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 methods from Edgington (1972) and Fisher (1934). The underlying data sets are simulated as described in Section 2 and Section 2.4. The resulting intervals are then compared to CIs constructed using the other methods listed in Section 3.1 using the measures defined in Section 3.4.

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:

update number

- Higgin's I^2 heterogeneity measure $\in \{0, 0.3, 0.6, 0.9\}$.
- We always use an additive heterogeneity model.
- Number of studies summarized by the meta-analysis $k \in \{3, 5, 10, 20, 50\}$.
- Publication bias is $\in$ {'none', 'moderate', '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).

Maybe remove this entirely?

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: $\text{se}_i \sim \sqrt{\frac{\chi^2(2n_i-2)}{(n_i-1)n_i}}$.

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
$$\exp(-4\,\Phi(-y_i/\text{se}_i)^\gamma), \tag{3}$$
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.
To obtain a similar scenario as in Henmi and Copas (2010) we set
$$\theta/\sqrt{2/n_i} \stackrel{!}{=} 1 \Rightarrow \theta = \sqrt{2/n_i}$$
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()`.

Where does this come from? Does this still apply when $\theta = 0.5$?

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 3.4

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 (with REML estimate of the heterogeneity variance).
3. Henmi and Copas (HC) (Henmi and Copas, 2010).
4. Harmonic mean analysis with alternative `none` (Held, 2020) and without variance adjustment.
5. Harmonic mean analysis with alternative `none`, additive variance adjustment with $\hat{\tau}^2$. An extension of the idea in Held (2020).
6. Harmonic mean analysis with alternative `none`, multiplicative variance adjustment (Mawdsley et al., 2017).
7. k-trials rule with alternative `none` and without variance adjustment.
8. k-trials rule with alternative `none`, additive variance adjustment with $\hat{\tau}^2$.
9. k-trials rule with alternative `none`, multiplicative variance adjustment.

3.2 Definition of the k-trials rule

Similar to the harmonic mean method, the k-trials rule takes a mean value under the null hypothesis μ_0 as well as effect estimates $\hat{\theta}_i, i = 1, \ldots, k$ and the corresponding standard errors $\text{se}(\hat{\theta}_i)$ from k different studies as input and calculates the resulting p-value according to Equation 4.

$$p(\mu_0) = \max\left(\Phi\left(\frac{\hat{\theta}_i - \mu_0}{\text{se}(\hat{\theta}_i)}\right)\right)^k \tag{4}$$

As the effect estimates $\hat{\theta}_i$ and the corresponding standard errors $\text{se}(\hat{\theta}_i)$ are usually given in the context of meta-analyses, the above p-value function only depends on μ_0. Therefore, CI limits are computed by searching for those values of μ_0 for which $p(\mu_0) = 0.05$. This may result in confidence sets containing more than one confidence interval.
In case of variance adjustments, the term $\text{se}(\hat{\theta}_i)$ in Equation 4 is replaced with $\text{se}_{\text{adj}}(\hat{\theta}_i)$, which is defined in Subsection 3.3.

3.3 Definition of the variance adjustments

As stated in Subsection 3.1, the harmonic mean and k-trials methods can be extended such that heterogeneity between the individual studies is taken into account. In scenarios where the additive variance adjustment is used, we estimate the between study variance τ^2 using the REML method implemented in the `metagen` `R`-package "meta" and adjust the study-specific standard errors such that $\text{se}_{\text{adj}}(\hat{\theta}_i) = \sqrt{\text{se}(\hat{\theta}_i)^2 + \tau^2}$.
In case of the multiplicative variance adjustment, we estimate the multiplicative parameter ϕ as described in Mawdsley et al. (2017) and adjust the study-specific standard errors such that $\text{se}_{\text{adj}}(\hat{\theta}_i) = \text{se}(\hat{\theta}_i) \cdot \sqrt{\phi}$.

3.4 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
2. CI coverage of study effects, i. e. , the proportion of intervals containing the true study-specific effects
3. CI coverage of all study effects, i. e. , whether or not the CI covers all of the study effects
4. CI coverage of at least one of the study effects, i. e. , whether or not the CI covers at least one of the study effects
5. Prediction Interval (PI) coverage, i. e. , the proportion of intervals containing the treatment effect of a newly simulated study. The newly simulated study has $n = 50$ and is not subject to publication bias. All other simulation parameters stay the same as for the simulation of the original studies (only for Harmonic mean, k-trials, REML, and HK methods)
6. CI width (Corresponds to the sum the width of the individual intervals in case of more than one interval)
7. Interval score (Gneiting and Raftery, 2007)
8. Number of CIs (only for Harmonic mean and k-trials methods).

For the Harmonic mean and k-trials methods, we also investigate the distribution of the lowest 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, harmonic mean and k-trials, 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 3.4) for the respective method. For assessments 1-8 in Subsection 3.4 we only store the mean value of all the 10'000 iterations in a specific scenario. Regarding the distribution of the lowest value of the p-value function, we store the summary measures mentioned in the respective paragraph of Subsection 3.4. We calculate the relative frequencies of the number of intervals $m = 1, 2, \ldots, 9, > 9$ 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 1-8 in Subsection 3.4 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

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

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 4 edition. 2

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

Held, L. (2020). The harmonic mean χ^2-test to substantiate scientific findings. *Journal of the Royal Statistical Society Series C*, 69(3):697–708. 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

Mawdsley, D., Higgins, J. P. T., Sutton, A. J., and Abrams, K. R. (2017). Accounting for heterogeneity in meta-analysis using a multiplicative model—an empirical study. *Research Synthesis Methods*, 8(1):43–52. 4, 5

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

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