

# Bayesian approaches to designing replication studies

## Supplementary materials

Samuel Pawel<sup>\*</sup>, Guido Consonni<sup>†</sup>, and Leonhard Held<sup>\*</sup>

<sup>\*</sup> Department of Biostatistics, University of Zurich

<sup>†</sup> Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore

E-mail: samuel.pawel@uzh.ch

October 10, 2022

In this document we provide additional information on methods for analyzing replication data. For each method we derive the success regions which are required for sample size determination of the replication study. For the two-trials rule and the replication Bayes factor provide additional derivation on how these methods can be generalized to the multisite replication setting.

## 1 The two-trials rule

Assume the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  where  $\hat{\theta}_i$  is an estimate of the effect size  $\theta$  from study  $i$  and  $\sigma_i$  is the corresponding standard error (assumed to be known). The  $p$ -value for testing  $H_0: \theta = 0$  versus  $H_1: \theta > 0$  is then given by  $p_i = 1 - \Phi(\hat{\theta}_i/\sigma_i)$ . Suppose the original effect estimate was positive ( $\hat{\theta}_o > 0$ ) and statistically significant at some level  $\alpha$ , i. e.,  $p_o \leq \alpha$ . Replication success at level  $\alpha$  is then established if also the replication effect estimate  $\hat{\theta}_r$  is positive and statistically significant at level  $\alpha$ , i. e.,  $p_r \leq \alpha$ . This implies that replication success is achieved if the replication effect estimate  $\hat{\theta}_r$  is contained in the success region

$$S_{2TR} = [z_\alpha \sigma_r, \infty).$$

Conversely, if the original effect estimate was negative ( $\hat{\theta}_o < 0$ ), the one-sided  $p$ -values are computed for the lower tail of the null distribution and the success region is therefore given by

$$S_{2TR} = [-\infty, -z_\alpha \sigma_r).$$

### 1.1 The multisite two-trials rule

If multiple replication studies of the same original study are conducted, the two-trials rule is usually modified in a way that the replication effect estimates are first synthesized using either fixed or random effects meta-analysis (see e. g., the “Many labs” projects from [Klein et al., 2014, 2018](#)). That is, a weighted average  $\hat{\theta}_{r*} = \{\sum_{i=1}^m \hat{\theta}_{ri}/(\sigma_{ri}^2 + \tau_r^2)\} / \sigma_{r*}^2$  with standard error  $\sigma_{r*} = 1/\sqrt{\{\sum_{i=1}^m 1/(\sigma_{ri}^2 + \tau_r^2)\}}$  is computed from the  $m$  replication effect estimates  $\hat{\theta}_{ri}$  and standard errors  $\sigma_{ri}$ . The between replication heterogeneity variance  $\tau_r^2$  can be either be set to zero (fixed effects) or estimated from the data. Assuming again that the Replication success at level  $\alpha$  is then established if the replication  $p$ -value is smaller than  $\alpha$ , i. e.,  $p_{r*} = 1 - \Phi(\hat{\theta}_{r*}/\sigma_{r*}) \leq \alpha$ . With some algebra one can show that this implies a success region

for the weighted average replication effect estimate  $\hat{\theta}_{r*}$  given by

$$S_{2TR} = [z_\alpha \sigma_{r*}, \infty).$$

for positive original effect estimates ( $\hat{\theta}_o > 0$ ) and

$$S_{2TR} = [-\infty, -z_\alpha \sigma_{r*})$$

for negative original effect estimates ( $\hat{\theta}_o < 0$ ).

## 2 Fixed effects meta-analysis

Assume again the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  where  $\hat{\theta}_i$  is an estimate of the effect size  $\theta$  from study  $i \in \{o, r\}$  and  $\sigma_i$  is the corresponding standard error (assumed to be known). In the fixed effects meta-analysis approach replicability is assessed in terms of the pooled effect estimate  $\hat{\theta}_m$  and standard error  $\sigma_m$  which are

$$\hat{\theta}_m = \left( \hat{\theta}_o / \sigma_o^2 + \hat{\theta}_r / \sigma_r^2 \right) \sigma_m^2 \quad \text{and} \quad \sigma_m = (1/\sigma_o^2 + 1/\sigma_r^2)^{-1/2},$$

which are also equivalent to the mean and standard deviation of a posterior distribution for the effect size  $\theta$  based on the data from original and replication study and an initial flat prior for  $\theta$ . Replication success at level  $\alpha$  is established if the one-sided meta-analytic  $p$ -value (in the direction of the original effect estimate  $\hat{\theta}$ ) is significant at level  $\alpha$ , i. e.,  $p_m = 1 - \Phi(\hat{\theta}_m / \sigma_m) \leq \alpha$  (assuming  $\hat{\theta}_o > 0$ ). With some algebraic manipulations one can show that this criterion implies a success region  $S_{MA}$  for the replication effect estimate  $\hat{\theta}_r$  given by

$$S_{MA} = \left[ \sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty \right),$$

respectively

$$S_{MA} = \left( -\infty, -\sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2 \right]$$

for negative original effect estimates ( $\hat{\theta}_o < 0$ ).

## 3 Effect size equivalence

Assume the data model  $\hat{\theta}_i | \theta_i \sim N(\theta_i, \sigma_i^2)$  for study  $i \in \{o, r\}$ . A  $(1 - \alpha)$  confidence interval for the effect size difference  $\delta = \theta_r - \theta_o$  is then given by

$$\hat{\theta}_r - \hat{\theta}_o \pm z_{\alpha/2} \sqrt{\sigma_r^2 + \sigma_o^2}.$$

Effect size equivalence is then established if the confidence interval is included in the equivalence region  $[-\Delta, \Delta]$  where  $\Delta > 0$  is a pre-specified margin. With some algebraic rearrangements one can show that the replication success criterion implies a success region  $S_E$  for the replication estimate  $\hat{\theta}_r$  given by

$$S_E = \left[ \hat{\theta}_o - \Delta + z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2}, \hat{\theta}_o + \Delta - z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2} \right].$$

## 4 The replication Bayes factor

Assume the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  for study  $i \in \{o, r\}$ . The replication Bayes factor uses the replication data  $x_r$  to contrast the null hypothesis  $H_0: \theta = 0$  to the alternative hypothesis  $H_1: \theta \sim f(\theta | \hat{\theta}_o, \sigma_o^2)$  postulating that the effect size  $\theta$  is distributed according to its posterior distribution based on the original data  $x_o = \{\hat{\theta}_o, \sigma_o\}$ . If under  $H_1$  a flat initial prior is specified for  $\theta$  and then updated by the original data  $x_o$ , the alternative is  $H_1: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$  so that the replication Bayes factor is given by

$$\text{BF}_R = \frac{f(\hat{\theta}_r | H_0)}{f(\hat{\theta}_r | H_1)} = \sqrt{1 + \sigma_o^2 / \sigma_r^2} \exp \left[ -\frac{1}{2} \left\{ \frac{\hat{\theta}_r^2}{\sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2} \right\} \right] \quad (1)$$

Replication success at some level  $\gamma \in (0, 1)$  is then achieved if  $\text{BF}_R \leq \gamma$ . With straightforward but tedious algebra one can show that this condition is equivalent to the replication effect estimate  $\hat{\theta}_r$  falling in the success region given by

$$S_{\text{BF}_R} = \left( -\infty, -\sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2 \right) \cup \left[ \sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty \right)$$

where  $A = \sigma_r^2 (1 + \sigma_r^2 / \sigma_o^2) \{ \hat{\theta}_o^2 / \sigma_o^2 - 2 \log \gamma + \log(1 + \sigma_o^2 / \sigma_r^2) \}$ .

### 4.1 The multisite replication Bayes factor

In the multisite setting the data are a vector of replication effect estimates  $\hat{\theta}_r = (\hat{\theta}_{r1}, \dots, \hat{\theta}_{rm})^\top$  with corresponding standard error vector  $\sigma_r = (\sigma_{r1}, \dots, \sigma_{rm})^\top$ . The sampling model is then  $\hat{\theta}_r | \theta \sim N_m\{\theta \mathbf{1}_m, \text{diag}(\sigma^2 + \tau_r^2 \mathbf{1}_m)\}$  where  $\mathbf{1}_m$  is a vector of  $m$  ones and  $\tau_r^2$  is a heterogeneity variance for the replication effect sizes (not to be confused with the heterogeneity variance  $\tau^2$  used in the design prior).

To contrast the hypothesis  $H_0: \theta = 0$  to the alternative  $H_1: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$  we need to derive the marginal density of the replication effect estimates  $\hat{\theta}_r | \theta \sim N_n\{\theta \mathbf{1}_m, \text{diag}(\sigma_r^2 + \tau_r^2 \mathbf{1}_m)\}$  under the alternative  $H_1: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$ . Let  $N(x; m, v)$  denote the normal density function mean  $m$  and variance  $v$  evaluated at  $x$ . Define also  $\hat{\theta}_{r*} = \left\{ \sum_{i=1}^n \hat{\theta}_{ri} / (\sigma_{ri}^2 + \tau_r^2) \right\} \sigma_{r*}^2$  and  $\sigma_{r*}^2 = 1 / \left\{ \sum_{i=1}^n 1 / (\sigma_{ri}^2 + \tau_r^2) \right\}$ , i. e., the weighted average of the replication effect estimates based on the heterogeneity  $\tau_r^2$  and its variance. The marginal density is then

$$\begin{aligned} f(\hat{\theta}_r | H_1) &= \int f(\hat{\theta}_r | \theta) f(\theta | H_1) d\theta \\ &= \int \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \theta)^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right]}{\left\{ 2\pi \sigma_o^2 \prod_{i=1}^n 2\pi (\sigma_{ri}^2 + \tau_r^2) \right\}^{1/2}} d\theta \\ &= \int \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right]}{\left\{ 2\pi \sigma_o^2 \prod_{i=1}^n 2\pi (\sigma_{ri}^2 + \tau_r^2) \right\}^{1/2}} d\theta \\ &= \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} \right\} \right]}{\left\{ 2\pi \sigma_o^2 \prod_{i=1}^n 2\pi (\sigma_{ri}^2 + \tau_r^2) \right\}^{1/2}} \underbrace{\int \exp \left[ -\frac{1}{2} \left\{ \frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right] d\theta}_{=N(\hat{\theta}_{r*}; m, \sigma_o^2 + \sigma_{r*}^2) 2\pi \sigma_o \sigma_{r*}} \\ &= \left\{ (1 + \sigma_o^2 / \sigma_{r*}^2) \prod_{i=1}^n 2\pi (\sigma_{ri}^2 + \tau_r^2) \right\}^{-1/2} \exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - \hat{\theta}_o)^2}{\sigma_{r*}^2 + \sigma_o^2} \right\} \right]. \end{aligned}$$

This then leads to the replication Bayes factor

$$\text{BF}_{01}(\hat{\theta}_r) = \frac{f(\hat{\theta}_r | H_0)}{f(\hat{\theta}_r | H_1)} = \sqrt{1 + \sigma_o^2 / \sigma_{r*}^2} \exp \left[ -\frac{1}{2} \left\{ \frac{\hat{\theta}_{r*}^2}{\sigma_{r*}^2} - \frac{(\hat{\theta}_{r*} - \hat{\theta}_o)^2}{\sigma_{r*}^2 + \sigma_o^2} \right\} \right],$$

which is equivalent to the replication Bayes factor for one replication study from (1) using the weighted average  $\hat{\theta}_{r*}$  and its standard error  $\sigma_{r*}$  as the replication effect estimate  $\hat{\theta}_r$  and standard error  $\sigma_r$ .

## 5 The sceptical $p$ -value

Held (2020) proposed a reverse-Bayes approach for assessing replicability. One assumes again the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  with  $i \in \{o, r\}$ , along with a zero-mean “sceptical” prior  $\theta \sim N(0, \sigma_s^2)$  for the effect size. In a first step, a level  $\alpha \geq 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$  is fixed and the variance  $\sigma_s^2$  of the sceptical prior is computed

$$\sigma_s^2 = \frac{\sigma_o^2}{(z_o^2 / z_\alpha^2) - 1}$$

where  $z_o = \hat{\theta}_o / \sigma_o$ . The sceptical prior variance  $\sigma_s^2$  renders the resulting posterior of  $\theta$  no longer “convincing” at level  $\alpha$ , that is, the posterior tail probability is fixed to  $\Pr(\theta \geq 0 = 1 - \alpha)$  for positive estimates and  $\Pr(\theta \leq 0 = 1 - \alpha)$  for negative estimates. In a second step, the conflict between the sceptical prior and the observed replication data is quantified.

## 6 The sceptical Bayes factor

Pawel and Held (2022) showed that the success region based on  $\text{BF}_R \leq \gamma$  is given by

$$S_{\text{BF}_S} = \begin{cases} (-\infty, -\sqrt{B} - M] \cup [\sqrt{B} - M, \infty) & \text{for } s_\gamma < 1 \\ [\hat{\theta}_o - \{(\sigma_o^2 + \sigma_r^2) \log \gamma\} / \hat{\theta}_o, \infty) & \text{for } s_\gamma = 1 \\ [-\sqrt{B} - M, \sqrt{B} - M] & \text{for } s_\gamma > 1 \end{cases} \quad (2)$$

with

$$\begin{aligned} B &= \left\{ \frac{\hat{\theta}_o^2}{\sigma_o^2(1 - s_\gamma)} + 2 \log \left( \frac{\sigma_r^2 + \sigma_o^2}{\sigma_r^2 + s\sigma_o^2} \right) - 2 \log \gamma \right\} \frac{(\sigma_r^2 + s\sigma_o^2)(\sigma_r^2 + \sigma_o^2)}{\sigma_o^2(1 - s_\gamma)} \\ M &= \frac{\hat{\theta}_o(\sigma_r^2 + s_\gamma \sigma_o^2)}{\sigma_o^2(1 - s_\gamma)} \\ s_\gamma &= \begin{cases} -\frac{z_o^2}{q} - 1 & \text{if } -\frac{z_o^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \\ \text{where } q &= W_{-1} \left( -\frac{z_o^2}{\gamma^2} \cdot \exp \{ -z_o^2 \} \right) \end{aligned}$$

## References

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448.

doi:10.1111/rssa.12493.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45:142–152. doi:10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490. doi:10.1177/2515245918810225.

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911. doi:10.1111/rssb.12491.