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); Wilkinson (1951); Pearson (1933); Tippett (1931); 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* (Gaujoux, 2023) 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 1080 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 $\in$ {'none', 'moderate', 'strong'} following the terminology of Henmi and Copas (2010).
- The average study effect $\theta \in \{0.1, 0.2, 0.5\}$.
- The distribution to draw the true study values δ_i is either 'Gaussian' or 'sn' ("skew-normal").
- 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, but do not use a skew-normal distribution for the δ_i's.

2.5 Simulation details

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

1. Compute the within-study variance

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

   For $n_i = 50$ we obtain $\epsilon^2 = 1/25$.

2. Compute the between-study variance

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

   For $I^2 \in \{0.0, 0.3, 0.6, 0.9\}$ and $\epsilon^2 = 1/25$ we obtain $\tau^2 \in \{0, 0.017, 0.06, 0.36\}$.

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 skew-normal distribution. The skew-normal distribution has three parameters. The shape parameter α is usually transformed to $\delta = \alpha/\sqrt{1+\alpha^2} \in (-1, 1)$. We use $\alpha = 8$ where $\delta \approx 0.992$. The other two parameter are

   - scale $\omega = \tau/\sqrt{1 - 2\delta^2/\pi}$
   - location $\xi = \theta - \omega\delta\sqrt{2/\pi}$

   to obtain mean θ and variance τ^2. Figure 1 compares Gaussian and skew-normal distribution for $\alpha = \pm 8$, where the skewness coefficient is ± 0.93.

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

Perhaps use skew-normal with median 0.2 (rather than mean)?

Note: The marginal variance
The marginal variance of the effect estimates y_i is $\tau^2 + 2/n_i$, 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.

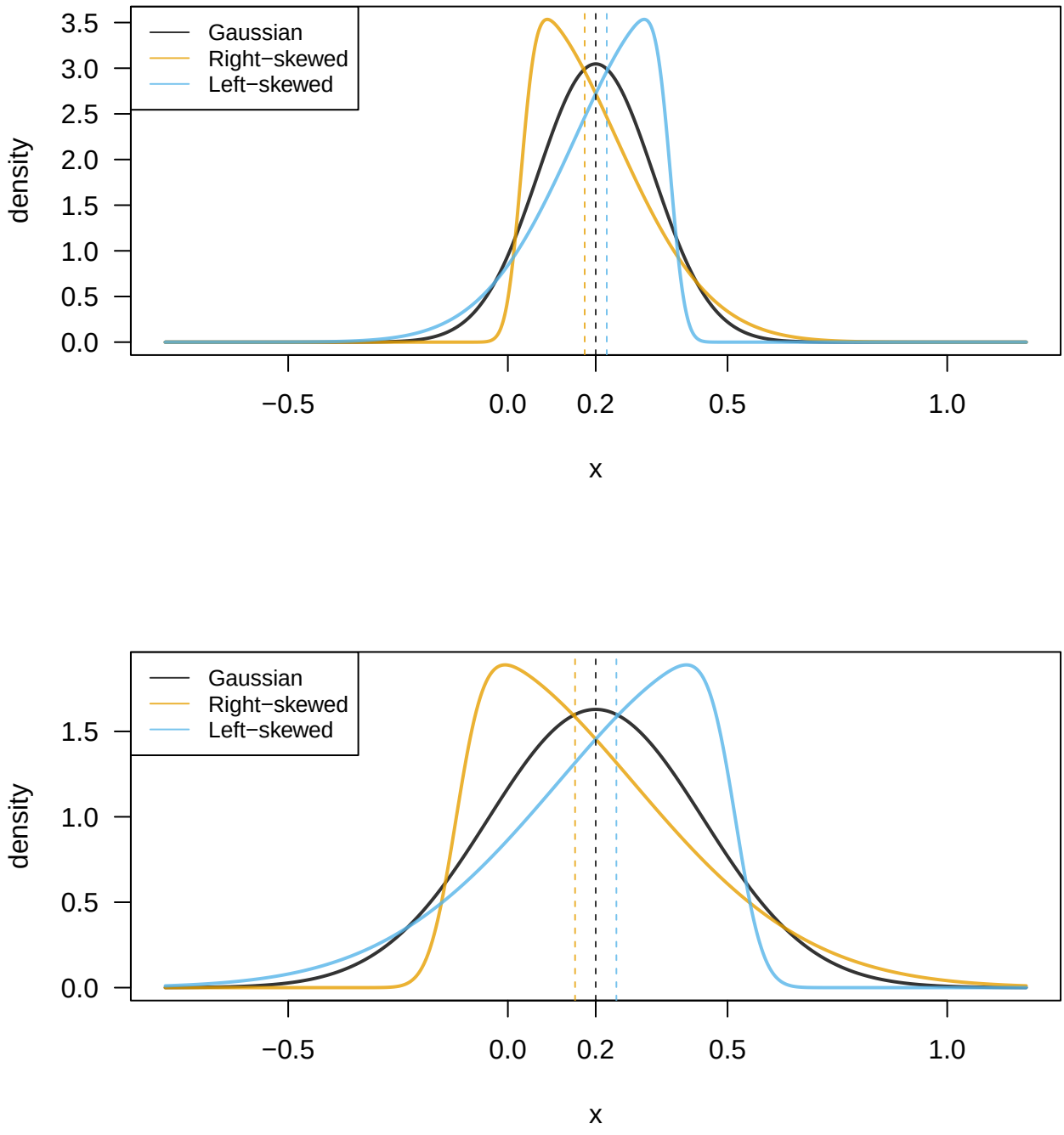

Figure 1: Gaussian and skew-normal distribution with mean $\theta = 0.2$ and variance $\tau^2 = 0.131^2$ (top) and $\tau^2 = 0.245^2$ (bottom). The median of each distribution is indicated with a dashed lined.

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 $\theta \in \{0.1, 0.2, 0.5\}$. See the R function `simREbias()`.
In order to check how this implementation of publication bias impacts the simulation performance, we keep track of the mean acceptance probability for each simulation scenario that is subject to publication bias. For the calculation of the mean, we also consider large studies with $n = 500$. Since such studies are not subject to publication bias, they have an acceptance probability of 1.

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. Edgington's method (Edgington, 1972).
5. Wilkinson's method (Wilkinson, 1951).
6. Pearson's method (Pearson, 1933).
7. Tippett's method (Tippett, 1931).
8. Fisher's method (Fisher, 1932).

3.2 Definition of the variance estimates

As we assume an additive heterogeneity model, we will calculate the confidence intervals and *Random effects* based on a suitable (DerSimonian and Laird, 1986) and the REML estimate of the between-study variance τ^2. The following estimates act thus as an additional scenario that is applied to all methods (except Hartung-Knapp and Henmi-Copas):

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

Calculation will be done using the `metagen` function from the `R` package *meta* (Balduzzi et al., 2019).

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

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 θ. 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).

Furthermore, we calculate the following measures related to the point estimates:

1. Mean squared error (MSE) of the estimator.
2. Bias of the estimator.
3. Variance of the estimator.

We also calculate the following measures related to skewness γ of the observed study effects and the skewness β of the 95% confidence interval. The first quantity γ is defined as Fisher's skewness coefficient of the observed study effects $\hat{\theta}_i = y_i$, weighted with the inverse squared standard errors $w_i = 1/\mathrm{se}^2_{\mathrm{adj}}(\hat{\theta}_i) = 1/(\mathrm{se}(\hat{\theta}_i)^2 + \tau^2)$. The skewness β of the 95% confidence interval $[\mathrm{lower}, \mathrm{upper}]$ is calculated as

$$\beta = \frac{\mathrm{upper} + \mathrm{lower} - 2\,\mathrm{estimate}}{\mathrm{upper} - \mathrm{lower}}.$$

We then calculate the following measures related to skewness:

1. Correlation between the (weighted) skewness coefficient γ of the observed study effects and the skewness β of the confidence interval.
2. Agreement of the sign of the (weighted) skewness coefficient γ of the observed study effects and the sign of the skewness β of the confidence interval. This will be quantified with Cohen's κ.

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.
Furthermore, we store the mean of the average acceptance probability in each of the 10'000 iterations for all simulation scenarios where there is either 'modest' or 'strong' publication bias.

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), bias, and variance 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 *Edgington* and *Fisher* 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

Balduzzi, S., Rücker, G., and Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health*, (22):153–160. 6

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. 6

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, 5

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 4 edition. 2, 5

Gaujoux, R. (2023). *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*. R package version 1.8.6. 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. 6

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.]. 6

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, 5

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, 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

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25:379–410. 2, 5

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

Tippett, L. H. C. (1931). *Methods of Statistics*. Williams Norgate. 2, 5

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, 48:156–158. 2, 5