Simulation protocol:

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

Florian Gerber, Leonhard Held, Lisa Hofer, Philip Heesen

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

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

1. Considered methods are
   (a) DerSimonian and Laird (DL) [DerSimonian and Laird, 1986]
   (b) Hartung Knapp Sidik Jonkman (HK) [IntHout et al., 2014]
   (c) Harmonic mean analysis with alternative `none` [Held, 2020]
   (d) Harmonic mean analysis with alternative `two.sided` [Held, 2020]
   (e) Harmonic mean analysis with alternative `none`, additive variance adjustment with $\hat{I}^2$. An extension of the idea in [Held, 2020].
   (f) Harmonic mean analysis with alternative `none`, multiplicative variance adjustment [Mawdsley et al., 2017]. An extension of the idea in [Held, 2020].
   (g) Henmi and Copas (HC) [Henmi and Copas, 2010].
2. We assess the CIs using the following criteria
   (a) CI coverage of combined effect, i. e. , the proportion of intervals containing the true effect
   (b) CI coverage of study effects, i. e. , the proportion of intervals containing the true study specific effect
   (c) CI width
   (d) CI score [Gneiting and Raftery, 2007]
   (e) number of CIs (only for Harmonic mean method with alternative `none`).

2 Simulation procedures

2.1 Allowance for failures

We expect no failures, i. e. , for all simulated datasets 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.

3 Scenarios to be investigated

The 360 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 $k \in \{2, 3, 5, 10, 20\}$.
- Publication bias is $\in$ {'none', 'moderate', strong'} following the terminology of [Henmi and Copas, 2010]. The average study effect also influences the publication bias, and we set it to $\theta = 0.2$ to obtain a similar scenario as used in [Henmi and Copas, 2010].
- The distribution to draw the true study values δ_i is either 'Gaussian' or 't'. The latter leads to more 'outliers'.
- Increase the number of patients for 0, 1, or 2 studies by a factor 10.

Sample size of the individual studies (number of patients per study) is fixed to $n = 50$. Note that [IntHout et al., 2014] use a similar setup.

3.1 Simulation of the data

For each scenario in Section 3 we

1. simulate 10000 meta-analysis datasets
2. compute the CI's listed in Section 1 for each meta-analysis
3. summarize the performance of the CI's by the criteria listed in Section 1

For the **Gaussian model without publication bias**, the simulation of one meta-analysis dataset is performed as follows:

1. Compute the within-study variance $\epsilon^2 = \frac{2}{n}$.
2. Compute the between-study variance $\tau^2 = \epsilon^2 \frac{I^2}{1-I^2}$.
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 students-t distribution such that the samples have mean θ and variance τ^2.
   (b) Simulate the effect estimates of each trial $y_i \sim \mathcal{N}(\delta_i, \frac{2}{n})$.
   (c) $\text{se}_i \sim \sqrt{\frac{\chi^2(2n-2)}{(n-1)n}}$ are the standard errors of trial outcomes

Note: Student's t instead of Gaussian
To generate more 'outlier' studies, we use the Student's t instead of the Gaussian distribution to simulate the true study effects δ_i. We use the Student's t with 4 degrees of freedom and choose the other parameter such that the variance of the samples is τ^2.

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

$$\exp(-4\,\Phi(-\theta_i/\text{se}_i)^\gamma),$$

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 θ_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.
The mean study effect θ and the sample size n have an influence on the acceptance probability. To obtain a similar scenario as in [Henmi and Copas, 2010] we set

$$\theta/\sqrt{2/n} \stackrel{!}{=} 1 \Rightarrow \theta = \sqrt{2/n}$$

This is, for $n = 50$ we use $\theta = 0.2$. See the R function `simREbias()`.

Note: Unbalanced sample sizes
To study the effect of unbalanced sample sizes we consider the following setup:

1. Increase the sample size of **one** of the k by a factor 10.
2. Increase the sample size of **two** of the k by a factor 10.

A possible publication bias is only applied to the small studies. See the argument `large` of `simREbias()`.

4 Statistical methods to be evaluated

1. DerSimonian and Laird (DL) [DerSimonian and Laird, 1986]
2. Hartung Knapp Sidik Jonkman (HK) [IntHout et al., 2014]
3. Harmonic mean analysis with alternative `none` [Held, 2020]
4. Harmonic mean analysis with alternative `two.sided` [Held, 2020]
5. Harmonic mean analysis with alternative `none`, additive variance adjustment with $\hat{I}^2$. An extension of the idea in [Held, 2020].
6. Harmonic mean analysis with alternative `none`, multiplicative variance adjustment [Mawdsley et al., 2017]. An extension of the idea in [Held, 2020].
7. Henmi and Copas (HC) [Henmi and Copas, 2010].

Extension: Multiplicative variance adjustment
For the harmonic mean analysis with alternative `none`, we also try a variant with multiplicative instead of additive variance adjustment [Mawdsley et al., 2017]. (Note: We think in [Mawdsley et al., 2017] v_i and v_i^2 are confused.)

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

For each simulated meta-analysis we store the assessments (Section 4) of all the CI methods (Section 4). Then we compute the average performance by taking the mean.

6 Number of simulations to be performed

We perform the simulation of each scenario 10000 times. We repeat the final simulation several times to assess the variability in the simulation results.

7 Criteria to evaluate the performance of statistical methods for different scenarios

Assess the CIs by 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 effect
3. CI width
4. CI score [Gneiting and Raftery, 2007]
5. number of CIs (only for Harmonic mean method with alternative `none`).

8 Presentation of the simulation results

We present the average of each performance measurement in figures. The figures have

- 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

References

[Burton et al., 2006] 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 and Laird, 1986] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188. 2, 4

[Gneiting and Raftery, 2007] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 2, 4

[Held, 2020] 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. 2, 4

[Henmi and Copas, 2010] 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 et al., 2014] 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 et al., 2017] 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. 2, 4

[Morris et al., 2019] 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 Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2