

Bayesian approaches to designing replication studies

Supplementary materials

Samuel Pawel^{*}, Guido Consonni[†], and Leonhard Held^{*}

^{*} Department of Biostatistics, University of Zurich

[†] Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore

E-mail: samuel.pawel@uzh.ch

October 12, 2022

In this document we provide additional information on methods for analyzing replication data. For each method we also derive the *success region* in terms of the effect estimate of the replication study $\hat{\theta}_r$, which is required for sample size determination as illustrated in the main manuscript. For the two-trials rule and the replication Bayes factor methods we additionally provide derivations on how these methods can be generalized to the multisite replication setting.

1 The two-trials rule

The two-trials rule is the most common analysis approach for replication studies. Replication success is declared if both original and replication study achieve statistical significance at some level α (and both estimates go in the same direction which can be taken into account by using one-sided p -values). We will study the two-trial under normality using the data model $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$ with $\hat{\theta}_i$ the estimate of the unknown effect size θ from study i and σ_i is the corresponding standard error (assumed to be known). The p -values for testing $H_0: \theta = 0$ versus $H_1: \theta > 0$ are then $p_i = 1 - \Phi(\hat{\theta}_i/\sigma_i)$ whereas for the alternative $H_1: \theta < 0$ they are $p_i = \Phi(\hat{\theta}_i/\sigma_i)$. Suppose the original effect estimate was statistically significant at level α , i. e., $p_o \leq \alpha$. Replication success at level α is then established if the replication effect estimate $\hat{\theta}_r$ is also statistically significant at level α , i. e., $p_r \leq \alpha$. By applying some algebraic manipulations to the success condition, one can show that this implies that replication success is achieved if the replication effect estimate $\hat{\theta}_r$ is contained in the success region

$$S_{2TR} = \begin{cases} [z_\alpha \sigma_r, \infty) & \text{for } \hat{\theta}_o > 0 \\ [-\infty, -z_\alpha \sigma_r) & \text{for } \hat{\theta}_o < 0. \end{cases}$$

1.1 The multisite two-trials rule

If multiple replication studies are conducted for one original study (a *multisite* replication), the two-trials rule is typically modified by meta-analyzing the effect estimates from all replications and then using the combined estimate as usual in the two-trials rule (see e. g., the “Many labs” projects from [Klein et al., 2014, 2018](#)). Suppose m replication studies are conducted and produce m effect estimates $\hat{\theta}_{r1}, \dots, \hat{\theta}_{rm}$ with standard errors $\sigma_{r1}, \dots, \sigma_{rm}$. Subsequently, 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 $\sigma_{r*} = 1/\sqrt{\{\sum_{i=1}^m 1/(\sigma_{ri}^2 + \tau_r^2)\}}$ can be computed. If the between-replication heterogeneity variance τ_r^2 is set to zero this corresponds to the fixed effects estimate of θ , while estimating τ_r^2

from the data corresponds to the random effects estimate. Replication success at level α is then established if the replication p -value is smaller than α , 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

$$S_{2TR} = \begin{cases} [z_\alpha \sigma_{r*}, \infty) & \text{for } \hat{\theta}_o > 0 \\ [-\infty, -z_\alpha \sigma_{r*}) & \text{for } \hat{\theta}_o < 0. \end{cases}$$

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 θ from study $i \in \{o, r\}$ and σ_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 σ_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 θ based on the data from original and replication study and an initial flat prior for θ . Fixed effects meta-analysis is typically used because estimating a heterogeneity variance from two studies is highly unstable. Replication success at level α is established if the one-sided meta-analytic p -value (in the direction of the original effect estimate $\hat{\theta}$) is significant at level α , i. e., $p_m = 1 - \Phi(\hat{\theta}_m/\sigma_m) \leq \alpha$ for $\hat{\theta}_o > 0$ and $p_m = \Phi(\hat{\theta}_m/\sigma_m) \leq \alpha$ for $\hat{\theta}_o < 0$. 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

$$S_{MA} = \begin{cases} [\sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2, \infty) & \text{for } \hat{\theta}_o > 0 \\ (-\infty, -\sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2] & \text{for } \hat{\theta}_o < 0. \end{cases}$$

3 Effect size equivalence

The effect size equivalence approach ([Anderson and Maxwell, 2016](#)) defines replication success via comparability of the effect estimates from both studies. Under normality we may assume the data model $\hat{\theta}_i | \theta_i \sim N(\theta_i, \sigma_i^2)$ for study $i \in \{o, r\}$, and we are interested in the true effect size difference $\delta = \theta_r - \theta_o$. A $(1 - \alpha)$ confidence interval for δ is then given by

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

Effect size equivalence is established if the confidence interval is fully included in an equivalence region $C_\alpha \subseteq [-\Delta, \Delta]$ with $\Delta > 0$ a pre-specified margin. Applying some algebraic manipulations to the success conditions one can show that the equivalence test 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

The replication Bayes factor approach uses the replication data x_r to quantify the evidence for the null hypothesis $H_0: \theta = 0$ relative to the alternative hypothesis $H_1: \theta \sim f(\theta | x_o)$, which postulates that the effect size θ is distributed according to its posterior distribution based on the original data x_o . Assume again a normal model $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$ with $\hat{\theta}_i$ an estimate of the effect size θ from study $i \in \{o, r\}$ and σ_i the corresponding standard error (assumed to be known), and that we use the alternative $H_1: N(\hat{\theta}_o, \sigma_o^2)$ which arises from updating an initial flat prior for θ the original data $x_o = \{\hat{\theta}_o, \sigma_o\}$. The replication Bayes factor is then

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

Replication success at level $\gamma \in (0, 1)$ is achieved if $\text{BF}_R \leq \gamma$. By applying some algebra to $\text{BF}_R \leq \gamma$, one can show that it is equivalent to the replication effect estimate $\hat{\theta}_r$ falling in the success region

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

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

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

As in the singlsite case, the replication Bayes factor quantifies the evidence that the data provide for the null hypothesis $H_0: \theta = 0$ relative to the alternative hypothesis $H_1: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$. The marginal density of the replication data under the null hypothesis is simply $\hat{\theta}_r | H_0 \sim N_m\{0 \mathbf{1}_m, \text{diag}(\sigma^2 + \tau_r^2 \mathbf{1}_m)\}$, whereas the marginal likelihood under the alternative H_1 is obtained from integrating the likelihood with respect to the prior distribution of θ under the alternative H_1 . 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 τ_r^2 and its variance. The marginal density is then

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

Dividing the marginal density of $\hat{\theta}_r$ under H_0 by the marginal density of $\hat{\theta}_r$ under H_1 leads to cancelation of several terms, and produces the replication Bayes factor

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

The multisite replication Bayes factor is therefore equivalent to the singlesite replication Bayes factor from (1) but using the weighted average $\hat{\theta}_{r*}$ and its standard error σ_{r*} as the replication effect estimate $\hat{\theta}_r$ and standard error σ_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 $\alpha \geq p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$ is fixed and the “sufficiently sceptical” prior variance σ_s^2 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 sufficiently sceptical prior variance σ_s^2 has the property that it renders the resulting posterior of θ no longer “credible” at level α , that is, the posterior tail probability is fixed to $\Pr(\theta \geq 0 | \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$ for positive estimates and $\Pr(\theta \leq 0 | \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$ for negative estimates. In a second step, the conflict between the sceptical prior and the observed replication data is quantified, larger conflict indicating a higher degree of replication success. For doing so, a prior predictive tail probability

$$p_{\text{Box}} = \begin{cases} 1 - \Phi \left\{ \hat{\theta}_r / (\sigma_r^2 + \sigma_s^2) \right\} & \text{if } \hat{\theta}_o > 0 \\ \Phi \left\{ \hat{\theta}_r / (\sigma_r^2 + \sigma_s^2) \right\} & \text{if } \hat{\theta}_o < 0 \end{cases}$$

is computed and replication success at level α is declared if $p_{\text{Box}} \leq \alpha$. The smallest level α at which replication success is achieved is called the *sceptical p-value* p_s and replication success at level α is equivalent with $p_s \leq \alpha$ (see [Held, 2020](#); [Held et al., 2022](#), for more details on p_s). By applying some algebraic manipulations to the condition $p_{\text{Box}} \leq \alpha$, one can show that it is equivalent to the replication effect estimate $\hat{\theta}_r$ falling in the success region

$$S_{p_s} = \begin{cases} [z_\alpha \sqrt{\{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2)-1}\}}, \infty) & \text{if } \hat{\theta}_o > 0 \\ (-\infty, -z_\alpha \sqrt{\{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2)-1}\}}] & \text{if } \hat{\theta}_o < 0. \end{cases}$$

6 The sceptical Bayes factor

[Pawel and Held \(2022\)](#) modified the reverse-Bayes assessment of replication success from [Held \(2020\)](#) to use Bayes factors ([Jeffreys, 1961](#); [Kass and Raftery, 1995](#)) instead of tail probabilities as measures of evidence and prior data conflict. The procedure assumes again the data model $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$ for study $i \in \{o, r\}$. In the first step the original data are used to contrast the evidence for the point null hypothesis $H_0: \theta = 0$ relative to the “sceptical” alternative $H_S: \theta \sim N(0, \sigma_s^2)$ with the Bayes factor

$$\text{BF}_{0S} = \frac{f(\hat{\theta}_o | H_0)}{f(\hat{\theta}_o | H_S)} = \sqrt{1 + \sigma_s^2/\sigma_o^2} \exp \left\{ -\frac{z_o^2}{2(1 + \sigma_o^2/\sigma_s^2)} \right\}.$$

where $z_o = \hat{\theta}_o/\sigma_o$. One then determines the sufficiently sceptical prior variance σ_s^2 so that the Bayes factor is fixed to a level $\gamma \in (0, 1)$ meaning that there is no longer evidence against the null hypothesis at level γ . The sufficiently sceptical prior variance can be computed by

$$\sigma_s^2 = \begin{cases} -\frac{\hat{\theta}_o^2}{q} - \sigma_o^2 & \text{if } -\frac{\hat{\theta}_o^2}{q} \geq \sigma_o^2 \\ \text{undefined} & \text{else} \end{cases} \quad (2)$$

$$\text{where } q = W_{-1} \left\{ -\frac{z_o^2}{\gamma^2} \exp(-z_o^2) \right\} \quad (3)$$

with $W_{-1}(\cdot)$ the branch of the Lambert W function with $W(y) \leq -1$ for $y \in [-1/e, 0)$.

In a second step the conflict between the sceptical prior and the replication data is quantified. To do so, the sceptic is contrasted to the “advocacy” alternative $H_A: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$ which represents the position of an advocate as the prior corresponds to the posterior distribution based on the original data $\{\hat{\theta}_o, \sigma_o\}$ and a flat prior for the effect size θ . This is done by computing the Bayes factor

$$\text{BF}_{SA} = \frac{f(\hat{\theta}_r | H_S)}{f(\hat{\theta}_r | H_A)} = \sqrt{\frac{\sigma_o^2 + \sigma_r^2}{\sigma_s^2 + \sigma_r^2}} \exp \left[-\frac{1}{2} \left\{ \frac{\hat{\theta}_r^2}{\sigma_s^2 + \sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2} \right\} \right]$$

and replication success at level γ is defined by $\text{BF}_{SA} \leq \gamma$ as the data favor the advocate over the sceptic at a higher level than the sceptic’s initial objection to the null hypothesis. The smallest level γ at which replication success is achievable is then called the *sceptical Bayes factor* BF_s , and replication success at level γ is equivalent to $\text{BF}_s \leq \gamma$ (see [Pawel and Held, 2022](#), for details on how to compute BF_s). To derive the success region of the sceptical Bayes factor one can apply algebraic manipulations to $\text{BF}_{SA} \leq \gamma$ the

condition for replication success at level γ , leading to

$$S_{\text{BF}_S} = \begin{cases} (-\infty, -\sqrt{B} - M] \cup [\sqrt{B} - M, \infty) & \text{for } \sigma_s^2 < \sigma_o^2 \\ [\hat{\theta}_o - \{(\sigma_o^2 + \sigma_r^2) \log \gamma\} / \hat{\theta}_o, \infty) & \text{for } \sigma_s^2 = \sigma_o^2 \\ [-\sqrt{B} - M, \sqrt{B} - M] & \text{for } \sigma_s^2 > \sigma_o^2 \end{cases} \quad (4)$$

with

$$B = \left\{ \frac{\hat{\theta}_o^2}{\sigma_o^2 - \sigma_s^2} + 2 \log \left(\frac{\sigma_o^2 + \sigma_r^2}{\sigma_s^2 + \sigma_r^2} \right) - 2 \log \gamma \right\} \frac{(\sigma_s^2 + \sigma_r^2)(\sigma_o^2 + \sigma_r^2)}{\sigma_o^2 - \sigma_s^2}$$

$$M = \frac{\hat{\theta}_o(\sigma_s^2 + \sigma_r^2)}{\sigma_o^2 - \sigma_s^2}$$

and the sufficiently sceptical prior variance σ_s^2 computed by (2).

References

- Anderson, S. F. and Maxwell, S. E. (2016). There’s more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12. doi:10.1037/met0000051.
- 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.
- Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:10.1214/21-aos1502.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, third edition.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:10.1080/01621459.1995.10476572.
- 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.