5, 10, 20, 50\}$.
- Publication bias is $\in \{\text{'none'}, \text{'moderate'}, \text{'strong'}\}$ following the terminology of Henmi and Copas (2010).
- The average study effect $\theta \in \{0.2, 0.5\}$.
- The distribution to draw the true study values δ_i is either 'Gaussian' or 't' with 4 degrees of freedom. The latter still has finite mean and variance, but leads to more 'outliers'.
- The sample size n_i of the i-th study (number of patients per study) is $n_i = 50$ (small study) except for 0, 1, or 2 studies where $n_i = 500$ (large study).

Note that IntHout et al. (2014) use a similar setup.

2.5 Simulation details

The simulation of one meta-analysis data set is performed as follows:

1. Compute the within-study variance

$$\epsilon^2 = \frac{2}{k} \sum_{i=1}^{k} \frac{1}{n_i}. \tag{1}$$

2. Compute the between-study variance

$$\tau^2 = \epsilon^2 \frac{I^2}{1 - I^2}. \tag{2}$$

3. For a trial i of the meta-analysis with k trials, $i = 1, \ldots, k$:

   (a) Simulate the true effect size using the Gaussian model: $\delta_i \sim \mathcal{N}(\theta, \tau^2)$ or using a Student-t distribution with 4 degrees of freedom such that the samples have mean θ and variance τ^2.

   (b) Simulate the effect estimates of each trial $y_i \sim \mathcal{N}(\delta_i, \frac{2}{n_i})$.

   (c) Simulate the standard errors of the trial outcomes: $\mathrm{se}_i \sim \sqrt{\frac{\chi^2(2n_i-2)}{(n_i-1)n_i}}$.

Note: The marginal variance
The marginal variance of this simulation procedure is τ^2+2/n_i, so follows the additive heterogeneity model as intended.

Note: Publication bias
To simulate studies under **publication bias**, we follow the suggestion of Henmi and Copas (2010) and accept each simulated study with probability

$$\exp(-4\,\Phi(-y_i/\mathrm{se}_i)^\gamma), \tag{3}$$

where $\gamma = 3$ and $\gamma = 1.5$ correspond to *moderate* and *strong* publication bias, respectively. This is, accepted studies are kept and for a rejected study we replace y_i and se_i by newly simulated values, which are then again accepted with the given probability above. This procedure is repeated until the required number of studies is simulated.
However, we assume that only small studies with $n_i = 50$ are subject to publication bias. Thus, larger studies with $n_i = 500$ are always accepted. As described in Section 2.4, we set $\theta \in \{0.2, 0.5\}$. See the R function `simREbias()`.

2.6 Simulation procedure

For each scenario in Section 2.4 we

1. simulate 10'000 meta-analysis data sets

2. compute the CIs listed in Section 3.1 for each meta-analysis

3. summarize the performance of the CIs by the criteria listed in Section 3.3

3 Analysis of the confidence intervals

This section contains an overview over the construction methods for CIs that we consider in this simulation. Moreover, we explain what measures we use in order to compare the different CIs with each other.

3.1 Construction methods for confidence intervals

For this project, we will calculate 95% CIs according to the following methods.

1. Hartung-Knapp-Sidik-Jonkman (HK) (IntHout et al., 2014).
2. Random effects model.
3. Henmi and Copas (HC) (Henmi and Copas, 2010).
4. Bayesian random effects meta analysis (Bayesmeta) with half-normal prior distribution with scale parameter $\sigma = 0.3$ for τ, the square root between-study heterogeneity (Röver, 2020; Lilienthal et al., 2023).
5. Edgington's method (Edgington, 1972).
6. Fisher's method (Fisher, 1932).

3.2 Definition of the variance adjustments

As we assume an additive heterogeneity model, we will calculate the confidence intervals for methods *Fisher*, *Edgington*, and *Random effects* based on the following estimators for the between-study variance τ^2. The estimator acts thus as an additional scenario that is only applied to the above mentioned methods.

1. No heterogeneity, i. e. $\tau^2 = 0$.
2. DerSimonian-Laird (DerSimonian and Laird, 1986).
3. Paule-Mandel (Paule and Mandel, 1982).
4. REML (Harville, 1977).

The calculation of the estimates in the simulation will be done using the `metagen` function from the `R` package "*meta*".

The adjusted study-specific standard errors are then given by $\text{se}_{\text{adj}}(\hat{\theta}_i) = \sqrt{\text{se}(\hat{\theta}_i)^2 + \tau^2}$.

3.3 Measures considered

We assess the CIs using the following criteria

1. CI coverage of combined effect, i. e. , the proportion of intervals containing the true effect. If the CI does not exist given a specific simulated data set, we treat the coverage as as missing (`NA`).

2. CI width. If there is more than one interval, the width is the sum of the lengths of the individual intervals. If the interval does not exist for a simulated data set, the width will be recorded as missing (`NA`).

3. Interval score (Gneiting and Raftery, 2007). If the interval does not exist for a simulated data set, the score will be recorded as missing (`NA`).

4. Number of CIs (only for Fisher and Edgington methods). If the interval does not exist for a simulated data set, the number of CIs will be recorded as 0.

Furthermore, we calculate the following measure related to the point estimates

1. Mean squared error (MSE).

For the Edgington and Fisher methods, we also investigate the distribution of the highest value of the p-value function between the lowest and the highest treatment effect of the simulated studies. In order to do so, we calculate the following measures:

- Minimum
- First quartile
- Mean
- Median
- Third quartile
- Maximum

As both methods can result in more than one CI for a given meta-analysis, we record the relative frequency of the number of intervals m over the 10'000 iterations for each of the different scenarios mentioned in Section 2.4. However, we truncate the distribution by summarising all events where the number of intervals is > 9.

4 Estimates to be stored for each simulation and summary measures to be calculated over all simulations

For each simulated meta-analysis we construct CIs according to all methods (Section 3.1) and calculate all available assessments (Section 3.3) for the respective method. For assessments 1-3 in Subsection 3.3 we only store the mean value of all the 10'000 iterations in a specific scenario. Possible missing values (`NA`) are removed before calculating the mean value. Regarding the distribution of the highest value of the p-value function, we store the summary measures mentioned in the respective paragraph of Subsection 3.3. We calculate the relative frequencies of the number of intervals $m = 1, 2, \dots, 9, > 9$ in each confidence set over the 10'000 iterations of the same scenario.

5 Presentation of the simulation results

For each of the performance measures 1-3 in Subsection 3.3 as well as the mean squared error (MSE) we construct plots with

- the number of studies k on the x-axis
- the performance measure on the y-axis
- one connecting line and color for each value of I^2
- one panel for each CI method

Regarding the distribution of the p-value function for the harmonic mean and k-trials methods, we will create plots that contain

- the number of studies k on the x-axis
- the value of the summary statistic on the y-axis
- one connecting line and color for each summary statistic
- one panel for each CI method

The plots for the relative frequencies of the number of intervals have

- the category (1 to 9 and > 9) indicating the number of intervals n on the x-axis
- the relative frequency on the y-axis
- a bar for each category indicating the relative frequency for the respective category
- one panel for each CI method

References

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292. 1

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188. 4

Edgington, E. S. (1972). An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology*, 80(2):351–363. Publisher: Routledge _eprint: https://doi.org/10.1080/00223980.1972.9924813. 2, 4

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 4 edition. 2, 4

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 5

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358):320–338. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. 4

Henmi, M. and Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, 29(29):2969–2983. 2, 3, 4

IntHout, J., Ioannidis, J. P., and Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14(25). 2, 4

Lilienthal, J., Sturtz, S., Schürmann, C., Maiworm, M., Röver, C., Friede, T., and Bender, R. (2023). Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. submitted for publication. 4

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102. 1

Paule, R. C. and Mandel, J. (1982). Consensus Values and Weighting Factors. *Journal of Research of the National Bureau of Standards (1977)*, 87(5):377–385. 4

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2

Röver, C. (2020). Bayesian Random-Effects Meta-Analysis Using the bayesmeta R Package. *Journal of Statistical Software*, 93:1–51. 4