Journal of Statistical Software

November 2021, Volume 100, Issue 17. *doi: 10.18637/jss.v100.i17*

Informed Bayesian Inference for the A/B Test

Quentin F. Gronau
University of Amsterdam

Akash Raj K. N.
University of Amsterdam

Eric-Jan Wagenmakers
University of Amsterdam

Abstract

Booming in business and a staple analysis in medical trials, the A/B test assesses the effect of an intervention or treatment by comparing its success rate with that of a control condition. Across many practical applications, it is desirable that (1) evidence can be obtained in favor of the null hypothesis that the treatment is ineffective; (2) evidence can be monitored as the data accumulate; (3) expert prior knowledge can be taken into account. Most existing approaches do not fulfill these desiderata. Here we describe a Bayesian A/B procedure based on Kass and Vaidyanathan (1992) that allows one to monitor the evidence for the hypotheses that the treatment has either a positive effect, a negative effect, or, crucially, no effect. Furthermore, this approach enables one to incorporate expert knowledge about the relative prior plausibility of the rival hypotheses and about the expected size of the effect, given that it is non-zero. To facilitate the wider adoption of this Bayesian procedure we developed the **abtest** package in R. We illustrate the package options and the associated statistical results with a fictitious business example and a real data medical example.

Keywords: model comparison, Bayes factor, prior elicitation, Bayesian estimation.

1. Introduction

Does the modification of a company website increase the number of online purchases? Does a new drug result in a lower mortality rate? These are just two examples of the kinds of questions that can be addressed with A/B testing, a procedure popular not only in business and medical clinical trials, but also in fields such as psychology, neuroscience, and biology. The A/B test set-up discussed in this article assumes that the outcome variable is binary; nevertheless, the outcome variable could in principle also be continuous. Based on a binary outcome variable, an A/B test compares the success rate of two options or treatment arms, A and B, and therefore can be conceptualized as a test for a difference between two proportions (Little 1989). Typically, options A and B correspond to a control condition and an intervention or treatment of interest.

Regardless of the specific field of application, we believe three general desiderata for A/B tests can be identified. First, we believe it is desirable that evidence can be obtained in favor of the null hypothesis that there is no difference between options A and B. For instance, suppose a programmer alters code that should leave the appearance of a website unaffected. An A/B test may be conducted to confirm that the code changes did not lead to unintended consequences. Alternatively, suppose that a cheaper drug is introduced as a replacement of the standard drug; here, an A/B test may confirm that the cheaper drug is as effective as the drug that is currently standard.

Second, we believe it is desirable that evidence can be monitored as the data accumulate. Data collection can be time-consuming and expensive, and interim tests allow one to assess whether the results in hand are already sufficiently compelling or whether additional data ought to be obtained. There is also an ethical aspect to this desideratum, one that is particularly pronounced in case of new clinical treatments that are potentially beneficial or harmful; it is unethical to withhold treatment that interim analysis shows to be beneficial, just as it is unethical to continue to administer a treatment that interim analysis shows to be harmful (e.g., Armitage 1960; see also Ware 1989 and the accompanying discussion).

Third, we believe it is desirable that expert knowledge can be taken into account (e.g., O'Hagan 2019). In many A/B testing applications, there exists considerable expert knowledge about what size of effect to expect. For instance, the effect of website changes on conversion rates is often less than 0.5% (Berman, Pekelis, Scott, and Van den Bulte 2018). Incorporating such expert knowledge into the statistical analysis will yield a more targeted test.

The majority of A/B testing procedures that are currently in vogue do not fulfill the above desiderata. Specifically, many companies apply standard p-value-based null hypothesis significance testing to assess whether or not options A and B differ. This procedure has the advantage that it is readily available in software such as R (R Core Team 2021, e.g., via the functions `prop.test()`, `fisher.test()`, and `chisq.test()`). However, this approach cannot distinguish between *absence of evidence* (i.e., the data are inconclusive) and *evidence of absence* (i.e., the data provide support for the null hypothesis that options A and B do not differ; e.g., Dienes 2014; Keysers, Gazzola, and Wagenmakers 2020). Furthermore, although common practice, sequentially monitoring the uncorrected p-value (and stopping data collection as soon as the p-value is smaller than some fixed α-level) invalidates the analysis (e.g., Feller 1940). However, there exist valid classical sequential procedures that enable one to monitor a corrected p-value as data accumulate (e.g., Malek, Katariya, Chow, and Ghavamzadeh 2017). For instance, *Optimizely*, one of the leading commercial A/B testing platforms, has recently implemented an alternative p-value-based approach that allows users to continuously monitor the test outcome (Johari, Koomen, Pekelis, and Walsh 2017). Nevertheless, these sequential p-value-based procedures retain the inability to quantify evidence for the absence of an effect. Furthermore, (sequential) p-value-based A/B testing does not allow one to incorporate expert knowledge into the statistical analysis in a straightforward manner.

An alternative A/B testing approach that has become more popular of late is Bayesian estimation. For instance, *VWO*, another leading A/B testing platform, has recently implemented a Bayesian estimation approach (Stucchio 2015). Bayesian estimation is also available via the **BayesianFirstAid** package (Bååth 2014) and the **bayesAB** package (Portman 2021).[1] Since

[1] The **bayesAB** package provides a range of functions for Bayesian A/B testing. One advantage is that users can choose from a range of different data distributions (e.g., Bernoulli, normal, Poisson, etc.).

Bayesian inference does not require sample sizes to be fixed a priori (Berger and Wolpert 1988), this approach allows one to monitor the analysis output as data accumulate. It also enables the incorporation of expert knowledge via the specification of a prior distribution that captures the expert's knowledge about a parameter of interest. However, this approach operates under the assumption that an effect exists – since a continuous prior assigns zero probability to a single null value – and consequently does not allow one to obtain evidence in favor of the null hypothesis of no effect. For instance, **bayesAB** and **BayesianFirstAid** provide the user with the posterior probability that one option yields more successes than the other, but this ignores the fact that both options could be equally effective. Furthermore, the currently used Bayesian estimation approaches – such as the one implemented in **bayesAB** and **BayesianFirstAid** – typically assign independent priors to the success probabilities of the control and treatment condition, a practice that was critiqued by Howard (1998).[2]

To overcome the limitations of the current A/B tests we developed the **abtest** package in R (R Core Team 2021). Package **abtest** (Gronau 2021) is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=abtest`. The **abtest** package implements one form of Bayesian inference for the A/B test, using informed prior distributions that induce a dependency between the two success probabilities. The analysis approach is based on a model by Kass and Vaidyanathan (1992); for alternative approaches see Deng, Lu, and Chen (2016), Jamil, Marsman, Ly, Morey, and Wagenmakers (2017), Pham-Gia, Van Thin, and Doan (2017), and Skorski (2019). The implemented Bayesian procedure allows users (1) to obtain evidence in favor of the null hypothesis (e.g., Berger and Delampady 1987; Wagenmakers *et al.* 2018); (2) monitor the evidence as the data accumulate (e.g., Rouder 2014); and (3) elicit and incorporate expert prior knowledge (e.g., O'Hagan 2019). The **abtest** package thus fulfills all three desiderata mentioned above.

The **abtest** package provides functionality for both hypothesis testing and parameter estimation. In line with Jeffreys (1939) and Fisher (1928), we believe that testing and estimation are complementary activities (Haaf, Ly, and Wagenmakers 2019): before a parameter is estimated, it should be tested whether there is anything to justify estimation at all. Jeffreys (1939, p. 345) related this principle to Occam's razor: "variation must be taken as random until there is positive evidence to the contrary" (see also Kass and Raftery 1995, Section 8.1). However, some researchers and practitioners oppose this idea, for instance because they believe that one should replace hypothesis testing with parameter estimation (e.g., Gelman and Rubin 1995; Cumming 2014). Nevertheless, the **abtest** package may also be useful for researchers without an interest in hypothesis testing, since the package can also be used exclusively for Bayesian parameter estimation (and prior elicitation).

This article is organized as follows: The next section introduces a fictitious business example. Afterwards, the implementation details of the Bayesian A/B test procedure used in **abtest** are discussed. Subsequently, the fictitious example is continued and the functionality of the **abtest** package and the practical benefits of the implemented approach are demonstrated. Next, a real data medical example is used to demonstrate further functionality of the package. The article ends with concluding comments.

[2] "do English or Scots cattle have a higher proportion of cows infected with a certain virus? Suppose we were informed (before collecting any data) that the proportion of English cows infected was 0.8. With independent uniform priors we would now give H_1 ($p_1 > p_2$) a probability of 0.8 (because the chance that $p_2 > 0.8$ is still 0.2). In very many cases this would not be appropriate. Often we will believe (for example) that if p_1 is 80%, p_2 will be near 80% as well and will be almost equally likely to be larger or smaller." (p. 363)

2. Example 1: Effectiveness of resilience training

Suppose the managers of a large consultancy firm are interested in reducing the number of employees who quit within the first six months, possibly due to the high stress involved in the job. A coaching company offers a resilience training and claims that this training greatly reduces the number of employees who quit. Implementing the training for all newly hired employees would be expensive and some of the managers are not completely convinced that the training is at all effective. Therefore, the managers decide to run an A/B test where half of a sample of newly hired employees will receive the training, the other half will not be trained. The outcome variable is whether or not an employee quit within the first six months ($1 =$ still on the job, $0 =$ quit).

The consultancy firm collects $1,000$ observations (500 in each group). These (fictitious) data[3] are included in the **abtest** package (i.e., `seqdata`). The number of employees still on the job after six months is 249 in the group without training and 269 in the trained group. Figure 1 provides an illustration of some of the information that can be obtained by analyzing these data using **abtest**. The figure displays the probability of the hypothesis that the training has a positive effect (i.e., $\mathcal{H}_+$), negative effect (i.e., $\mathcal{H}_-$), and no effect (i.e., $\mathcal{H}_0$) as a function of the number of observations across the two groups. The top part of the figure displays the probability of the three hypotheses before and after taking into account the observed data (i.e., prior and posterior probabilities) as probability wheels (e.g., Tversky 1969; Lipkus and Hollands 1999). Before providing more details about how to obtain and interpret this result as well as providing additional analyses, we discuss the implementation details of the A/B test procedure used by **abtest**.

[3] The data set is structured such that the sequential nature of the data is retained: The data set contains the number of observations and the number of successes in each of the two groups after each observation.

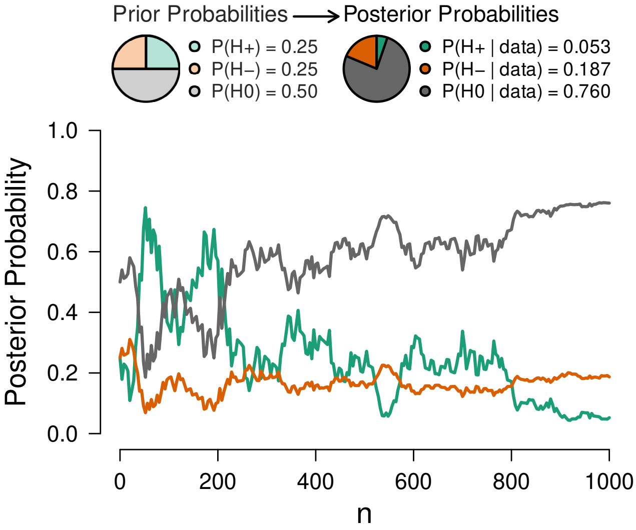

Figure 1: The posterior probability of the hypothesis that the training in Example 1 has a positive effect (i.e., $\mathcal{H}_+$), negative effect (i.e., $\mathcal{H}_-$), and no effect (i.e., $\mathcal{H}_0$) is plotted as a function of the number of observations across groups. On top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities after taking into account all observations.

3. Implementation details

The Bayesian A/B test implemented in the **abtest** package is based on Kass and Vaidyanathan (1992, Section 3, "Testing Equality of Two Binomial Proportions"). Appendices A–C provide detailed derivations.

3.1. Model

Let y_1 denote the number of successes for option A with n_1 denoting the corresponding total number of observations for option A. Similarly, y_2 denotes the number of successes for option B with n_2 denoting the corresponding total number of observations for option B. The Bayesian A/B test model based on Kass and Vaidyanathan (1992) is specified as follows:[4]

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2}\\ \log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2}\\ y_1 &\sim \text{Binomial}(n_1, p_1)\\ y_2 &\sim \text{Binomial}(n_2, p_2).\end{aligned} \tag{1}$$

Therefore, the model assumes that y_1 and y_2 follow binomial distributions with success probabilities p_1 and p_2. These probabilities are functions of the two model parameters, β and ψ. Specifically, the log odds corresponding to p_1 are given by $\beta - \psi/2$ and the log odds corresponding to p_2 are given by $\beta + \psi/2$. The nuisance parameter β corresponds to the grand mean of the log odds and the test-relevant parameter ψ corresponds to the log odds ratio. When ψ is positive, this implies that $p_2 > p_1$ (i.e., option B has a higher success probability than option A); when ψ is negative this implies that $p_2 < p_1$ (i.e., option B has a lower success probability than option A).

3.2. Hypotheses

The **abtest** package enables both estimation of the model parameters and testing of hypotheses about the test-relevant log odds ratio parameter ψ. There are four hypotheses that are of potential interest:

1. The null hypothesis $\mathcal{H}_0$ which states that the success probabilities p_1 and p_2 are identical, that is, $p_1 = p_2$. This is equivalent to $\mathcal{H}_0 : \psi = 0$. This hypothesis corresponds to the claim that there is no difference between options A and B (i.e., the "A/A test").

2. The two-sided alternative hypothesis $\mathcal{H}_1$ which states that the two success probabilities p_1 and p_2 are not equal (i.e., $p_1 \neq p_2$), but does not specify which of the two is larger. This is equivalent to $\mathcal{H}_1 : \psi \neq 0$. This hypothesis corresponds to the claim that options A and B differ but it is not specified which one yields more successes.

3. The one-sided hypothesis $\mathcal{H}_+$ which states that the second success probability p_2 is larger than the first success probability p_1. This is equivalent to $\mathcal{H}_+ : \psi > 0$. This hypothesis corresponds to the claim that option B yields more successes than option A.

[4] Note that this is equivalent to a logistic regression model with a binary covariate (i.e., group membership) that is coded using ± 0.5.

	Test				
Hypothesis	Default	Undirected	Positive	Negative	Direction
$\mathcal{H}_0$	*0.50*	*0.50*	*0.50*	*0.50*	0.00
$\mathcal{H}_1$	0.00	*0.50*	0.00	0.00	0.00
$\mathcal{H}_+$	*0.25*	0.00	*0.50*	0.00	*0.50*
$\mathcal{H}_-$	*0.25*	0.00	0.00	*0.50*	*0.50*

Table 1: Changing the prior probability assignments across rival hypotheses produces different tests.

4. The one-sided hypothesis $\mathcal{H}_-$ which states that the first success probability p_1 is larger than the second success probability p_2. This is equivalent to $\mathcal{H}_- : \psi < 0$. This hypothesis corresponds to the claim that option A yields more successes than option B.

Researchers who conduct an A/B test are usually interested in answering the question: Does option B yield more successes than option A (i.e., $\mathcal{H}_+$), fewer successes than option A (i.e., $\mathcal{H}_-$), or is there no difference between options A and B (i.e., $\mathcal{H}_0$)? Therefore, it may be argued that the hypotheses of interest are typically $\mathcal{H}_+$, $\mathcal{H}_-$, and $\mathcal{H}_0$. Consequently, by default, only these three hypotheses are assigned non-zero prior probability in the **abtest** package. Specifically, a default prior probability of 0.50 is assigned to the hypothesis that there is no effect (i.e., $\mathcal{H}_0$), and the remaining prior probability is split evenly across the hypothesis that there is a positive effect (i.e., $\mathcal{H}_+$ receives 0.25) and a negative effect (i.e., $\mathcal{H}_-$ also receives 0.25). The user may change these default prior probabilities to custom values. Table 1 provides an overview of five qualitatively different tests that can be conducted by assigning prior probabilities to hypotheses in certain ways.[5] The first column displays the default setting that assigns probability 0.50 to the null hypothesis and splits the remaining probability evenly across $\mathcal{H}_+$ and $\mathcal{H}_-$. The second column displays a prior probability assignment that implements an undirected test (i.e., $\mathcal{H}_0$ is compared to the undirected $\mathcal{H}_1$). The third column displays a prior probability assignment for testing whether the effect is non-existent or positive. The fourth column displays a prior probability assignment for testing whether the effect is non-existent or negative. Finally, the fifth column displays a prior probability assignment for a test of direction, that is, for testing whether the effect is positive or negative. This last setting may be of interest whenever the null hypothesis is a priori deemed implausible, uninteresting, or irrelevant.

3.3. Parameter priors

The **abtest** package assigns normal priors to the model parameters: $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ and $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)$. As illustrated in the example below, these priors result in a dependency in the implied prior for the success probabilities p_1 and p_2, which is generally desirable (Howard 1998).

For the one-sided hypotheses $\mathcal{H}_+$ and $\mathcal{H}_-$, the prior on ψ is truncated at zero. Specifically, for

[5] Note that, except for the first column of Table 1 which displays the default setting, the remaining examples use equal prior probabilities for all hypotheses that are assigned non-zero prior probability. However, the user can of course also assign prior probability unevenly to the hypotheses of interest (e.g., if prior knowledge exists about the relative plausibility of the rival hypotheses).

$\mathcal{H}_+$, the prior on ψ is a truncated normal distribution with parameters μ_ψ and σ_ψ and lower bound at zero. For $\mathcal{H}_-$, the prior on ψ is a truncated normal distribution with parameters μ_ψ and σ_ψ and upper bound at zero. These normal priors are computationally convenient and sufficiently flexible to encode a wide range of prior information.

By default, the **abtest** package assigns standard normal priors to both β and ψ. For the nuisance parameter β, a standard normal prior results in a relatively flat implied prior on p_1 and p_2 when $\psi = 0$. Generally, the choice of a prior for the nuisance parameter β is relatively inconsequential (Kass and Vaidyanathan 1992). In contrast, the prior on the test-relevant parameter ψ is consequential, as it defines the extent to which the hypotheses of interest differ from $\mathcal{H}_0$. Our choice for a default standard normal prior on the test-relevant parameter ψ is motivated by the fact that a zero-centered prior does not favor any of the two options A or B a priori. Furthermore, the standard deviation of 1 results in a prior distribution that assigns mass to a wide range of reasonable log odds ratios (Chen, Cohen, and Chen 2010) without being so uninformative that the results unduly favor $\mathcal{H}_0$ (Bartlett 1957; Lindley 1957).[6] However, large changes in the prior standard deviation of the test-relevant parameter may result in large changes in the results, as the prior standard deviation governs the degree to which the hypothesis of interest makes predictions that differ from $\mathcal{H}_0$. To include prior knowledge about the expected results, the **abtest** package allows the user to change the default values of the prior distributions for the nuisance parameter β and the test-relevant parameter ψ, either by changing the location of the normal prior distribution, the scale, or both.

3.4. Encoding prior information

A straightforward way to encode prior information about the model parameters is to set μ_β, σ_β, μ_ψ, and σ_ψ directly. However, it may sometimes be easier to specify prior distributions based on quantities such as the (log) odds ratio, relative risk (i.e., p_2/p_1, the ratio of the success probability in condition B and condition A), and absolute risk (i.e., $p_2 - p_1$, the difference of the success probability in condition B and condition A). The `elicit_prior()` function allows users to encode prior information about a quantity of interest (either log odds ratio, odds ratio, relative risk, or absolute risk). The function assumes that the prior on β is not the primary target of prior elicitation and is fixed by the user a priori (using the arguments `mu_beta` and `sigma_beta`) – for instance, to a standard normal prior which corresponds to a relatively flat implied prior on p_1 and p_2 when $\psi = 0$.

To encode prior information, the user needs to provide quantiles for a quantity of interest. Let $q_i, i = 1, \ldots, I$ denote the values of I quantiles provided by the user and let $\text{prob}_i, i = 1, \ldots, I$ denote the corresponding probabilities (e.g., for the median, $\text{prob}_i = 0.5$). Least-squares minimization is used to obtain μ_ψ and σ_ψ as follows:

$$(\mu_\psi, \sigma_\psi) = \underset{\mu_\psi, \sigma_\psi}{\arg\min} \sum_{i=1}^{I} \left(F(q_i; \mu_\psi, \sigma_\psi) - \text{prob}_i \right)^2, \tag{2}$$

where $F(\cdot; \mu_\psi, \sigma_\psi)$ corresponds to the cumulative distribution function (CDF) for the quantity of interest implied by the normal prior on ψ. For some quantities, this CDF also depends on the prior for β; however, as described above, it is assumed that μ_β and σ_β are fixed a priori.

[6] Note that the default implied prior on the absolute risk $p_2 - p_1$ is considerably more narrow than the prior induced by the popular default choice that assigns p_1 and p_2 independent uniform distributions (Jeffreys 1935).

3.5. Hypothesis testing

To quantify the evidence that the data provide for $\mathcal{H}_0$, $\mathcal{H}_1$, $\mathcal{H}_+$, and $\mathcal{H}_-$, one can compute Bayes factors (Jeffreys 1939; Kass and Raftery 1995) and posterior probabilities of the rival hypotheses. The posterior probability of hypothesis $\mathcal{H}_j$, $j \in \{0, 1, +, -\}$ is given by:

$$\overbrace{p(\mathcal{H}_j \mid \text{data})}^{\text{posterior probability}} = \overbrace{\frac{p(\text{data} \mid \mathcal{H}_j)}{\sum_k p(\text{data} \mid \mathcal{H}_k)\, p(\mathcal{H}_k)}}^{\text{updating factor}} \times \overbrace{p(\mathcal{H}_j)}^{\text{prior probability}} . \tag{3}$$

The Bayes factor for comparing hypotheses $\mathcal{H}_j$ and $\mathcal{H}_k$ equals the change from prior to posterior odds:

$$\underbrace{\frac{p(\mathcal{H}_j \mid \text{data})}{p(\mathcal{H}_k \mid \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\text{data} \mid \mathcal{H}_j)}{p(\text{data} \mid \mathcal{H}_k)}}_{\text{Bayes factor BF}_{jk}} \times \underbrace{\frac{p(\mathcal{H}_j)}{p(\mathcal{H}_k)}}_{\text{prior odds}} . \tag{4}$$

In order to obtain posterior probabilities of the hypotheses and Bayes factors one needs to evaluate the marginal likelihood $p(\text{data} \mid \mathcal{H}_j)$ for each hypothesis $j \in \{0, 1, +, -\}$. For $\mathcal{H}_0$ and $\mathcal{H}_1$, we evaluate the marginal likelihood using Laplace approximations as suggested by Kass and Vaidyanathan (1992). Specifically, the marginal likelihood for $\mathcal{H}_0$ is approximated by:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_0) &= \int \underbrace{p(\text{data} \mid \beta)}_{\text{likelihood}} \underbrace{\pi_0(\beta)}_{\text{prior}} \, \mathrm{d}\beta \\ &\approx (2\pi\sigma_0^2)^{\frac{1}{2}} \exp\left\{l_0^*(\beta_0^*)\right\}, \end{aligned} \tag{5}$$

where $l_0^*(\beta) = \log\{p(\text{data} \mid \beta)\, \pi_0(\beta)\}$, β_0^* corresponds to the mode of $l_0^*(\beta)$, and $\sigma_0^2 = \left(-\frac{d^2}{d\beta^2} l_0^*(\beta)\right)^{-1}\Big|_{\beta=\beta_0^*}$ denotes the inverse of the negative second derivative of $l_0^*(\beta)$ evaluated at the mode β_0^*.

The marginal likelihood for $\mathcal{H}_1$ is approximated by:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_1) &= \int\int \underbrace{p(\text{data} \mid \beta, \psi)}_{\text{likelihood}} \underbrace{\pi(\beta, \psi)}_{\text{prior}} \, \mathrm{d}\beta \mathrm{d}\psi \\ &\approx 2\pi \det(\mathbf{\Sigma}_1)^{\frac{1}{2}} \exp\left\{l^*(\beta^*, \psi^*)\right\}, \end{aligned} \tag{6}$$

where $l^*(\beta, \psi) = \log\{p(\text{data} \mid \beta, \psi)\, \pi(\beta, \psi)\}$, (β^*, ψ^*) denotes the mode of $l^*(\beta, \psi)$, and $\mathbf{\Sigma}_1 = (-\boldsymbol{H}_1)^{-1}\big|_{(\beta,\psi)=(\beta^*,\psi^*)}$ denotes the inverse of the negative Hessian $\boldsymbol{H}_1$ (i.e., the matrix with second-order partial derivatives) of $l^*(\beta, \psi)$ evaluated at the mode (β^*, ψ^*).

These Laplace approximations work well in practice, even for sample sizes that are extremely small. As a demonstration, for a range of synthetic data sets we computed the (log of the) Bayes factor BF_{10} which compares $\mathcal{H}_1$ to $\mathcal{H}_0$ using the above Laplace approximations and, as a comparison, also using bridge sampling (Meng and Wong 1996; Gronau, Singmann, and Wagenmakers 2020). The priors on β and ψ were standard normal distributions. Figure 2 displays the results and confirms that the Laplace approximation yields accurate results, even for sample sizes as small as $n_1 = n_2 = 5$.

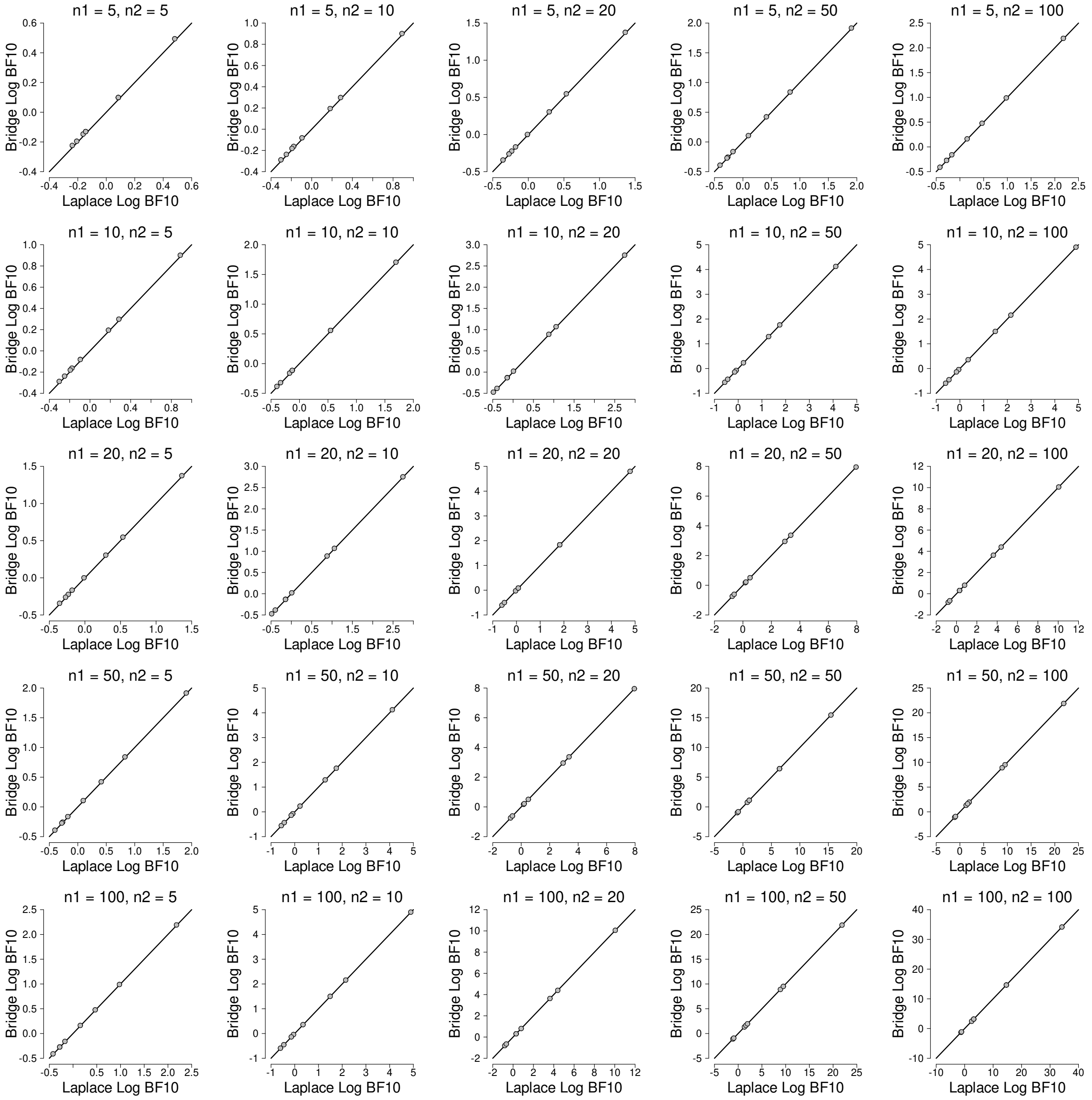

Figure 2: Comparison of the Laplace approximation and bridge sampling for computing the (log of the) Bayes factor BF_{10}. We considered all possible combinations of $n_1 \in \{5, 10, 20, 50, 100\}$ and $n_2 \in \{5, 10, 20, 50, 100\}$. For each of the n_1-n_2 combinations, we considered all possible combinations of $y_1 \in \{\frac{1}{5}n_1, \frac{2}{5}n_1, \frac{3}{5}n_1, \frac{4}{5}n_1\}$ and $y_2 \in \{\frac{1}{5}n_2, \frac{2}{5}n_2, \frac{3}{5}n_2, \frac{4}{5}n_2\}$. The results reveal that the two methods yield highly similar results, even when sample size is very small.

For the one-sided hypotheses $\mathcal{H}_+$ and $\mathcal{H}_-$, Laplace approximations did not appear to yield accurate results for small sample sizes, even after removing the constraint on ψ through the parameterization $(\beta, \xi) = (\beta, \log(\psi))$ for $\mathcal{H}_+$ and $(\beta, \xi) = (\beta, \log(-\psi))$ for $\mathcal{H}_-$. The **abtest** package therefore uses importance sampling to increase the accuracy of the Laplace approximations when computing the marginal likelihoods for $\mathcal{H}_+$ and $\mathcal{H}_-$. Specifically, a Laplace approximation is used to approximate the mode and covariance matrix of the posterior. The importance density is then given by a multivariate t distribution with location set to the approximated posterior mode, scale matrix set to the approximated posterior covariance matrix, and five degrees of freedom (note that the user can change the degrees of freedom). The marginal likelihood for $\mathcal{H}_+$ is then estimated as follows:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_+) &= \int\int \underbrace{p(\text{data} \mid \beta, \xi)}_{\text{likelihood}} \, \underbrace{\pi_+(\beta, \xi)}_{\text{prior}} \, \text{d}\beta \text{d}\xi \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \frac{p(\text{data} \mid \tilde{\beta}_s, \tilde{\xi}_s) \, \pi_+(\tilde{\beta}_s, \tilde{\xi}_s)}{g_{\text{is}}(\tilde{\beta}_s, \tilde{\xi}_s)}, \end{aligned} \tag{7}$$

where $\left\{\tilde{\beta}_s, \tilde{\xi}_s\right\}_{s=1}^{S}$ denotes S samples from the multivariate t importance density g_{is}, and

$$\pi_+(\beta, \xi) = \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) \, \mathcal{N}_+(\exp(\xi); \mu_\psi, \sigma_\psi^2) \, \exp(\xi), \tag{8}$$

where $\mathcal{N}(x; y, z)$ denotes the probability density function of a normal distribution with mean y and variance z that is evaluated at x. Furthermore, $\mathcal{N}_+(x; y, z)$ denotes the density of a normal distribution that is truncated to allow only positive values for x. The marginal likelihood for $\mathcal{H}_-$ is computed analogously.

3.6. Obtaining posterior samples

In a Bayesian A/B test application, one may not only be interested in testing hypotheses, but also in obtaining posterior samples for the model parameters under $\mathcal{H}_1$, $\mathcal{H}_+$, and $\mathcal{H}_-$. The **abtest** package allows the user to obtain posterior samples using sampling importance resampling (e.g., Robert and Casella 2010). Specifically, posterior samples for $\mathcal{H}_+$ are obtained as follows (samples for the other hypotheses are obtained in an analogous manner):

1. Generate S samples from the multivariate t proposal distribution mentioned before, denoted by $\left\{\tilde{\beta}_s, \tilde{\xi}_s\right\}_{s=1}^{S}$.

2. Compute the importance weights:

$$w_s = \frac{p(\text{data} \mid \tilde{\beta}_s, \tilde{\xi}_s) \, \pi_+(\tilde{\beta}_s, \tilde{\xi}_s)}{g_{\text{is}}(\tilde{\beta}_s, \tilde{\xi}_s)}, \qquad s = 1, 2, \ldots, S. \tag{9}$$

3. Re-normalize the importance weights: $v_s = w_s / \sum_{t=1}^{S} w_t$, $s = 1, 2, \ldots, S$.

4. Re-sample (with replacement) from the samples obtained from the importance density according to the normalized importance weights v_s which yields (approximate) samples from the posterior distribution.

4. Example 1: Effectiveness of resilience training (continued)

Next we continue the effectiveness of resilience training example and show how expert prior information can be taken into account, how the hypotheses of interest can be tested, and how one can estimate the model parameters using the **abtest** package.

4.1. Prior specification

Before commencing the A/B test, the managers asked the coaching company to specify how effective they believe the training will be. The coaching company claimed that, based on past experience with the training, they expect the proportion of employees who do not quit within the first six months to be 15% larger for the group who received the training, with a 95% uncertainty interval ranging from a 2.5% benefit to a 27.5% benefit. Assuming that the claimed 15% corresponds to the prior median, this expectation corresponds to a median absolute risk (i.e., $p_2 - p_1$) of 0.15 with a 95% uncertainty interval ranging from 0.025 to 0.275. The `elicit_prior()` function can be used to encode this prior information:[7]

```
R> library("abtest")
R> prior_par <- elicit_prior(q = c(0.025, 0.15, 0.275),
+    prob = c(0.025, 0.5, 0.975),
+    what = "arisk")
```

The obtained prior on the absolute risk can be visualized as follows:

```
R> plot_prior(prior_par, what = "arisk")
```

The resulting graph is shown in the top panel of Figure 3.

The user can also visualize the (implied) prior for other quantities. For instance, the prior on the log odds ratio (middle panel of Figure 3) is obtained as follows:

```
R> plot_prior(prior_par, what = "logor")
```

The implied prior on the success probabilities p_1 and p_2 (bottom panel of Figure 3) is obtained as follows:

```
R> plot_prior(prior_par, what = "p1p2")
```

The bottom panel of Figure 3 illustrates that there is a dependency between p_1 and p_2 which is arguably desirable (Howard 1998): When one of the success probabilities is very (small) large, it is likely that the other one will also be (small) large.

4.2. Hypothesis testing

Since the number of employees still on the job after six months is 249 in the group without training and 269 in the trained group, the observed success probabilities are $\hat{p}_1 = 0.498$ in the control group and $\hat{p}_2 = 0.538$ in the group that received training. Consequently, the observed success probabilities suggest that there is a positive effect of the training of 4%; however, a statistical analysis is required to assess whether this observed difference is statistically compelling. The `ab_test()` function can be used to conduct a Bayesian A/B test as follows:

```
R> data("seqdata", package = "abtest")
R> set.seed(1)
R> ab <- ab_test(data = seqdata, prior_par = prior_par)
```

[7] All code and plots are also available at https://osf.io/t3ajr/.

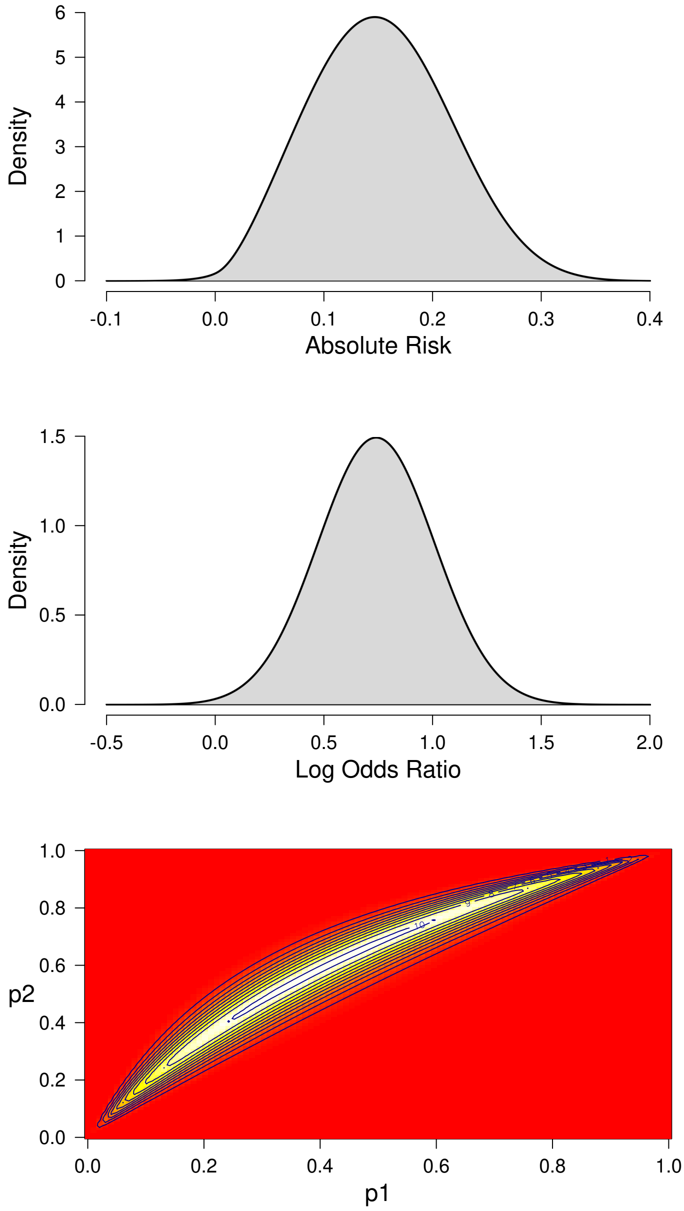

Figure 3: Elicited (implied) prior distributions for the effectiveness of the resilience training in Example 1. The top panel displays the prior distribution for the absolute risk which corresponds to the difference between the probability of still being on the job for the trained and the non-trained employees (i.e., $p_2 - p_1$). The middle panel shows the prior distribution for the log odds ratio parameter ψ. The bottom panel displays the implied joint prior distribution for the success probabilities p_1 and p_2. The bottom panel illustrates that the two success probabilities are assigned dependent priors. Furthermore, most prior mass is above the main diagonal which represents the coaching company's prior expectation that the training is successful.

This yields the following output:

```
R> print(ab)

Bayesian A/B Test Results:

 Bayes Factors:

 BF10: 0.1406443
 BF+0: 0.13823
 BF-0: 0.4920187

 Prior Probabilities Hypotheses:

 H+: 0.25
 H-: 0.25
 H0: 0.5

 Posterior Probabilities Hypotheses:

 H+: 0.0526
 H-: 0.1871
 H0: 0.7604
```

The first part of the output presents Bayes factors in favor of the hypotheses $\mathcal{H}_1$, $\mathcal{H}_+$, and $\mathcal{H}_-$, where the reference hypothesis (i.e., denominator of the Bayes factor) is $\mathcal{H}_0$. Since all three Bayes factors are smaller than 1, they all indicate evidence in favor of the null hypothesis of no effect. The next part of the output displays the prior probabilities of the hypotheses with non-zero prior probability. As explained before, the default setting assigns probability 0.50 to the null hypothesis and splits the remaining probability evenly across $\mathcal{H}_+$ and $\mathcal{H}_-$. The user can change this default setting via the `prior_prob` argument (e.g., to assign non-zero probability to $\mathcal{H}_1$). The final part of the output displays the posterior probabilities of the hypotheses with non-zero prior probability. The posterior probability of the null hypothesis $\mathcal{H}_0$ indicates that the data have increased the plausibility of the null hypothesis from 0.50 to 0.76. Furthermore, the data have decreased the plausibility of both $\mathcal{H}_+$ and $\mathcal{H}_-$.

As an aside, it may appear paradoxical that the data indicate a 4% positive effect of the training and yet the posterior probability of $\mathcal{H}_-$ is larger than that of $\mathcal{H}_+$. The reason for this result is that the company's prior was overly ambitious, and $\mathcal{H}_+$ is penalized for having predicted effects that are much too large. Furthermore, note that the test-relevant prior distribution under $\mathcal{H}_-$ is obtained by truncating the prior on ψ at zero and re-normalizing. Since the company's prior assigns almost all mass to positive log odds ratio values, re-normalizing the negative part of the distribution results in a prior that is highly similar to $\mathcal{H}_0$; this explains why $\mathcal{H}_-$ receives non-trivial posterior probability. These considerations underscore the fact that the outcome of a Bayesian analysis is always relative to the specific set of models (and associated prior distributions) under consideration. Because highly informed priors can exert a large influence on the results, it is generally wise to examine the robustness of the conclusions by executing the default analysis as well. This analysis is reported in Appendix D.

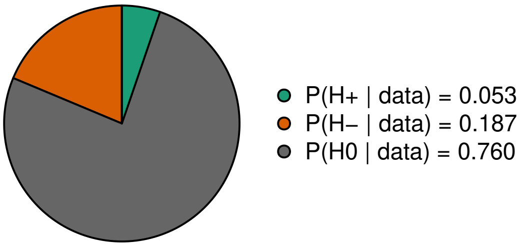

Figure 4: Posterior probabilities of the hypotheses visualized as a probability wheel for Example 1.

The **abtest** package allows users to visualize the posterior probabilities of the hypotheses by means of a probability wheel (Figure 4):

```
R> prob_wheel(ab)
```

Overall, the data support the hypothesis that the training is ineffective over the company's hypothesis that the training is highly effective. The Bayes factor for $\mathcal{H}_0$ over $\mathcal{H}_+$ equals $1/0.138 \approx 7.2$, which indicates moderate evidence (Jeffreys 1939, Appendix I).

Since the data set is of a sequential nature, it may be of interest to consider not only the result based on all observations, but to conduct also a sequential analysis that tracks the evidential flow as a function of the total number of observations (i.e., the number of observations across both groups). This sequential analysis can be conducted as follows:

```
R> plot_sequential(ab, thin = 4)
```

Setting the `thin` argument to `4` indicates that the evidence is computed after every $4th$ observation. Thinning can be useful to speed up the analysis in case the data set is very large or in case observations arrive in batches. Figure 1 displays the result of the sequential analysis. The posterior probability of each hypothesis with non-zero prior probability is plotted as a function of the total number of observations. At the top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities of the hypotheses based on all available data. Figure 1 shows that after some initial fluctuation, adding more observations increased the probability of the null hypothesis that there is no effect of the training.

4.3. Parameter estimation

The data indicate evidence in favor of the null hypothesis versus the hypothesis that the training is highly effective, leaving open the possibility that the training does have an effect, but of a more modest size than the company anticipated. To assess this possibility one may investigate the potential size of the effect under the assumption that the effect is non-zero.[8] For parameter estimation, we generally prefer to investigate the posterior distribution for the unconstrained alternative hypothesis $\mathcal{H}_1$; however, the **abtest** package also provides posterior samples and plotting functionality for the constrained hypotheses $\mathcal{H}_+$ and $\mathcal{H}_-$.

[8] For consistency, we continue this analysis with the company's prior; an analysis with the less enthusiastic default prior is provided in Appendix D.

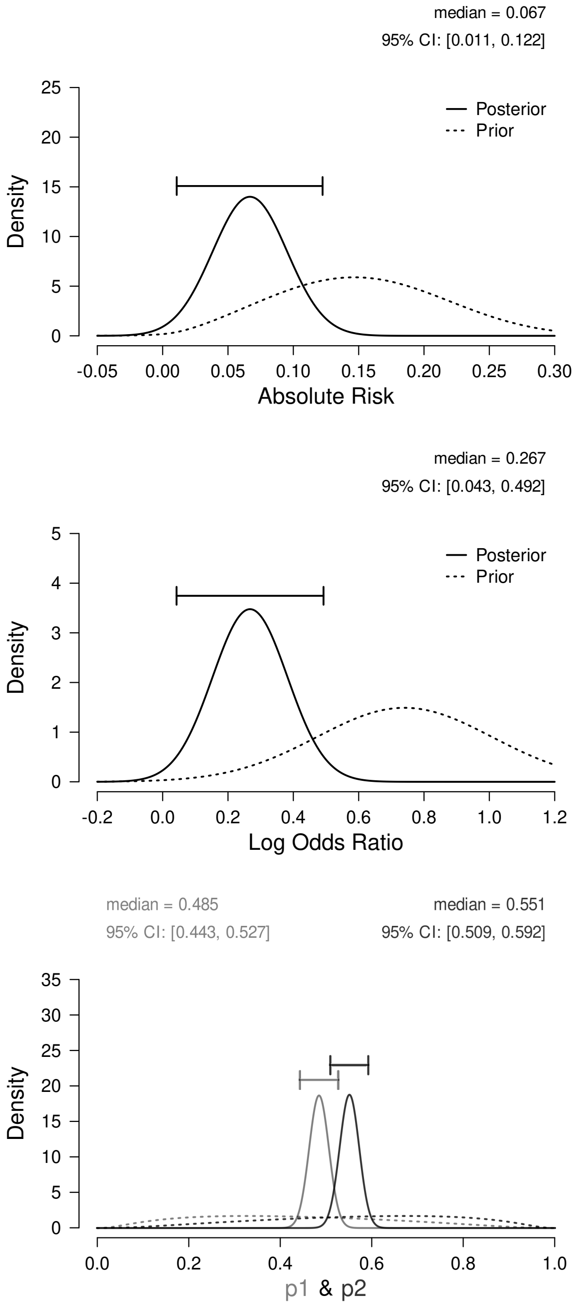

Figure 5: (Implied) prior and posterior distributions under $\mathcal{H}_1$ for Example 1. The dotted lines display the prior distributions, the solid lines display the posterior distributions (with 95% central credible intervals). The medians and the bounds of the 95% central credible intervals are displayed on top of each panel. The top panel displays the posterior distribution for the absolute risk (i.e., $p_2 - p_1$); the middle panel shows the posterior distribution for the log odds ratio parameter ψ; the bottom panel displays the marginal posterior distributions for the success probabilities p_1 and p_2.

The top panel of Figure 5 displays the posterior distribution for the absolute risk (i.e., $p_2 - p_1$) that can be obtained as follows:

```
R> plot_posterior(ab, what = "arisk")
```

The top panel of Figure 5 shows the prior distribution as a dotted line and the posterior distribution (with 95% central credible interval) as a solid line. The plot indicates that, under the assumption that the difference between the two success probabilities is not exactly zero, it is likely to be smaller than expected: The posterior median is 0.067 and the 95% central credible interval ranges from 0.011 to 0.122.

The middle panel of Figure 5 displays the posterior distribution for the log odds ratio ψ that can be obtained as follows:

```
R> plot_posterior(ab, what = "logor")
```

The middle panel of Figure 5 indicates that, given the log odds ratio is not exactly zero, it is likely to be between 0.043 and 0.492, where the posterior median is 0.267.

It may also be of interest to consider the marginal posterior distributions of the success probabilities p_1 and p_2. This plot can be produced as follows:

```
R> plot_posterior(ab, what = "p1p2")
```

The bottom panel of Figure 5 displays the resulting plot. In this example, p_1 and p_2 correspond to the probability of still being on the job after six month for the non-trained employees and the employees that received the training, respectively. The bottom panel of Figure 5 indicates that the posterior median for p_1 is 0.485, with 95% credible ranging from 0.443 to 0.527, and the posterior median for p_2 is 0.551, with 95% credible interval ranging from 0.509 to 0.592.

In sum, this fictitious data set offers modest evidence in favor of the null hypothesis which states that the training is not effective over the hypothesis that the training is highly effective; nevertheless, the consultancy firm should probably continue to collect data in order to obtain more compelling evidence before deciding whether or not the training should be implemented. If the true effect is as small as 4%, continued testing will ultimately show compelling evidence for $\mathcal{H}_+$ over $\mathcal{H}_0$. Note that continued testing is trivial in the Bayesian framework: The results can simply be updated as new observations arrive.

5. Example 2: Progesterone in women with bleeding

As a second example application of the **abtest** package, here we present a reanalysis of a recent medical trial[9]. Coomarasamy *et al.* (2019) assessed the effectiveness of progesterone in preventing miscarriages. The number of live births was 74.7% (1513/2025) in the progesterone group and 72.5% (1459/2013) in the placebo group ($p = 0.08$). The authors concluded: “The incidence of adverse events did not differ significantly between the groups” (Coomarasamy *et al.* 2019, p. 1815).

This conclusion leaves unaddressed the degree to which the data undercut or support the no-effect hypothesis $\mathcal{H}_0$ over the positive-effect hypothesis $\mathcal{H}_+$. To quantify such evidence we can use the **abtest** package. A default analysis can be conducted as follows:

[9] This reanalysis is also available in Gronau and Wagenmakers (2019).

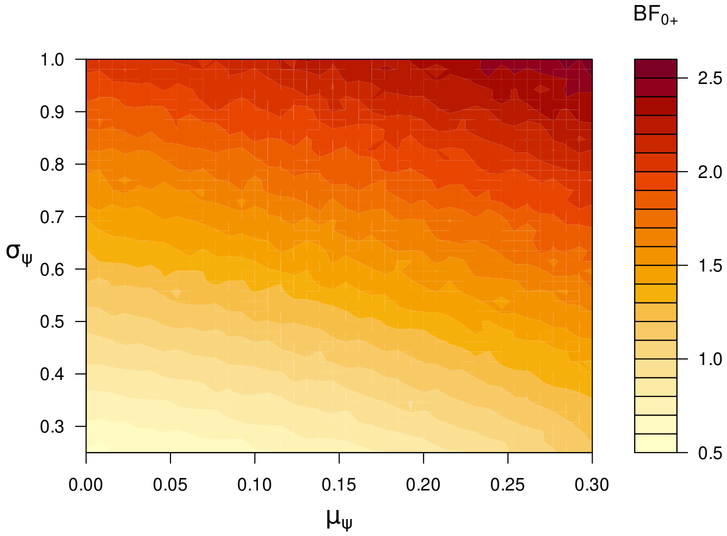

Figure 6: Prior robustness analysis for Example 2. The heat map displays the Bayes factor BF_{0+} as a function of the test-relevant prior parameters μ_ψ and σ_ψ. Across different prior settings, the evidence for the no-effect hypothesis $\mathcal{H}_0$ over the positive-effect hypothesis $\mathcal{H}_+$ is weak.

```
R> data <- list(y1 = 1459, n1 = 2013, y2 = 1513, n2 = 2025)
R> set.seed(1)
R> ab <- ab_test(data = data)
```

This yields the following output:

```
R> print(ab)

Bayesian A/B Test Results:

 Bayes Factors:

 BF10: 0.259709
 BF+0: 0.4866008
 BF-0: 0.02796485

 Prior Probabilities Hypotheses:

 H+: 0.25
 H-: 0.25
 H0: 0.5

 Posterior Probabilities Hypotheses:

 H+: 0.1935
 H-: 0.0111
 H0: 0.7954
```

A Bayes factor of $\text{BF}_{0+} = 1/\text{BF}_{+0} \approx 2$ indicates that there is only weak evidence in favor of the no-effect hypothesis $\mathcal{H}_0$ over the positive-effect hypothesis $\mathcal{H}_+$ (Jeffreys 1939). To alleviate concerns about the choice of the prior distribution for the test-relevant log odds ratio parameter ψ one can conduct a prior robustness analysis as follows:

```
R> plot_robustness(ab, bftype = "BF0+")
```

Note that the `bftype` argument is used to indicate which Bayes factor is plotted (in this case BF_{0+}). Figure 6 displays the results and shows that the evidence is weak for all combinations of $\mu_\psi \in [0, 0.30]$ and $\sigma_\psi \in [0.25, 1]$.

In sum, these data neither undercut nor support the progesterone hypothesis in compelling fashion.

6. Concluding comments

In this article, we have introduced the **abtest** package that implements both Bayesian hypothesis testing and Bayesian estimation for the A/B test using informed priors. The procedure allows users to (1) obtain evidence in favor of the null hypothesis; (2) monitor the evidence as data accumulate; and (3) elicit and incorporate expert prior distributions. We hope that the provided analysis approach is useful across different fields that apply A/B testing on a routine basis, particularly business and medicine.

We have introduced the approach implemented in **abtest** as testing hypotheses of interest about the test-relevant log odds ratio parameter ψ for the model in Equation 1. However, it should be pointed out that an alternative interpretation is to view the procedure as estimating a mixture model, where the mixture components correspond to the different hypotheses of interest, and the mixture weights are given by the prior/posterior probabilities of the hypotheses (e.g., Mitchell and Beauchamp 1988). This interpretation is illustrated with a fictitious example in Figure 7. For simplicity, the plot assumes that the user has set the prior probabilities of $\mathcal{H}_+$ and $\mathcal{H}_-$ to zero, whereas the prior probabilities of $\mathcal{H}_1$ and $\mathcal{H}_0$ are both set to 0.50. The left panel illustrates the mixture representation before having observed any data. Specifically, the height of the spike at zero corresponds to the prior probability of $\mathcal{H}_0$ whereas the shape of the slab corresponds to the continuous default prior distribution for ψ under $\mathcal{H}_1$. The maximum height of this continuous distribution corresponds to the prior probability of $\mathcal{H}_1$.[10] The right panel illustrates the mixture representation after having observed 20 successes out of 40 observations in the control condition and 30 successes out of 40 observations in the experimental condition (these are fictitious data). The height of the spike corresponds to the posterior probability of $\mathcal{H}_0$, and the maximum height of the continuous posterior distribution under $\mathcal{H}_1$ (i.e., the slab) corresponds to the posterior probability of $\mathcal{H}_1$. In this fictitious example, the data have decreased the plausibility of $\mathcal{H}_0$ and have increased the plausibility of $\mathcal{H}_1$.

Despite the practical benefits that the package offers right now, there are areas for future improvement. For instance, **abtest** currently allows users to compare two groups; however, there are applications in which one may be interested in simultaneously comparing more than two groups. Furthermore, at the moment, **abtest** expects the outcome variable to be binary.

[10] This scaling method is inspired by the **BAS** package (Clyde 2021).

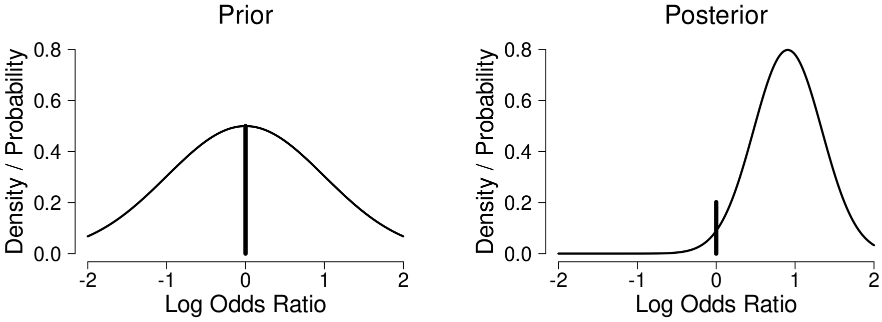

Figure 7: Mixture representation of the A/B test procedure. The left panel illustrates the mixture representation before having observed any data, the right panel illustrates the mixture representation after having observed 20 successes out of 40 observations in the control condition and 30 successes out of 40 observations in the experimental condition. The height of the spike at zero corresponds to the prior/posterior probability of $\mathcal{H}_0$ whereas the shape of the slab corresponds to the continuous default prior/posterior distribution for ψ under $\mathcal{H}_1$. The maximum height of this continuous distribution corresponds to the prior/posterior probability of $\mathcal{H}_1$.

Nevertheless, in certain scenarios, it may be more natural to compare the two groups based on a continuous outcome variable. This scenario resembles an independent samples t-test for which well-established Bayesian procedures exist (e.g., Rouder, Speckman, Sun, Morey, and Iverson 2009; Ly, Verhagen, and Wagenmakers 2016) which are available, for instance, in the **BayesFactor** package (Morey and Rouder 2018) and **JASP** (Love, Selker, Marsman, Jamil, Dropmann, Verhagen, Ly, Gronau, Šmíra, Epskamp, Matzke, Wild, Knight, Rouder, Morey, and Wagenmakers 2019).[11] Moreover, currently, the **abtest** package does not provide functions for generating predictions. Note, however, that users can generate predictions in a straightforward manner themselves based on the posterior samples that are provided by **abtest**. The implementation also does not allow users to incorporate utilities explicitly (e.g., Lindley 1985; for alternative approaches see also Azevedo, Alex, Montiel Olea, Rao, and Weyl 2019 and Feit and Berman 2019). However, again, based on the provided posterior probabilities and posterior samples, users who wish to take into account utilities may do so in a relatively straightforward way. Furthermore, users interested in adjusting the model used in **abtest** (e.g., to account for hierarchically-structured data or covariates) are referred to general-purpose Bayesian software such as Stan (Carpenter *et al.* 2017; Stan Development Team 2020) and the related R package **brms** (Bürkner 2017). In combination with the **bridgesampling** package (Gronau *et al.* 2020), this enables the user to compare custom models using Bayes factors and posterior model probabilities. A more structural limitation of **abtest** is that it has been developed to analyze A/B test data, but not to run the A/B test experiment itself.

In sum, A/B testing is ubiquitous in business and medicine. Here we have demonstrated how the **abtest** package enables relatively complete Bayesian inference including the capability to obtain support for the null, continuously monitor the results, and elicit and incorporate expert prior knowledge. Hopefully, this approach forms a basis for evidence-based conclusions that will benefit both businesses and patients.

[11] For a list of Bayesian R packages, see Park (2021).

Computational details

The results in this paper have been obtained with R 4.0.4, **g++** compiler version 10.3.0 for C++, **rstan** 2.21.2 with **StanHeaders** 2.21.0.7, **bridgesampling** 1.1.2 on a Pop OS 21.04 (based on Ubuntu 21.04) system. The random seed was set to 1. Note, however, that for replicating the results exactly, the same operating system, same compiler, and compiler version would be required for **Stan**. In other setups the results will be very similar, though, leading to the same conclusions qualitatively.

Acknowledgments

This research was supported by a Netherlands Organization for Scientific Research (NWO) grant to QFG (406.16.528) and by an NWO Vici grant to EJW (016.Vici.170.083).

References

Armitage P (1960). *Sequential Medical Trials.* Thomas, Springfield.

Azevedo EM, Alex D, Montiel Olea J, Rao JM, Weyl EG (2019). "A/B Testing with Fat Tails." *SSRN*. doi:10.2139/ssrn.3171224.

Bååth R (2014). "Bayesian First Aid: A Package That Implements Bayesian Alternatives to the Classical *.test Functions in R." In *useR! 2014 – The International R User Conference.*

Bartlett MS (1957). "A Comment on D. V. Lindley's Statistical Paradox." *Biometrika*, **44**, 533–534. doi:10.1093/biomet/44.3-4.533.

Berger JO, Delampady M (1987). "Testing Precise Hypotheses." *Statistical Science*, **2**, 317–352. doi:10.1214/ss/1177013238.

Berger JO, Wolpert RL (1988). *The Likelihood Principle.* 2nd edition. Institute of Mathematical Statistics, Hayward.

Berman R, Pekelis L, Scott A, Van den Bulte C (2018). "p-Hacking and False Discovery in A/B Testing." *SSRN*. doi:10.2139/ssrn.3204791.

Bürkner PC (2017). "**brms**: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, **80**(1), 1–28. doi:10.18637/jss.v080.i01.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, **76**(1), 1–32. doi:10.18637/jss.v076.i01.

Chen H, Cohen P, Chen S (2010). "How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies." *Communications in Statistics—Simulation and Computation®*, **39**, 860–864. doi:10.1080/03610911003650383.

Clyde M (2021). ***BAS**: Bayesian Variable Selection and Model Averaging Using Bayesian Adaptive Sampling.* R package version 1.6.0, URL https://CRAN.R-project.org/package=BAS.

Coomarasamy A, Devall AJ, Cheed V, Harb H, Middleton LJ, Gallos ID, Williams H, Eapen AK, Roberts T, Ogwulu CC, Goranitis I, Daniels JP, Ahmed A, Bender-Atik R, Bhatia K, Bottomley C, Brewin J, Choudhary M, Crosfill F, Deb S, Duncan WC, Ewer A, Hinshaw K, Holland T, Izzat F, Johns J, Kriedt K, Lumsden MA, Manda P, Norman JE, Nunes N, Overton CE, Quenby S, Rao S, Ross J, Shahid A, Underwood M, Vaithilingam N, Watkins L, Wykes C, Horne A, Jurkovic D (2019). "A Randomized Trial of Progesterone in Women with Bleeding in Early Pregnancy." *New England Journal of Medicine*, **380**, 1815–1824. doi:10.1097/01.aoa.0000603772.96889.98.

Cumming G (2014). "The New Statistics: Why and How." *Psychological Science*, **25**, 7–29. doi:10.1177/0956797613504966.

Deng A, Lu J, Chen S (2016). "Continuous Monitoring of A/B Tests Without Pain: Optional Stopping in Bayesian Testing." In *2016 IEEE International Conference on Data Science and Advanced Analytics*, pp. 243–252.

Dienes Z (2014). "Using Bayes to Get the Most out of Non-Significant Results." *Frontiers in Psycholology*, **5:781**. doi:10.3389/fpsyg.2014.00781.

Feit EM, Berman R (2019). "Test & Roll: Profit-Maximizing A/B Tests." *Marketing Science*, **38**, 1038–1058. doi:10.1287/mksc.2019.1194.

Feller W (1940). "Statistical Aspects of ESP." *Journal of Parapsychology*, **4**, 271–298. doi:10.1002/9780470317020.ch9.

Fisher RA (1928). *Statistical Methods for Research Workers*. 2nd edition. Oliver and Boyd, Edinburgh.

Gelman A, Rubin DB (1995). "Avoiding Model Selection in Bayesian Social Research." *Sociological Methodology*, **25**, 165–173. doi:10.2307/271064.

Gronau QF (2021). ***abtest**: Bayesian A/B Testing*. R package version 1.0.1, URL https://CRAN.R-project.org/package=abtest.

Gronau QF, Singmann H, Wagenmakers EJ (2020). "**bridgesampling**: An R Package for Estimating Normalizing Constants." *Journal of Statistical Software*, **92**(10), 1–29. doi:10.18637/jss.v092.i10.

Gronau QF, Wagenmakers EJ (2019). "Progesterone in Women with Bleeding in Early Pregnancy: Absence of Evidence, Not Evidence of Absence." *PsyArXiv etk7g*, PsyArXiv Preprints. doi:10.31234/osf.io/etk7g.

Haaf J, Ly A, Wagenmakers EJ (2019). "Retire Significance, But Still Test Hypotheses." *Nature*, **567**, 461. doi:10.1038/d41586-019-00972-7.

Howard JV (1998). "The 2 × 2 Table: A Discussion from a Bayesian Viewpoint." *Statistical Science*, **13**, 351–367. doi:10.1214/ss/1028905830.

Jamil T, Marsman M, Ly A, Morey RD, Wagenmakers EJ (2017). "What Are the Odds? Modern Relevance and Bayes Factor Solutions for MacAlister's Problem from the 1881 *Educational Times*." *Educational and Psychological Measurement*, **77**, 819–830. doi:10.1177/0013164416667980.

Jeffreys H (1935). "Some Tests of Significance, Treated by the Theory of Probability." *Proceedings of the Cambridge Philosophy Society*, **31**, 203–222. doi:10.1017/s030500410001330x.

Jeffreys H (1939). *Theory of Probability*. 1st edition. Oxford University Press, Oxford.

Johari R, Koomen P, Pekelis L, Walsh D (2017). "Peeking at A/B Tests: Why It Matters, and What to Do About It." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1517–1525. ACM, New York. URL http://doi.acm.org/10.1145/3097983.3097992.

Kass RE, Raftery AE (1995). "Bayes Factors." *Journal of the American Statistical Association*, **90**, 773–795. doi:10.1080/01621459.1995.10476572.

Kass RE, Vaidyanathan SK (1992). "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions." *Journal of the Royal Statistical Society B*, **54**, 129–144. doi:10.1111/j.2517-6161.1992.tb01868.x.

Keysers C, Gazzola V, Wagenmakers EJ (2020). "Using Bayes Factor Hypothesis Testing in Neuroscience to Establish Evidence of Absence." *Nature Neuroscience*, **23**, 788–799. doi:10.1038/s41593-020-0660-4.

Lindley DV (1957). "A Statistical Paradox." *Biometrika*, **44**, 187–192. doi:10.1093/biomet/44.1-2.187.

Lindley DV (1985). *Making Decisions*. 2nd edition. John Wiley & Sons, London.

Lipkus IM, Hollands JG (1999). "The Visual Communication of Risk." *Journal of the National Cancer Institute Monographs*, **25**, 149–163. doi:10.1093/oxfordjournals.jncimonographs.a024191.

Little RJA (1989). "Testing the Equality of Two Independent Binomial Proportions." *The American Statistician*, **43**, 283–288. doi:10.1080/00031305.1989.10475676.

Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau QF, Šmíra M, Epskamp S, Matzke D, Wild A, Knight P, Rouder JN, Morey RD, Wagenmakers EJ (2019). "**JASP**: Graphical Statistical Software for Common Statistical Designs." *Journal of Statistical Software*, **88**(2), 1–17. doi:10.18637/jss.v088.i02.

Ly A, Verhagen AJ, Wagenmakers EJ (2016). "Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology." *Journal of Mathematical Psychology*, **72**, 19–32. doi:10.1016/j.jmp.2015.06.004.

Malek A, Katariya S, Chow Y, Ghavamzadeh M (2017). "Sequential Multiple Hypothesis Testing with Type I Error Control." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1468–1476.

Meng XL, Wong WH (1996). "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration." *Statistica Sinica*, **6**, 831–860. doi:10.1214/ss/1028905934.

Mitchell TJ, Beauchamp JJ (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, **83**, 1023–1032. doi:10.1080/01621459.1988.10478694.

Morey RD, Rouder JN (2018). ***BayesFactor**: Computation of Bayes Factors for Common Designs.* R package version 0.9.12-4.2, URL https://CRAN.R-project.org/package=BayesFactor.

O'Hagan A (2019). "Expert Knowledge Elicitation: Subjective but Scientific." *The American Statistician*, **73**, 69–81. doi:10.1080/00031305.2018.1518265.

Park JH (2021). *CRAN Task View: Bayesian Inference.* Version 2021-11-04, URL https://CRAN.R-project.org/view=Bayesian.

Pham-Gia T, Van Thin N, Doan PP (2017). "Inferences on the Difference of Two Proportions: A Bayesian Approach." *Open Journal of Statistics*, **7**, 1–15. doi:10.4236/ojs.2017.71001.

Portman F (2021). ***bayesAB**: Fast Bayesian Methods for AB Testing.* R package version 1.1.3, URL https://CRAN.R-project.org/package=bayesAB.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stan Development Team (2020). "**rstan**: The R Interface to Stan." R package version 2.21.2, URL https://CRAN.R-project.org/package=rstan.

Robert C, Casella G (2010). *Introducing Monte Carlo Methods with R.* Springer-Verlag, New York.

Rouder JN (2014). "Optional Stopping: No Problem for Bayesians." *Psychonomic Bulletin & Review*, **21**, 301–308. doi:10.3758/s13423-014-0595-4.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009). "Bayesian T Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin & Review*, **16**, 225–237. doi:10.3758/pbr.16.2.225.

Skorski M (2019). "Bounds on Bayes Factors for Binomial A/B Testing." *arXiv 1903.00049*, arXiv.org E-Print Archive. URL https://arxiv.org/abs/1903.00049.

Stucchio C (2015). "Bayesian A/B Testing at VWO." *Technical report*, VWO. URL https://www.chrisstucchio.com/pubs/VWO_SmartStats_technical_whitepaper.pdf.

Tversky A (1969). "Intransitivity of Preferences." *Psychological Review*, **76**, 31–48. doi:10.1037/h0026750.

Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen AJ, Love J, Selker R, Gronau QF, Šmíra M, Epskamp S, Matzke D, Rouder JN, Morey RD (2018). "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications." *Psychonomic Bulletin & Review*, **25**, 35–57. doi:10.3758/s13423-017-1343-3.

Ware JH (1989). "Investigating Therapies of Potentially Great Benefit: ECMO." *Statistical Science*, **4**, 298–340. doi:10.1214/ss/1177012384.

A. Interpretation of the parameters

Here we show that β corresponds to the grand mean of the log odds and that ψ corresponds to the log odds ratio (for the model definition, see Equation 1). The nuisance parameter β corresponds to the grand mean of the log odds since

$$\frac{1}{2}\log\left(\frac{p_1}{1-p_1}\right)+\frac{1}{2}\log\left(\frac{p_2}{1-p_2}\right)=\frac{1}{2}\beta-\frac{1}{4}\psi+\frac{1}{2}\beta+\frac{1}{4}\psi=\beta.$$

The test-relevant parameter ψ corresponds to the log odds ratio since

$$\log\left(\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}\right)=\log\left(\frac{p_2}{1-p_2}\right)-\log\left(\frac{p_1}{1-p_1}\right)=\beta+\frac{\psi}{2}-\left(\beta-\frac{\psi}{2}\right)=\psi.$$

B. Prior elicitation: Implied distributions

The prior elicitation approach described in Equation 2 requires the CDF's for the quantities of interest. Here, we derive the implied CDF's for these quantities; we also derive the corresponding probability density functions (PDF's). Additionally, we derive four further implied distributions of interest: The joint PDF of p_1 and p_2, the conditional PDF of p_2 given p_1 is fixed to a particular value, the marginal distribution for p_1, and the marginal distribution for p_2. A few of these expressions will contain a one-dimensional integral which can easily be evaluated using numerical integration.

B.1. Log odds ratio

Since ψ itself corresponds to the log odds ratio, $F(\cdot;\mu_\psi,\sigma_\psi)$ corresponds in this case to the CDF of a normal distribution with mean μ_ψ and standard deviation σ_ψ. The corresponding PDF is the normal probability density function.

B.2. Odds ratio

The implied prior on the odds ratio $\omega = \exp(\psi)$ is a log-normal distribution. Hence, $F(\cdot;\mu_\psi,\sigma_\psi)$ corresponds in this case to the CDF of a log-normal distribution with parameters μ_ψ and σ_ψ. The corresponding PDF is the log-normal probability density function.

B.3. Relative risk

The relative risk is given by $\Lambda = \frac{p_2}{p_1}$. We use a capital letter (i.e., Λ) to refer to the random variable and use a lower-case letter (i.e., λ) to refer to a concrete realization. Note that so far, we have abused notation by only using lower-case letters, but it should be clear from the context when we referred to a random variable or a concrete realization. However, for deriving the following CDF, we need the distinction to keep the notation clear. To derive the implied CDF for the relative risk, we proceed as follows:

$$\begin{aligned}
\mathsf{P}(\Lambda \leq \lambda) &= \mathsf{P}\left(\frac{p_2}{p_1} \leq \lambda\right) \\
&= \mathsf{P}\left(p_2 \leq \lambda p_1\right) \\
&= \mathsf{P}\left(\frac{1}{1+\exp\left(-\beta-\frac{\psi}{2}\right)} \leq \frac{\lambda}{1+\exp\left(-\beta+\frac{\psi}{2}\right)}\right).
\end{aligned}$$

Taking reciprocals and some algebra yields

$$\mathsf{P}\left(\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1-\lambda)\exp(\beta)\exp\left(\frac{\psi}{2}\right) - \lambda \leq 0\right).$$

When we set

$$\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1-\lambda)\exp(\beta)\exp\left(\frac{\psi}{2}\right) - \lambda = 0,$$

we can solve for ψ using the fact that this is a quadratic equation in $\exp\left(\frac{\psi}{2}\right)$ and we obtain:

$$\exp\left(\frac{\psi}{2}\right) = \frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2},$$

where we took into account that $\exp\left(\frac{\psi}{2}\right)$ needs to be positive (i.e., we omitted the solution corresponding to minus the square root). Hence,

$$\psi = 2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2}\right).$$

Therefore, $\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1-\lambda)\exp(\beta)\exp\left(\frac{\psi}{2}\right) - \lambda \leq 0$ whenever

$$\psi \leq 2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2}\right).$$

Hence, the desired CDF can be written as

$$\begin{aligned}
&\mathsf{P}\left(\psi \leq 2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2}\right)\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{2\log\left(\frac{-(1-\lambda)\exp(\beta)+\sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2}\right)} \mathcal{N}(\psi;\mu_\psi,\sigma_\psi^2)\mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\mathrm{d}\psi\mathrm{d}\beta \\
&= \int_{-\infty}^{\infty} \mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\,\Phi\left(2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}{2}\right);\mu_\psi,\sigma_\psi^2\right)\mathrm{d}\beta,
\end{aligned} \tag{10}$$

where $\Phi\left(\cdot; \mu_\psi, \sigma_\psi^2\right)$ denotes the CDF of a normal distribution with mean μ_ψ and variance σ_ψ^2, and $\mathcal{N}(\cdot; \mu_\beta, \sigma_\beta^2)$ denotes the corresponding PDF.

The PDF of the relative risk is obtained by taking the derivative with respect to λ:

$$\begin{aligned}
&\frac{d}{d\lambda}\left[\int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2)\, \Phi\left(2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta) + 4\lambda}}{2}\right); \mu_\psi, \sigma_\psi^2\right) \mathrm{d}\beta\right] \\
&\quad = \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2)\, \mathcal{N}\left(2\log\left(\frac{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta) + 4\lambda}}{2}\right); \mu_\psi, \sigma_\psi^2\right) \\
&\qquad\quad \times 2\left[\frac{\exp(\beta) + \frac{2-(1-\lambda)\exp(2\beta)}{\sqrt{(1-\lambda)^2\exp(2\beta)+4\lambda}}}{-(1-\lambda)\exp(\beta) + \sqrt{(1-\lambda)^2\exp(2\beta) + 4\lambda}}\right] \mathrm{d}\beta.
\end{aligned} \tag{11}$$

B.4. Absolute risk

The absolute risk is given by $\Upsilon = p_2 - p_1$. We use the upper-case letter Υ to refer to the random variable and the lower-case letter υ to refer to a concrete realization. To derive the implied CDF for the absolute risk, we proceed as follows:

$$\begin{aligned}
\mathsf{P}(\Upsilon \leq \upsilon) &= \mathsf{P}\left(p_2 - p_1 \leq \upsilon\right) \\
&= \mathsf{P}\left(p_2 \leq \upsilon + p_1\right) \\
&= \mathsf{P}\left(\frac{1}{1 + \exp\left(-\beta - \frac{\psi}{2}\right)} \leq \upsilon + \frac{1}{1 + \exp\left(-\beta + \frac{\psi}{2}\right)}\right).
\end{aligned}$$

After some algebra, we obtain

$$\mathsf{P}\left(\exp(\beta)(1-\upsilon)\left(\exp\left(\frac{\psi}{2}\right)\right)^2 - \upsilon\left(\exp(2\beta) + 1\right)\exp\left(\frac{\psi}{2}\right) - \exp(\beta)(\upsilon + 1) \leq 0\right).$$

When we set

$$\exp(\beta)(1-\upsilon)\left(\exp\left(\frac{\psi}{2}\right)\right)^2 - \upsilon\left(\exp(2\beta) + 1\right)\exp\left(\frac{\psi}{2}\right) - \exp(\beta)(\upsilon + 1) = 0,$$

we can solve for ψ using the fact that this is a quadratic equation in $\exp\left(\frac{\psi}{2}\right)$ and we obtain:

$$\exp\left(\frac{\psi}{2}\right) = \frac{\upsilon\left(\exp(2\beta) + 1\right) + \sqrt{\upsilon^2\left(\exp(2\beta) - 1\right)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)},$$

where we took into account that $\exp\left(\frac{\psi}{2}\right)$ needs to be positive (i.e., we omitted the solution corresponding to minus the square root). Hence,

$$\psi = 2\log\left(\frac{\upsilon\left(\exp(2\beta) + 1\right) + \sqrt{\upsilon^2\left(\exp(2\beta) - 1\right)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right).$$

Therefore, $\exp(\beta)(1-\upsilon)\left(\exp\left(\frac{\psi}{2}\right)\right)^2 - \upsilon(\exp(2\beta)+1)\exp\left(\frac{\psi}{2}\right) - \exp(\beta)(\upsilon+1) \leq 0$ whenever

$$\psi \leq 2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right).$$

Hence, the desired CDF can be written as

$$\begin{aligned}
&\mathsf{P}\left(\psi \leq 2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right)\right)\\
&=\int_{-\infty}^{\infty}\int_{-\infty}^{2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right)} \mathcal{N}(\psi;\mu_\psi,\sigma_\psi^2)\mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\mathrm{d}\psi\mathrm{d}\beta\\
&=\int_{-\infty}^{\infty}\mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\,\Phi\left(2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right);\mu_\psi,\sigma_\psi^2\right)\mathrm{d}\beta.
\end{aligned} \tag{12}$$

The PDF of the absolute risk is obtained by taking the derivative with respect to υ:

$$\begin{aligned}
&\frac{d}{d\upsilon}\left[\int_{-\infty}^{\infty}\mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\,\Phi\left(2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right);\mu_\psi,\sigma_\psi^2\right)\mathrm{d}\beta\right]\\
&=\int_{-\infty}^{\infty}\mathcal{N}(\beta;\mu_\beta,\sigma_\beta^2)\,\mathcal{N}\left(2\log\left(\frac{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}{2\exp(\beta)(1-\upsilon)}\right);\mu_\psi,\sigma_\psi^2\right)\\
&\qquad\times 2\left[\frac{\exp(2\beta)+\frac{\upsilon(\exp(2\beta)-1)^2}{\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}+1}{\upsilon(\exp(2\beta)+1)+\sqrt{\upsilon^2(\exp(2\beta)-1)^2+4\exp(2\beta)}}+\frac{1}{1-\upsilon}\right]\mathrm{d}\beta.
\end{aligned} \tag{13}$$

B.5. Joint distribution of p_1 and p_2

Another distribution of interest is the implied joint distribution of the two success probabilities p_1 and p_2. This distribution will not be used to elicit the prior on ψ which is the reason why we only derive the PDF and not the CDF. The model parameters β and ψ are related to p_1 and p_2 as follows:

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2}\\
\log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2}.
\end{aligned}$$

Hence, the inverse transformation is given by:

$$\begin{aligned}
\beta &= \frac{1}{2}\log\left(\frac{p_1}{1-p_1}\right) + \frac{1}{2}\log\left(\frac{p_2}{1-p_2}\right)\\
\psi &= \log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right).
\end{aligned}$$

The corresponding Jacobian is:

$$
\begin{aligned}
|J| &= \left| \begin{pmatrix} \frac{\partial \beta}{\partial p_1} & \frac{\partial \beta}{\partial p_2} \\ \frac{\partial \psi}{\partial p_1} & \frac{\partial \psi}{\partial p_2} \end{pmatrix} \right| \\
&= \left| \begin{pmatrix} \frac{1}{2}\frac{1}{p_1(1-p_1)} & \frac{1}{2}\frac{1}{p_2(1-p_2)} \\ -\frac{1}{p_1(1-p_1)} & \frac{1}{p_2(1-p_2)} \end{pmatrix} \right| \\
&= \frac{1}{p_1 p_2 (1-p_1)(1-p_2)}.
\end{aligned}
$$

Therefore, the joint PDF of p_1 and p_2 is given by:

$$
\begin{aligned}
p(p_1, p_2) = \frac{1}{p_1 p_2 (1-p_1)(1-p_2)} & \mathcal{N}\left(\frac{1}{2}\left[\log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right)\right]; \mu_\beta, \sigma_\beta^2\right) \\
& \times \mathcal{N}\left(\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right).
\end{aligned} \tag{14}
$$

B.6. Marginal distribution of p_1

The marginal distribution of p_1 is given by:

$$
\begin{aligned}
p(p_1) &= \int_0^1 p(p_1, p_2')\mathrm{d}p_2' \\
&= \int_0^1 \frac{1}{p_1 p_2'(1-p_1)(1-p_2')} \mathcal{N}\left(\frac{1}{2}\left[\log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2'}{1-p_2'}\right)\right]; \mu_\beta, \sigma_\beta^2\right) \\
&\qquad \times \mathcal{N}\left(\log\left(\frac{p_2'}{1-p_2'}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right) \mathrm{d}p_2'.
\end{aligned} \tag{15}
$$

B.7. Marginal distribution of p_2

The marginal distribution of p_2 is given by:

$$
\begin{aligned}
p(p_2) &= \int_0^1 p(p_1', p_2)\mathrm{d}p_1' \\
&= \int_0^1 \frac{1}{p_1' p_2(1-p_1')(1-p_2)} \mathcal{N}\left(\frac{1}{2}\left[\log\left(\frac{p_1'}{1-p_1'}\right) + \log\left(\frac{p_2}{1-p_2}\right)\right]; \mu_\beta, \sigma_\beta^2\right) \\
&\qquad \times \mathcal{N}\left(\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1'}{1-p_1'}\right); \mu_\psi, \sigma_\psi^2\right) \mathrm{d}p_1'.
\end{aligned} \tag{16}
$$

B.8. Conditional distribution of p_2 given p_1

Another distribution of interest is the conditional distribution of the second success probability p_2 given a particular value of p_1. This distribution will not be used for prior elicitation which is the reason why we only present the expression for the PDF which is given by:

$$\begin{aligned} p(p_2 \mid p_1) &= \frac{p(p_1, p_2)}{\int_0^1 p(p_1, p_2')\mathrm{d}p_2'} \\ &= \frac{\frac{1}{p_2(1-p_2)} \mathcal{N}\left(\frac{1}{2}\left[\log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right)\right]; \mu_\beta, \sigma_\beta^2\right) \mathcal{N}\left(\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right)}{\int_0^1 \frac{1}{p_2'(1-p_2')} \mathcal{N}\left(\frac{1}{2}\left[\log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2'}{1-p_2'}\right)\right]; \mu_\beta, \sigma_\beta^2\right) \mathcal{N}\left(\log\left(\frac{p_2'}{1-p_2'}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right) \mathrm{d}p_2'}. \end{aligned} \tag{17}$$

B.9. Implied distributions for truncated priors on the log odds ratio

Note that the above expressions can be all easily modified in case the prior on the log odds ratio ψ is a truncated normal distribution (e.g., restricting ψ to be larger/smaller than zero) which is the case for the hypotheses $\mathcal{H}_+$ and $\mathcal{H}_-$. In this case, the normal prior density function and cumulative distribution function for ψ simply need to be replaced by the truncated versions. For the implied log-normal prior on the odds ratio, the truncation bounds simply need to be exponentiated to obtain the truncation bounds with respect to the log-normal prior.

C. Laplace approximation details

The Laplace approximations require first-order and second-order derivatives. Let us first state explicitly the functions for which we need to find the derivatives. For $\mathcal{H}_0$ we have:

$$\begin{aligned} l_0^*(\beta) &= \log\left\{p(y \mid \beta)\, \pi_0(\beta)\right\} \\ &= (y_1 + y_2)\log\left(\frac{\exp(\beta)}{1+\exp(\beta)}\right) + (n_1 + n_2 - y_1 - y_2)\log\left(1 - \frac{\exp(\beta)}{1+\exp(\beta)}\right) \\ &\qquad - \frac{1}{2}\log\left(2\pi\sigma_\beta^2\right) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2. \end{aligned} \tag{18}$$

For $\mathcal{H}_1$ we have:

$$\begin{aligned} l^*(\beta, \psi) &= \log\left\{p(y \mid \beta, \psi)\, \pi(\beta, \psi)\right\} \\ &= y_1 \log\left(\frac{\exp(\beta - \frac{\psi}{2})}{1 + \exp(\beta - \frac{\psi}{2})}\right) + (n_1 - y_1)\log\left(1 - \frac{\exp(\beta - \frac{\psi}{2})}{1 + \exp(\beta - \frac{\psi}{2})}\right) \\ &\qquad + y_2 \log\left(\frac{\exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})}\right) + (n_2 - y_2)\log\left(1 - \frac{\exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})}\right) \\ &\qquad - \frac{1}{2}\log\left(2\pi\sigma_\beta^2\right) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 - \frac{1}{2}\log\left(2\pi\sigma_\psi^2\right) - \frac{1}{2\sigma_\psi^2}(\psi - \mu_\psi)^2. \end{aligned} \tag{19}$$

For $\mathcal{H}_+$ we have:

$$\begin{aligned}
l_+^*(\beta,\xi) &= \log\{p(y \mid \beta,\xi)\,\pi_+(\beta,\xi)\} \\
&= y_1 \log\left(\frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1+\exp(\beta - \frac{\exp(\xi)}{2})}\right) + (n_1 - y_1)\log\left(1 - \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1+\exp(\beta - \frac{\exp(\xi)}{2})}\right) \\
&\quad + y_2 \log\left(\frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1+\exp(\beta + \frac{\exp(\xi)}{2})}\right) + (n_2 - y_2)\log\left(1 - \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1+\exp(\beta + \frac{\exp(\xi)}{2})}\right) \\
&\quad - \frac{1}{2}\log\left(2\pi\sigma_\beta^2\right) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 \\
&\quad - \frac{1}{2}\log\left(2\pi\sigma_\psi^2\right) - \frac{1}{2\sigma_\psi^2}(\exp(\xi) - \mu_\psi)^2 - \log(1 - \Phi\left(0;\mu_\psi,\sigma_\psi^2\right)) + \xi.
\end{aligned} \tag{20}$$

Finally, for $\mathcal{H}_-$ we have

$$\begin{aligned}
l_-^*(\beta,\xi) &= \log\{p(y \mid \beta,\xi)\,\pi_-(\beta,\xi)\} \\
&= y_1 \log\left(\frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1+\exp(\beta + \frac{\exp(\xi)}{2})}\right) + (n_1 - y_1)\log\left(1 - \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1+\exp(\beta + \frac{\exp(\xi)}{2})}\right) \\
&\quad + y_2 \log\left(\frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1+\exp(\beta - \frac{\exp(\xi)}{2})}\right) + (n_2 - y_2)\log\left(1 - \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1+\exp(\beta - \frac{\exp(\xi)}{2})}\right) \\
&\quad - \frac{1}{2}\log\left(2\pi\sigma_\beta^2\right) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 \\
&\quad - \frac{1}{2}\log\left(2\pi\sigma_\psi^2\right) - \frac{1}{2\sigma_\psi^2}(-\exp(\xi) - \mu_\psi)^2 - \log(\Phi\left(0;\mu_\psi,\sigma_\psi^2\right)) + \xi.
\end{aligned} \tag{21}$$

C.1. First-order derivatives

The first-order derivatives are used to find the modes for the Laplace approximations. As shown below, we can find these derivatives analytically; however, setting the derivatives equal to zero and solving for the parameters is not straightforward. Nevertheless, having these derivatives is useful not only as an intermediate step to finding the second-order derivatives but also for finding the modes: This allows us to provide numerical optimizers with the analytic expressions for the derivatives which can increase speed and accuracy for numerically finding the modes of the relevant functions.

The first-order derivative for $l_0(\beta)$ is given by:

$$\frac{d}{d\beta}\, l_0^*(\beta) = \frac{y_1 + y_2 - (n_1 + n_2 - y_1 - y_2)\exp(\beta)}{1+\exp(\beta)} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}. \tag{22}$$

The first-order partial derivatives for $l^*(\beta,\psi)$ are given by

$$\frac{\partial}{\partial\beta}\, l^*(\beta,\psi) = \frac{y_1 - (n_1 - y_1)\exp(\beta - \frac{\psi}{2})}{1+\exp(\beta - \frac{\psi}{2})} + \frac{y_2 - (n_2 - y_2)\exp(\beta + \frac{\psi}{2})}{1+\exp(\beta + \frac{\psi}{2})} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \tag{23}$$

and

$$\frac{\partial}{\partial \psi}\, l^*(\beta, \psi) = \frac{1}{2}\left(\frac{(n_1 - y_1)\exp(\beta - \frac{\psi}{2}) - y_1}{1 + \exp(\beta - \frac{\psi}{2})} + \frac{y_2 - (n_2 - y_2)\exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})}\right) - \frac{\psi - \mu_\psi}{\sigma_\psi^2}. \tag{24}$$

The first-order partial derivatives for $l_+^*(\beta, \xi)$ are given by:

$$\frac{\partial}{\partial \beta}\, l_+^*(\beta, \xi) = \frac{y_1 - (n_1 - y_1)\exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2)\exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \tag{25}$$

and

$$\begin{aligned}\frac{\partial}{\partial \xi}\, l_+^*(\beta, \xi) = \frac{\exp(\xi)}{2}&\left(\frac{(n_1 - y_1)\exp(\beta - \frac{\exp(\xi)}{2}) - y_1}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} + \frac{y_2 - (n_2 - y_2)\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})}\right)\\ &- \exp(\xi)\frac{\exp(\xi) - \mu_\psi}{\sigma_\psi^2} + 1.\end{aligned} \tag{26}$$

The first-order partial derivatives for $l_-^*(\beta, \xi)$ are given by:

$$\frac{\partial}{\partial \beta}\, l_-^*(\beta, \xi) = \frac{y_1 - (n_1 - y_1)\exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2)\exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \tag{27}$$

and

$$\begin{aligned}\frac{\partial}{\partial \xi}\, l_-^*(\beta, \xi) = \frac{\exp(\xi)}{2}&\left(\frac{y_1 - (n_1 - y_1)\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} + \frac{(n_2 - y_2)\exp(\beta - \frac{\exp(\xi)}{2}) - y_2}{1 + \exp(\beta - \frac{\exp(\xi)}{2})}\right)\\ &+ \exp(\xi)\frac{-\exp(\xi) - \mu_\psi}{\sigma_\psi^2} + 1.\end{aligned} \tag{28}$$

C.2. Second-order derivatives

For the Laplace approximations, we also need the inverse of the negative Hessians. The Hessian is the matrix with the second-order partial derivatives which is the reason why we now present expressions for the second-order partial derivatives. Note that under all hypotheses there are either one or two parameters. Hence, the Hessians will be at most 2 by 2 matrices. For matrices up to 2 by 2, it is straightforward to find the inverse and the determinant which makes it easy to obtain the quantities needed for the Laplace approximations once we have the required derivatives.

For $l_0^*(\beta)$, there is only one parameter and the second-order derivative is given by:

$$\frac{d^2}{d\beta^2}\, l_0^*(\beta) = -\frac{(n_1 + n_2)\exp(\beta)}{(1 + \exp(\beta))^2} - \frac{1}{\sigma_\beta^2}. \tag{29}$$

For $l^*(\beta, \psi)$ the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial \beta^2} l^*(\beta, \psi) = -\frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} - \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} - \frac{1}{\sigma_\beta^2}, \tag{30}$$

and

$$\frac{\partial^2}{\partial \beta \partial \psi} l^*(\beta, \psi) = \frac{1}{2} \left(\frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} - \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} \right), \tag{31}$$

and

$$\frac{\partial^2}{\partial \psi^2} l^*(\beta, \psi) = -\frac{1}{4} \left(\frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} + \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} \right) - \frac{1}{\sigma_\psi^2}. \tag{32}$$

For $l_+^*(\beta, \xi)$ the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial \beta^2} l_+^*(\beta, \xi) = -\frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{\sigma_\beta^2}, \tag{33}$$

and

$$\frac{\partial^2}{\partial \beta \partial \xi} l_+^*(\beta, \xi) = \frac{\exp(\xi)}{2} \left(\frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} \right), \tag{34}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial \xi^2} l_+^*(\beta, \xi) = \frac{\exp(\xi)}{2} \Bigg(& \frac{(n_1 - y_1) \exp(\beta - \frac{\exp(\xi)}{2}) - y_1}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} + \frac{y_2 - (n_2 - y_2) \exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \\ & - \frac{1}{2} \exp(\xi) \frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{2} \exp(\xi) \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} \Bigg) \\ & - \exp(\xi) \frac{2\exp(\xi) - \mu_\psi}{\sigma_\psi^2}.\end{aligned} \tag{35}$$

For $l_-^*(\beta, \xi)$ the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial \beta^2} l_-^*(\beta, \xi) = -\frac{n_1 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{\sigma_\beta^2}, \tag{36}$$

and

$$\frac{\partial^2}{\partial \beta \partial \xi} l_-^*(\beta, \xi) = -\frac{\exp(\xi)}{2} \left(\frac{n_1 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} \right), \tag{37}$$

and

$$\frac{\partial^2}{\partial \xi^2}\, l_-^*(\beta,\xi) = -\frac{\exp(\xi)}{2}\Bigg(\frac{(n_1-y_1)\exp(\beta+\frac{\exp(\xi)}{2})-y_1}{1+\exp(\beta+\frac{\exp(\xi)}{2})}+\frac{y_2-(n_2-y_2)\exp(\beta-\frac{\exp(\xi)}{2})}{1+\exp(\beta-\frac{\exp(\xi)}{2})}$$
$$-\frac{1}{2}\exp(\xi)\frac{n_1\exp\left(\beta+\frac{\exp(\xi)}{2}\right)}{\left(1+\exp\left(\beta+\frac{\exp(\xi)}{2}\right)\right)^2}-\frac{1}{2}\exp(\xi)\frac{n_2\exp\left(\beta-\frac{\exp(\xi)}{2}\right)}{\left(1+\exp\left(\beta-\frac{\exp(\xi)}{2}\right)\right)^2}\Bigg)$$
$$+\exp(\xi)\frac{2\exp(\xi)-\mu_\psi}{\sigma_\psi^2}. \tag{38}$$

C.3. Hessians

Having derived the relevant second-order partial derivatives, we can simply build the Hessian matrices of interest by inserting the relevant expressions. Next, we present symbolically the Hessians of interest, that is, we show which of the second-order partial derivatives need to be inserted where. Note that we omit the one for $\mathcal{H}_0$ since this is a single number which is simply the second-order derivative of $l_0^*(\beta)$.

The Hessian for $\mathcal{H}_1$ is given by:

$$\boldsymbol{H}_1 = \begin{pmatrix} \frac{\partial^2}{\partial\beta^2}\, l^*(\beta,\psi) & \frac{\partial^2}{\partial\beta\partial\psi}\, l^*(\beta,\psi) \\ \frac{\partial^2}{\partial\beta\partial\psi}\, l^*(\beta,\psi) & \frac{\partial^2}{\partial\psi^2}\, l^*(\beta,\psi) \end{pmatrix}. \tag{39}$$

The Hessian for $\mathcal{H}_+$ is given by:

$$\boldsymbol{H}_+ = \begin{pmatrix} \frac{\partial^2}{\partial\beta^2}\, l_+^*(\beta,\xi) & \frac{\partial^2}{\partial\beta\partial\xi}\, l_+^*(\beta,\xi) \\ \frac{\partial^2}{\partial\beta\partial\xi}\, l_+^*(\beta,\xi) & \frac{\partial^2}{\partial\xi^2}\, l_+^*(\beta,\xi) \end{pmatrix}. \tag{40}$$

The Hessian for $\mathcal{H}_-$ is given by:

$$\boldsymbol{H}_- = \begin{pmatrix} \frac{\partial^2}{\partial\beta^2}\, l_-^*(\beta,\xi) & \frac{\partial^2}{\partial\beta\partial\xi}\, l_-^*(\beta,\xi) \\ \frac{\partial^2}{\partial\beta\partial\xi}\, l_-^*(\beta,\xi) & \frac{\partial^2}{\partial\xi^2}\, l_-^*(\beta,\xi) \end{pmatrix}. \tag{41}$$

Computing the inverse of the negative Hessians

Note that computing the inverses of the 2 by 2 negative Hessians is straightforward: We simply need to attach minus signs to each element of the Hessians and then make use of the fact that the inverse of a 2 by 2 matrix $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given by $\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, where $\det(\boldsymbol{A}) = ad - bc$.

D. Example 1: Effectiveness of resilience training (defaults)

Here we present the results for the resilience training example obtained using the default prior setting.

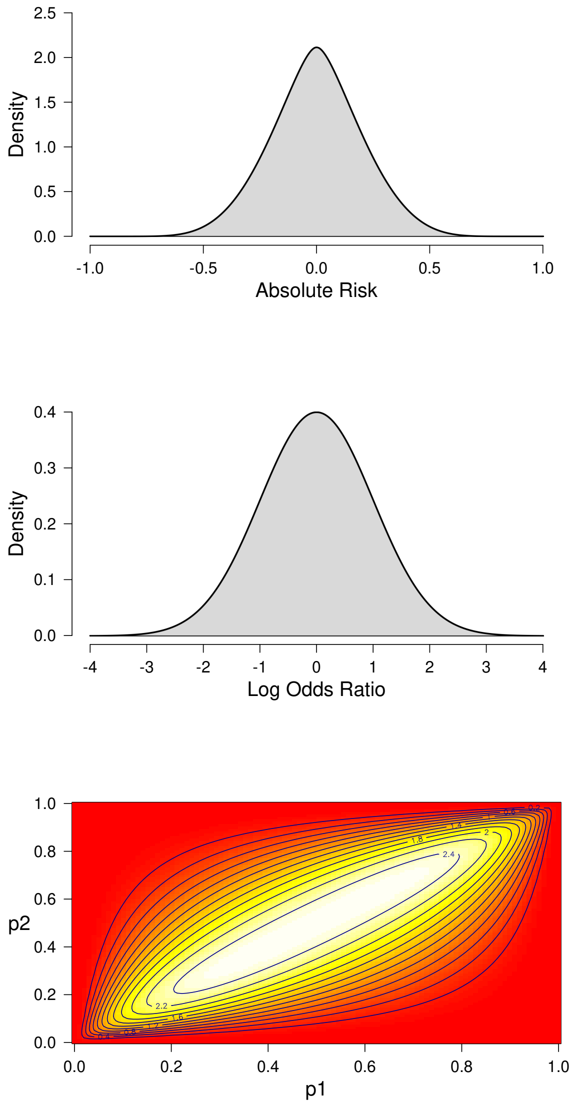

Figure 8: Default (implied) prior distributions. The top panel displays the prior distribution for the absolute risk which corresponds to the difference between the probability of still being on the job for the trained and the non-trained employees (i.e., $p_2 - p_1$). The middle panel shows the prior distribution for the log odds ratio parameter ψ. The bottom panel displays the implied joint prior distribution for the success probabilities p_1 and p_2. The bottom panel illustrates that the two success probabilities are assigned dependent priors.

D.1. Prior specification

We use the default prior setting in the **abtest** package that assigns both β and ψ standard normal prior distributions. The implied prior on the absolute risk can be visualized as follows:

```
R> library("abtest")
```

```
R> plot_prior(what = "arisk")
```

The resulting graph is shown in the top panel of Figure 8.

The user can also visualize the (implied) prior for other quantities. For instance, the prior on the log odds ratio (middle panel of Figure 8) is obtained as follows:

```
R> plot_prior(what = "logor")
```

The implied prior on the success probabilities p_1 and p_2 (bottom panel of Figure 8) is obtained as follows:

```
R> plot_prior(what = "p1p2")
```

The bottom panel of Figure 8 illustrates that there is a dependency between p_1 and p_2 which is arguably desirable (Howard 1998): When one of the success probabilities is very (small) large, it is likely that the other one will also be (small) large.

D.2. Hypothesis testing

The `ab_test()` function can be used to conduct a Bayesian A/B test using the default prior setting as follows:

```
R> data("seqdata", package = "abtest")
R> set.seed(1)
R> ab_default <- ab_test(data = seqdata)
```

This yields the following output:

```
R> print(ab_default)

Bayesian A/B Test Results:

Bayes Factors:

BF10: 0.2767214
BF+0: 0.4890489
BF-0: 0.05778357

Prior Probabilities Hypotheses:

H+: 0.25
H-: 0.25
H0: 0.5

Posterior Probabilities Hypotheses:

H+: 0.192
H-: 0.0227
H0: 0.7853
```

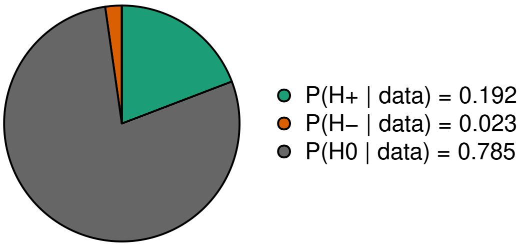

Figure 9: Posterior probabilities of the hypotheses visualized as a probability wheel.

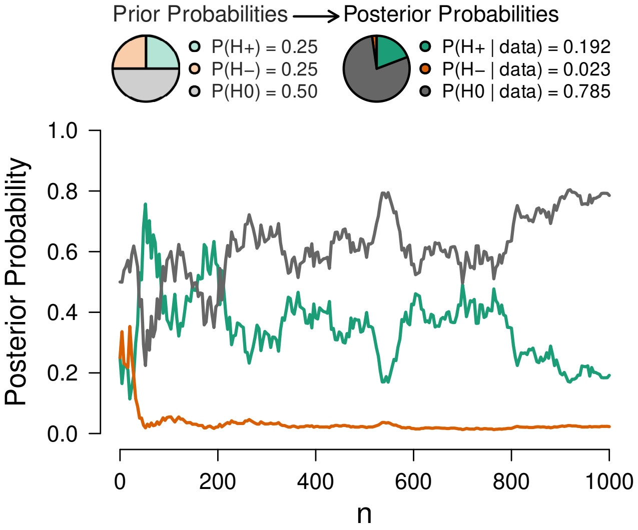

Figure 10: Sequential analysis results. The posterior probability of each hypothesis is plotted as a function of the number of observations across groups. On top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities after taking into account all observations.

The first part of the output presents Bayes factors in favor of the hypotheses $\mathcal{H}_1$, $\mathcal{H}_+$, and $\mathcal{H}_-$, where the reference hypothesis (i.e., denominator of the Bayes factor) is $\mathcal{H}_0$. Since all three Bayes factors are smaller than 1, they all indicate evidence in favor of the null hypothesis of no effect. The next part of the output displays the prior probabilities of the hypotheses with non-zero prior probability. The final part of the output displays the posterior probabilities of the hypotheses with non-zero prior probability. The posterior probability of the null hypothesis $\mathcal{H}_0$ indicates that the data have increased the plausibility of the null hypothesis from 0.50 to 0.79. Furthermore, the data have decreased the plausibility of both $\mathcal{H}_+$ and $\mathcal{H}_-$.

The **abtest** package allows users to visualize the posterior probabilities of the hypotheses by means of a probability wheel (Figure 9):

```
R> prob_wheel(ab_default)
```

Overall, the data support the hypothesis that the training is ineffective over the hypothesis that the training has a positive effect. The Bayes factor for $\mathcal{H}_0$ over $\mathcal{H}_+$ equals $1/0.489 \approx 2.04$; however, this indicates only anecdotal evidence (Jeffreys 1939, Appendix I).

Since the data set is of a sequential nature, it may be of interest to consider not only the result based on all observations, but to conduct also a sequential analysis that tracks the evidential flow as a function of the total number of observations (i.e., the number of observations across both groups). This sequential analysis can be conducted as follows:

```
R> plot_sequential(ab_default, thin = 4)
```

Figure 10 displays the result of the sequential analysis. The sequential analysis indicates that after some initial fluctuation, adding more observations increased the probability of the null hypothesis that there is no effect of the training.

D.3. Parameter estimation

The data indicate only anecdotal evidence in favor of the null hypothesis versus the hypothesis that the training is effective, leaving open the possibility that the training does have an effect. To assess this possibility one may investigate the potential size of the effect under the assumption that the effect is non-zero. For parameter estimation, we generally prefer to investigate the posterior distribution for the unconstrained alternative hypothesis $\mathcal{H}_1$.

The top panel of Figure 11 displays the posterior distribution for the absolute risk (i.e., $p_2 - p_1$) that can be obtained as follows:

```
R> plot_posterior(ab_default, what = "arisk")
```

The top panel of Figure 11 shows the prior distribution as a dotted line and the posterior distribution (with 95% central credible interval) as a solid line. The plot indicates that, under the assumption that the difference between the two success probabilities is not exactly zero, the posterior median is 0.039 and the 95% central credible interval ranges from -0.022 to 0.101.

The middle panel of Figure 11 displays the posterior distribution for the log odds ratio ψ that can be obtained as follows:

```
R> plot_posterior(ab_default, what = "logor")
```

The middle panel of Figure 11 indicates that, given the log odds ratio is not exactly zero, it is likely to be between -0.089 and 0.406, where the posterior median is 0.159.

It may also be of interest to consider the marginal posterior distributions of the success probabilities p_1 and p_2. This plot can be produced as follows:

```
R> plot_posterior(ab_default, what = "p1p2")
```

The bottom panel of Figure 11 displays the resulting plot. In this example, p_1 and p_2 correspond to the probability of still being on the job after six month for the non-trained employees and the employees that received the training, respectively. The bottom panel of Figure 11 indicates that the posterior median for p_1 is 0.498, with 95% credible ranging from 0.455 to 0.542, and the posterior median for p_2 is 0.537, with 95% credible interval ranging from 0.494 to 0.581.

In sum, based on a default prior analysis, this fictitious data set offers anecdotal evidence in favor of the null hypothesis which states that the training is not effective over the hypothesis

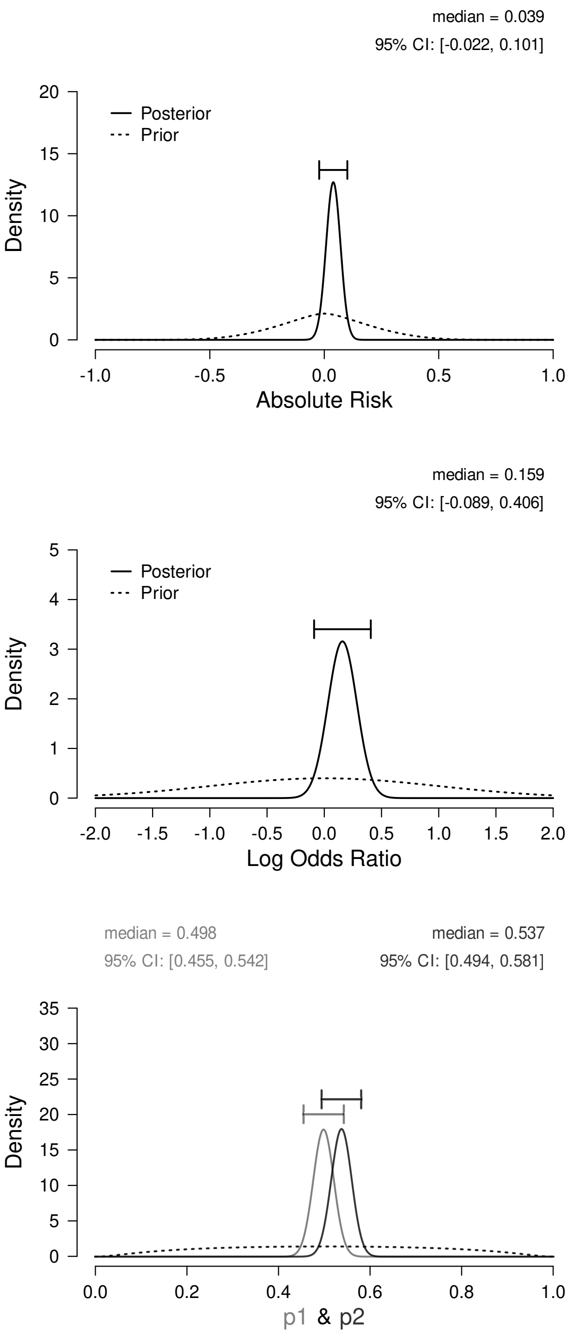

Figure 11: (Implied) prior and posterior distributions under $\mathcal{H}_1$. The dotted lines display the prior distributions, the solid lines display the posterior distributions (with 95% central credible intervals). The medians and the bounds of the 95% central credible intervals are displayed on top of each panel. The top panel displays the posterior distribution for the absolute risk (i.e., $p_2 - p_1$); the middle panel shows the posterior distribution for the log odds ratio parameter ψ; the bottom panel displays the marginal posterior distributions for the success probabilities p_1 and p_2.

that the training is effective; the consultancy firm should probably continue to collect data in order to obtain more compelling evidence before deciding whether or not the training should be implemented. If the true effect is as small as 4%, continued testing will ultimately show compelling evidence for $\mathcal{H}_+$ over $\mathcal{H}_0$. Note that continued testing is trivial in the Bayesian framework: The results can simply be updated as new observations arrive.

Affiliation:

Quentin F. Gronau
Department of Psychological Methods
University of Amsterdam
Nieuwe Achtergracht 129 B
1018 WT Amsterdam, The Netherlands
E-mail: Quentin.F.Gronau@gmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2021, Volume 100, Issue 17 *Submitted:* 2019-05-06
doi:10.18637/jss.v100.i17 *Accepted:* 2020-11-19