Closed-Form Power and Sample Size Calculations for Bayes Factors

Samuel Pawel **Leonhard Held**

Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich
E-mail: samuel.pawel@uzh.ch

Working paper version May 14, 2024

Abstract

An important aspect of study design is the determination of an appropriate sample size. The choice of sample size should be consistent with the planned analysis. If the planned analysis involves Bayes factor hypothesis testing, the sample size is usually desired to ensure a sufficiently high probability of obtaining compelling evidence for a hypothesis, given that the hypothesis is true. In practice, sample size determination based on Bayes factor analyses are typically performed using computationally intensive Monte Carlo simulation methods. Here we summarize how, under approximate normality assumptions, sample sizes can be determined numerically without simulation, and provide the R package `bfpwr` for doing so. We also identify conditions under which sample sizes can be determined in closed-form, leading to novel formulas that are easy-to-use, help fostering intuition, and enable asymptotic analysis. Case studies from medicine and psychology illustrate how researchers can use our methods for planning informative but cost-efficient studies.

Keywords: Bayesian hypothesis testing, evidence, likelihood ratio, numerical methods, study design

1 Introduction

A key aspect of study design is determining an appropriate sample size. Choosing a sample size that is too small may lead to inconclusive study results, while choosing a sample size that is too large may be unethical (e.g., for animal studies) or inefficient because the samples could be of better use in other studies. Whether or not a certain sample size can ensure sufficiently conclusive results depends on the planned analysis. Therefore, the sample size calculation should be aligned with the planned analysis (Anderson and Kelley, 2022), or in other words: '*As ye shall analyse is as ye shall design*' (Julious, 2023, p. 179).

A widely used formula for the sample size per group for data with continuous outcome based on a frequentist hypothesis test of a mean difference for two equally sized groups is given by

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu - \theta_0)^2}, \tag{1}$$

where z_q denotes the $q \times 100\%$ quantile of the standard normal distribution, α is the level of the test, $1 - \beta$ is the desired power, μ is the assumed mean difference under the alternative, θ_0 is the mean difference under the null hypothesis, and σ^2 is the variance of one observation (Matthews, 2006, p. 34). There exist various refinements of (1) that adapt it to unequal randomization, special study designs (e.g., cross-over studies), or other data types (e.g., binary data), see, for example, Kieser (2020) or Julious (2023). For many analysis methods, however, no closed-form formula exist and iterative or simulation methods have to be used. Nevertheless, while typically only being an approximation, the formula (1) allows for quick calculations and is often accurate enough for practical purposes. It also helps fostering intuition and is therefore useful, for example, in teaching of statistics or the study of asymptotics.

An alternative to frequentist hypothesis testing is Bayesian hypothesis testing. There are different flavors of Bayesian hypothesis testing, one of the most popular being the approach centered around the *Bayes factor*, which is the data-based updating factor of the prior to posterior odds of two competing hypotheses. Bayes factor approaches were pioneered by Jeffreys (1939) and are now in use in various scientific domains such as medicine (Goodman, 1999), psychology (Wagenmakers, 2007), or physics (Trotta, 2008). Bayes factor tests are conceptually different from frequentist tests in several ways. For example, they allow us to quantify evidence in favor of a null hypothesis or they can incorporate external information via a prior distribution. For an overview of Bayes factors see e.g., Kass and Raftery (1995); Held and Ott (2018).

Also if Bayes factors are used in the analysis, the design of the study should match the analysis. Fortunately, there is methodology for design based on Bayes factors (Weiss, 1997; Gelfand and Wang, 2002; De Santis, 2004, 2007; Schönbrodt and Wagenmakers, 2017; Schönbrodt et al., 2017; Pawel et al., 2023; Stefan et al., 2024). However, to our knowledge, there are no simple formulae such as (1) for sample size determination based on Bayes factor analyses. In practice, sample size determination is often performed by Monte Carlo simulation (Gelfand and Wang, 2002; Schönbrodt and Wagenmakers, 2017; Stefan et al., 2024), but this can be inaccurate, time-consuming, and less intuitive than a formula.

The goal of this paper is therefore to investigate whether, under approximate normality assumptions similar to those underlying the formula (1), a sample size formula can be derived for a planned Bayes factor analysis. As we will show, the answer is affirmative under certain assumptions about the analysis prior (the prior distribution for the parameter under the alternative used in the analysis) and the design prior (the prior distribution for the parameter used in the design). Specifically, for Bayes factors with point analysis priors (i.e., likelihood ratios) and point or normal design priors, and for Bayes factors with local normal analysis and design priors (i.e., normal priors centered around the null value) closed-form sample sizes are available. We also summarize sample size determination for Bayes factors analysis based on numerical root-finding (Weiss, 1997) that, while not being available in closed-form, is much faster and deterministic compared to simulation-based approaches. To facilitate reuse of our results, all methods are made available through the R package `bfpwr`.

This paper is organized as follows: We begin by defining the type of Bayes factor underlying our sample size calculations (Section 2), followed by deriving its distribution when new data are generated under different design priors (Section 3). Combining these results, we derive several formulae for the sample size under different constellations of design and analysis priors (Section 4). Exam-

ples from medicine and the social sciences then illustrates how our formulae can be implemented by researchers in practice (Section 5). The paper ends with a closing discussion of our results and final remarks on limitations and extensions (Section 6). Our R package `bfpwr` that implements the developed methodology, is illustrated in Appendix A.

2 Assumed Bayes factor analysis

To derive a formula for sample size calculation, we must first clarify how the future data will be analyzed. Denote by $\hat{\theta}$ the estimate of an unknown parameter θ that will result from the statistical analysis of n effective future samples, each with unit variance $\sigma^2_{\hat{\theta}}$. Assume that the estimate is approximately normally distributed $\hat{\theta} \mid \theta \sim \mathrm{N}(\theta, \sigma^2_{\hat{\theta}}/n)$. Table 1 shows common types of parameter estimates and the resulting interpretation of n and the unit variance $\sigma^2_{\hat{\theta}}$. For example, the frequentist sample size (1) assumes continuous outcome data and a mean difference parameter. The formula could be generalized to other settings by replacing $2\sigma^2$ in the numerator with other unit variances from Table 1. It is important to note that these are only approximations and they may be inadequate in certain situations, such as small sample sizes or rare events (Spiegelhalter et al., 2004).

Table 1: Different types of parameter estimates $\hat{\theta}$ with approximate variance $\mathrm{Var}(\hat{\theta}) = \sigma^2_{\hat{\theta}}/n$ and corresponding interpretation of sample size n and unit variance $\sigma^2_{\hat{\theta}}$ (adapted from Chapter 2.4 in Spiegelhalter et al., 2004 and Chapter 1 in Grieve, 2022). The variance of one continuous outcome sample is denoted by σ^2 and parameter estimates based on two groups assume equal number of samples per group.

Outcome	Parameter estimate $\hat{\theta}$	Interpretation of n	Unit variance $\sigma^2_{\hat{\theta}}$
Continuous	Mean	Number of samples	σ^2
Continuous	Mean difference	Number of samples per group	$2\sigma^2$
Continuous	Standardized mean difference	Number of samples per group	2
Continuous	z-transformed correlation	Number of samples minus 3	1
Binary	Log odds ratio	Total number of events	4
Binary	Arcsine difference	Number of samples per group	1/2
Survival	Log hazard ratio	Total number of events	4
Count	Log rate ratio	Total count	4

Assume now a point null hypothesis that states that θ equals a certain null value $H_0 \colon \theta = \theta_0$ and the alternative hypothesis that θ does not equal the null value $H_1 \colon \theta \neq \theta_0$, with prior $\theta \mid H_1 \sim \mathrm{N}(\mu, \tau^2)$ assigned to θ under H_1. The mean of the prior μ determines the most plausible parameter value under the alternative while the variance τ^2 determines its uncertainty. Informally, a point alternative at μ can be obtained by setting the variance of the prior to zero. The Bayes factor is then given by the updating factor of the prior to posterior odds of H_0 versus H_1, i.e.,

$$\mathrm{BF}_{01} = \frac{\Pr(H_0 \mid \hat{\theta})}{\Pr(H_1 \mid \hat{\theta})} \Big/ \frac{\Pr(H_0)}{\Pr(H_1)} = \sqrt{1 + \frac{n\tau^2}{\sigma^2_{\hat{\theta}}}} \exp\left[-\frac{1}{2} \left\{ \frac{(\hat{\theta} - \theta_0)^2}{\sigma^2_{\hat{\theta}}/n} - \frac{(\hat{\theta} - \mu)^2}{\tau^2 + \sigma^2_{\hat{\theta}}/n} \right\} \right]. \tag{2}$$

A Bayes factor smaller than one ($\mathrm{BF}_{01} < 1$) indicates evidence for the alternative H_1, while a Bayes

factor greater than one ($\text{BF}_{01} > 1$) indicates evidence for the null hypothesis H_0. The larger the deviation of the Bayes factor from one, the stronger the evidence.

The Bayes factor (2) has already appeared in the literature in one form or another (e.g., in Weiss, 1997; De Santis, 2004; Spiegelhalter et al., 2004; Bartoš and Wagenmakers, 2023), with perhaps the first proposal of a Bayes factor based on an approximately normally distributed parameter estimate and its standard error dating back to Jeffreys (1936), see also Wagenmakers (2022) for some historical notes on Jeffreys' approach. The Bayes factor (2) may be thought of as a 'Bayesian z-test', that is, a test of a normal mean based on an asymptotically normal statistic assuming that the variance of the statistic is known. Of course, the latter assumption is not true in most applications, but it makes the test widely applicable and is, for practical purposes, often close enough to Bayes factors based on the exact distribution of the data, which may or may not be available. Finally, the Bayes factor (2) can also be used with parameter estimates where a standard error is available but not of the form $\sigma_{\hat{\theta}}/\sqrt{n}$, e.g, a parameter estimate from a generalized linear model where the estimate is adjusted for covariates and the standard error is obtained numerically (e.g., with Fisher scoring). In this case, $\sigma_{\hat{\theta}}/\sqrt{n}$ in (2) can be replaced by the observed standard error. However, as we will show now, assuming such a particular dependence on the sample size n allows us to perform closed-form power and sample size calculations under certain additional assumptions.

3 Distribution and power function of the Bayes factor

Suppose now that we are interested in finding compelling evidence in favor of the alternative H_1 over the null hypothesis H_0 with a Bayes factor (2) smaller than some threshold $k < 1$. For example, the evidence thresholds could be $k = 1/3$ or $k = 1/10$, the levels from Jeffreys (1939) for 'substantial' and 'strong' evidence, respectively. To determine a sample size that ensures compelling evidence with a desired probability we need to know the distribution of the Bayes factor (2) for a given sample size.

Assume a so-called 'design prior' for the parameter θ that is used for the design of the study (O'Hagan et al., 2001, 2005). This prior should represent the state of knowledge and uncertainty about θ at the design stage and does not necessarily have to correspond to the 'analysis prior' $\theta \mid H_1 \sim \text{N}(\mu, \tau^2)$ used for the Bayes factor (2). In fact, the analysis prior is often set to a certain 'default' or 'objective' prior that is conventionally used in the field. Here, we will focus on normal design priors $\theta \sim \text{N}(\mu_d, \tau_d^2)$, as they are flexible enough to specify varying degrees of uncertainty about the parameter, and at the same time mathematically convenient for obtaining closed-form solutions for power and, in some cases, sample size.

Such a normal design prior induces a predictive distribution $\hat{\theta} \mid n, \mu_d, \tau_d \sim \text{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$ for the future parameter estimate $\hat{\theta}$. Under this normal sampling distribution of the future estimate, the distribution of the Bayes factor with normal alternative (2) can be derived in closed-form (Weiss, 1997; De Santis, 2004). We now rederive and extend this result in our setting and notation. The two cases of the Bayes factor with $\tau^2 = 0$ (point analysis prior under the alternative) and $\tau^2 > 0$ (normal analysis prior under the alternative) need to be distinguished, as the resulting distributions take a different form, and only the latter has been considered previously in the Bayes factor literature.

For the Bayes factor with point alternative ($\tau^2 = 0$), the cumulative distribution or 'power function' is

$$\Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau^2 = 0) = \begin{cases} 1 - \Phi(Z) & \text{if } \mu - \theta_0 > 0 \\ \Phi(Z) & \text{if } \mu - \theta_0 < 0 \end{cases} \tag{3}$$

with $\Phi(\cdot)$ the standard normal cumulative distribution function and

$$Z = \frac{1}{\sqrt{\tau_d^2 + \sigma_{\hat{\theta}}^2/n}} \left\{ \frac{\sigma_{\hat{\theta}}^2 \log k}{n(\theta_0 - \mu)} + \frac{\theta_0 + \mu}{2} - \mu_d \right\},$$

see Appendix B for details. In standard frequentist sample size determination, the power can typically be increased arbitrarily close to one by increasing the sample size. However, with the Bayes factor based on a point alternative, depending on the assumed design prior, one may not be able to approach a power of one with the power function (3) by increasing the sample size n. That is, the limiting power value is given by

$$\lim_{n \to \infty} \Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau^2 = 0) = \begin{cases} 1 - \Phi(Z_{\text{lim}}) & \text{if } \mu - \theta_0 > 0 \\ \Phi(Z_{\text{lim}}) & \text{if } \mu - \theta_0 < 0 \end{cases} \tag{4}$$

with $Z_{\text{lim}} = (\theta_0 + \mu - 2\mu_d)/(2\tau_d)$. Clearly, when the design prior is a point prior ($\tau_d^2 = 0$), Z_{lim} diverges and the limitting power approaches one or zero, depending on whether the location of the design prior μ_d is closer to the alternative μ or to the null θ_0. In case it is just in between the two, the limiting power approaches a half. On the other hand, for a normal design prior ($\tau_d^2 > 0$), the limiting power is bounded by a value in between (and not including) zero and one given by (4). The intuition behind these results is that for a normal design prior, there is always parameter uncertainty, even if the sample size becomes arbitrarily large, while for point design priors, the parameter uncertainty can be arbitrarily reduced by increasing the sample size. This parallels similar results on bounds for Bayesian/frequentist hybrid power of significance tests (Spiegelhalter et al., 2004; Micheloud and Held, 2022; Grieve, 2022).

For the Bayes factor with normal prior under the alternative ($\tau^2 > 0$), the cumulative distribution or power function is given by

$$\Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau^2 > 0) = \Phi(-\sqrt{X} - M) + \Phi(-\sqrt{X} + M) \tag{5}$$

with

$$M = \left\{ \mu_d - \theta_0 - \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}(\theta_0 - \mu) \right\} \frac{1}{\sqrt{\tau_d^2 + \sigma_{\hat{\theta}}^2/n}}$$

and

$$X = \left\{ \log\left(1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}\right) + \frac{(\theta_0 - \mu)^2}{\tau^2} - \log k^2 \right\} \left(1 + \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}\right) \frac{\sigma_{\hat{\theta}}^2}{n\tau_d^2 + \sigma_{\hat{\theta}}^2},$$

see Appendix B for details. Unlike the power function based on the Bayes factor with point alternative (3), the power function based on the Bayes factor with normal alternative (5) can be increased arbitrarily close to one by increasing the sample size n (see Appendix C). That is, we have that

$$\lim_{n \to \infty} \Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau^2 > 0) = 1$$

regardless of whether the design prior is a point prior ($\tau_d^2 = 0$) or a normal prior ($\tau_d^2 > 0$), provided that the design prior is not equal to the point null hypothesis itself. This is expected because Bayes factors testing point nulls against composite alternatives are 'consistent' in the sense that as the sample size increases, the probability of the Bayes factor favoring the hypothesis under which the data were generated tends to one (Dawid, 2011; Bayarri et al., 2012; Ly and Wagenmakers, 2022).

4 Sample size determination

Both power functions (3) and (5) are straightforward to implement and can be used to obtain 'power curves' as a function of the sample size, or of other parameters. We provide R implementations of both in our package `bfpwr` (see Appendix A for an illustration). Iterative root finding can then be applied to determine the sample size such that compelling evidence is obtained with a desired target power under a specified design prior (Weiss, 1997). This is also implemented in our R package.

We will now investigate situations where the sample size can be obtained in closed-form. As for the distribution of the Bayes factor in the previous section, there is again a distinction between sample size determination for Bayes factors with point alternatives ($\tau^2 = 0$) and normal alternatives ($\tau^2 > 0$), we start again with the former.

4.1 Bayes factor with point alternative

Assuming that the alternative μ is larger than the null θ_0 and setting the power function (3) equal to a target power of $1 - \beta$ lower than the limitting power (4), we obtain a quadratic equation in the sample size n

$$n^2 \underbrace{\left\{ \left(\frac{\theta_0 + \mu}{2} - \mu_d \right)^2 - z_{1-\beta}^2 \tau_d^2 \right\}}_{=a} + n \underbrace{\sigma_{\hat{\theta}}^2 \left\{ \frac{(\theta_0 + \mu - 2\mu_d) \log k}{\theta_0 - \mu} - z_{1-\beta}^2 \right\}}_{=b} + \underbrace{\left(\frac{\sigma_{\hat{\theta}}^2 \log k}{\theta_0 - \mu} \right)^2}_{=c} = 0.$$

Its solution is given by

$$n = \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \tag{6}$$

Of note, the second solution of the quadratic equation gives the sample size that leads to a power of β instead of $1 - \beta$. The sample size (6) also typically needs to be rounded to the next larger integer in order to be an evaluable sample size in practice.

We will now investigate the sample size (6) for two special cases. First, suppose that the design

prior is a point prior ($\tau_d^2 = 0$) at μ_d, not necessarily the same as the alternative μ. This leads to

$$n = \frac{\left[z_{1-\beta} + \sqrt{z_{1-\beta}^2 - \log k^2\{(\theta_0 + \mu - 2\mu_d)/(\theta_0 - \mu)\}}\right]^2}{(\mu + \theta_0 - 2\mu_d)^2/\sigma_{\hat{\theta}}^2}. \tag{7}$$

Assuming now that the tested parameter is a mean difference with unit variance $\sigma_{\hat{\theta}}^2 = 2\sigma^2$ (see Table 1, second row), we can see that the formula (7) represents a modification of the frequentist formula (1): The 'standardized effect size' $(\mu - \theta_0)/\sigma$ in the denominator is replaced by a standardized effect size that takes into account the mean of the design prior $(\mu + \theta_0 - 2\mu_d)/\sigma$, but reduces to the standardized effect size when the assumed parameter in the design equals the assumed parameter in the analysis (when $\mu_d = \mu$). Moreover, the quantile $z_{1-\alpha/2}$ is replaced by $\sqrt{[z_{1-\beta}^2 - \log k^2\{(\theta_0 + \mu - 2m)/(\theta_0 - \mu)\}]}$, reflecting the fact that we are interested in a Bayes factor hypothesis test and not a frequentist significance test.

Second, assume that the design prior is also equal to the alternative ($\mu_d = \mu$), so that the formula (7) further reduces to

$$n = \frac{\left\{z_{1-\beta} + \sqrt{z_{1-\beta}^2 - \log k^2}\right\}^2}{(\mu - \theta_0)^2/\sigma_{\hat{\theta}}^2}. \tag{8}$$

The same formula (8) was also found by Strug et al. (2007) for 'evidential' sample size calculations, but is unfortunately not well known. Unsurprisingly the two formulae coincide as Bayes factors and likelihood ratios – the measure of evidence used in 'evidential statistics' (see e.g., Royall, 1997; Blume, 2002; Strug, 2018; Perneger, 2021) – coincide when the Bayes factor involves only point hypotheses. Our more general formula (6) thus also generalizes evidential sample size calculations to incorporate parameter uncertainty, similar to how Bayesian design priors can be used for incorporating parameter uncertainty in a planned frequentist analysis (see e.g., Grieve, 2022, for an overview of Bayesian/frequentist hybrid design approaches).

To illustrate formula (8), we now assume that $\hat{\theta}$ is a standardized mean difference with unit variance $\sigma_{\hat{\theta}}^2 = 2$ so that n can be interpreted as the number of samples per group (see the third row in Table 1). Table 2 shows the sample size (8) based on an assumed standardized mean difference of one ($\mu - \theta_0 = 1$). We see, for instance, that $n = 20$ samples per group are required to achieve $1 - \beta = 80\%$ power for a Bayes factor threshold $k = 1/10$. If the assumed standardized mean difference were smaller, the required sample size would become larger. For example, for a half as large standardized mean difference, the sample sizes from Table 2 quadruple, e.g., requiring $n = 80$ samples per group to have $1 - \beta = 80\%$ power for a Bayes factor threshold of $k = 1/10$.

4.2 Bayes factor with normal alternative

Finding a sample size formula becomes more difficult when we move from point to normal alternatives. The technical reason is that when we set the power function (5) equal to a target power and try to solve for the sample size n, we have n appearing both in and outside logarithms. This forms a transcendental equation that cannot be solved in terms of elementary functions. In general, there is no closed-form solution. However, as we will show now, solutions can, under certain conditions, be

Table 2: Sample size per group n to obtain a Bayes factor BF_{01} smaller than k with at least a power of $1 - \beta$ assuming a standardized mean difference of one under the alternative and for the design.

	k											
$1 - \beta$	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10	1/30	1/100	1/300	1/1000
50%	5	6	7	8	8	9	9	10	14	19	23	28
55%	6	7	8	9	9	10	10	11	15	21	25	30
60%	7	8	9	10	11	11	12	12	17	22	27	32
65%	8	9	10	11	12	13	13	14	19	24	29	34
70%	9	11	12	13	14	14	15	15	21	26	32	37
75%	11	13	14	15	16	16	17	18	23	29	34	40
80%	13	15	16	17	18	19	20	20	26	32	38	44
85%	17	18	20	21	22	23	23	24	30	37	42	48
90%	22	23	25	26	27	28	28	29	36	42	48	55
95%	30	32	34	35	36	37	38	38	45	52	59	66

expressed in terms of the Lambert W function (Corless et al., 1996). The Lambert W function is the function $\text{W}(\cdot)$ that satisfies $\text{W}(x)\exp\{\text{W}(x)\} = x$, and it is therefore sometimes also called 'product logarithm'. It has many fundamental applications, such as the solution of the Schrödinger equation in quantum-mechanics, and has also previously appeared in the context of Bayes factor hypothesis testing (Pawel and Held, 2022; Wagenmakers, 2022; Pawel et al., 2024).

Suppose now that the design prior and the alternative are both centered around the null value ($\mu_d = \mu = \theta_0$). Centering the prior around the null value is commonly done in 'default' Bayes factor tests (Berger and Delampady, 1987). It encodes the assumption that some parameters are larger while others are smaller than the null, the standard deviation of the distribution determining the variability, yet the average parameter equals the null value. We then have that $M = 0$ and hence the power (5) reduces to

$$\Pr\{\text{BF}_{01} \leq k \mid n, \tau_d^2, \mu_d = \mu = \theta_0\} = 2\Phi(-\sqrt{X}). \tag{9}$$

Further, assume that the variance of the design prior corresponds to the variance of the normal prior under the alternative ($\tau_d^2 = \tau^2$). We then have that

$$X = \left\{\log\left(1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}\right) - \log k^2\right\}\frac{\sigma_{\hat{\theta}}^2}{n\tau^2}.$$

Setting the power function (9) equal to a target power of $1 - \beta$ and assuming that $\log\{1 + (n\tau^2)/\sigma_{\hat{\theta}}^2\} \approx \log\{(n\tau^2)/\sigma_{\hat{\theta}}^2\}$, we obtain the following approximate sample size formula

$$n = \frac{\sigma_{\hat{\theta}}^2}{\tau^2}\underbrace{k^2 \exp\left\{-\text{W}_{-1}(-k^2 z_{(1-\beta)/2}^2)\right\}}_{=n_{k,\beta}} \tag{10}$$

with $\text{W}_{-1}(\cdot)$ the branch of the Lambert W function that satisfies $\text{W}(x) < -1$ for $y \in (-1/e, 0)$ (Corless et al., 1996), see Appendix D for details.

We can see that the sample size (10) depends on the ratio of the unit variance $\sigma^2_{\hat{\theta}}$ to the prior variance τ^2 multiplied by a 'unit information sample size' $n_{k,\beta}$ which depends only on the Bayes factor threshold k and the target power $1-\beta$. The unit information sample size is the sample size that is obtained when a unit information prior (Kass and Wasserman, 1995) is specified, which is a prior with variance equal to the unit variance ($\tau^2 = \sigma^2_{\hat{\theta}}$). As in classical sample size calculations, smaller unit variances $\sigma^2_{\hat{\theta}}$ reduce the sample size. The prior variance τ^2 determines how large parameters are expected under the alternative, and as such, larger prior variances lead to a reduction of sample size similar to the effect size in classical sample size determination. Finally, the formula (10) allows us to study the potential existence of sample size that can achieve the target power: Since the argument of the Lambert W function has to be at least $-1/e$ for it to be defined, we can infer that only combinations of Bayes factor thresholds k and power values $1-\beta$ that satisfy $k^2 z^2_{(1-\beta)/2} \leq 1/e$ can actually be achievable with a finite sample size. For example, it is impossible to find a sample size that guarantees a power of $1-\beta = 50\%$ for a threshold of $k = 1$ since then $z_{0.25} = -0.67 < 1/\sqrt{e} = 0.607$.

Table 3: Required unit information sample size $n_{k,\beta}$ to obtain a Bayes factor smaller than k with at least a power of $1-\beta$ with a unit information analysis and design prior.

	k											
$1-\beta$	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10	1/30	1/100	1/300	1/1000
50%	10	12	13	14	15	16	16	17	22	28	33	39
55%	14	16	17	19	20	21	21	22	29	36	43	50
60%	19	22	24	25	27	28	29	29	38	48	57	66
65%	27	30	33	35	37	38	40	41	53	66	77	89
70%	40	45	48	51	53	56	57	59	75	93	109	126
75%	63	70	75	79	82	85	88	90	114	140	163	188
80%	108	118	126	132	138	143	147	150	188	229	265	305
85%	212	230	244	256	265	274	281	287	355	427	493	564
90%	538	579	610	636	658	677	693	708	859	1023	1170	1331
95%	2554	2716	2841	2943	3029	3103	3168	3226	3829	4481	5071	5714

Table 3 shows unit information sample sizes $n_{k,\beta}$ for a range of powers $1-\beta$ and Bayes factor thresholds k. Compared to the sample sizes from Table 2 the sample sizes are quite a bit larger. This is because the design and analysis priors underlying each of these calculations encode vastly different assumptions: The normal alternative represents a parameter distribution that is centered around the null value while the point alternative from Table 2 represents a mean difference of one standard deviation away from the null. The former is a more pessimistic assumption than the latter. To incorporate more optimistic beliefs into the calculations we may increase the standard deviation of the distribution τ as this encodes the assumption of potentially larger parameters. For example, doubling the standard deviation τ leads to a four-fold decrease of the sample size (10).

The formula (10) is, to our knowledge, the first closed-form sample size formula for Bayes factor analysis with normal alternatives. While interesting from a technical point of view, its practical use is perhaps more limited than the sample size formula for Bayes factors with point alternatives (6). This is because it makes the restrictive assumption that the design prior and the analysis prior are both centered around the null ($\mu_d = \mu = \theta_0$). This seems unrealistic in practice, since researchers

designing a study usually have good reasons to expect parameters to be different from the null and would like to account for this in the sample size calculation. However, we have not been able to derive a sample size formula for this more general setting due to the transcendental nature of the power equation, even in terms of the Lambert W function. Fortunately, the sample size can still be easily calculated numerically with our R package which is quicker and more reliable than computing it with a simulation approach.

5 Application

We will now illustrate Bayes factor sample size and power calculations using examples from medicine and psychology.

5.1 Randomized controlled clinical trial on treatment of influenza

MIST Study Group (1998) conducted a randomized controlled clinical trial to evaluate the efficacy of the drug 'zanamivir' in the treatment of influenza A and B. The primary endpoint of the trial was the time in days to relief of clinically important symptoms of influenza. This outcome was treated as continuous by the investigators, and the parameter of interest was the mean difference in time to relief between the treatment and placebo groups θ. The null hypothesis was defined as no difference in efficacy of the treatment ($H_0 \colon \theta = 0$), whereas the alternative was defined as a clinically relevant difference of one day ($H_1 \colon \theta = 1$). For sample size calculations, the investigators used a standard deviation of $\sigma = 2.75$ estimated from a previous study.

While this study was analyzed using frequentist methodology, we will now examine what the power and sample size calculations might look like if the planned analysis were performed using Bayes factors. For doing so, we assume a point alternative for the Bayes factor that equals the alternative specified by the trial investigators ($\mu = 1$ and $\tau = 0$), so that the Bayes factor corresponds with a likelihood ratio. Hence, we can use the power function (3) and sample size formula (6) to compute power and sample size under point and normal design priors, which we illustrate in the following.

Figure 1 shows power curves based on the Bayes factor providing strong evidence in favor of the alternative ($\text{BF}_{01} < 1/10$) in the top plot, and based on the Bayes factor providing strong evidence in favor of the null ($\text{BF}_{01} > 10$) in the bottom plot. The colors indicate under which data distribution the power was computed.

Focusing on the top plot, we can see from the green curve that at least $n = 217$ samples per group are required to ensure a target power of $1 - \beta = 90\%$ assuming that the same point prior is used in the design as for the analysis (i.e., a point prior at $\mu = 1$). This number increases to $n = 384$ when we move to a design prior that incorporates parameter uncertainty (blue curve), i.e., a prior that is still centered around $\mu = 1$ but with a standard deviation of $\tau_d = 0.25$. Finally, if we look at the orange power curve computed assuming that the null hypothesis is true ($\theta = 0$), we can see that the probability of misleading evidence for the alternative when the null hypothesis is actually true is very low and appears to be adequately controlled by conventional standards (i.e., below 5%) for each of the two sample sizes.

Focusing now on the bottom plot, we can see that the sample size $n = 217$ also ensures a power of

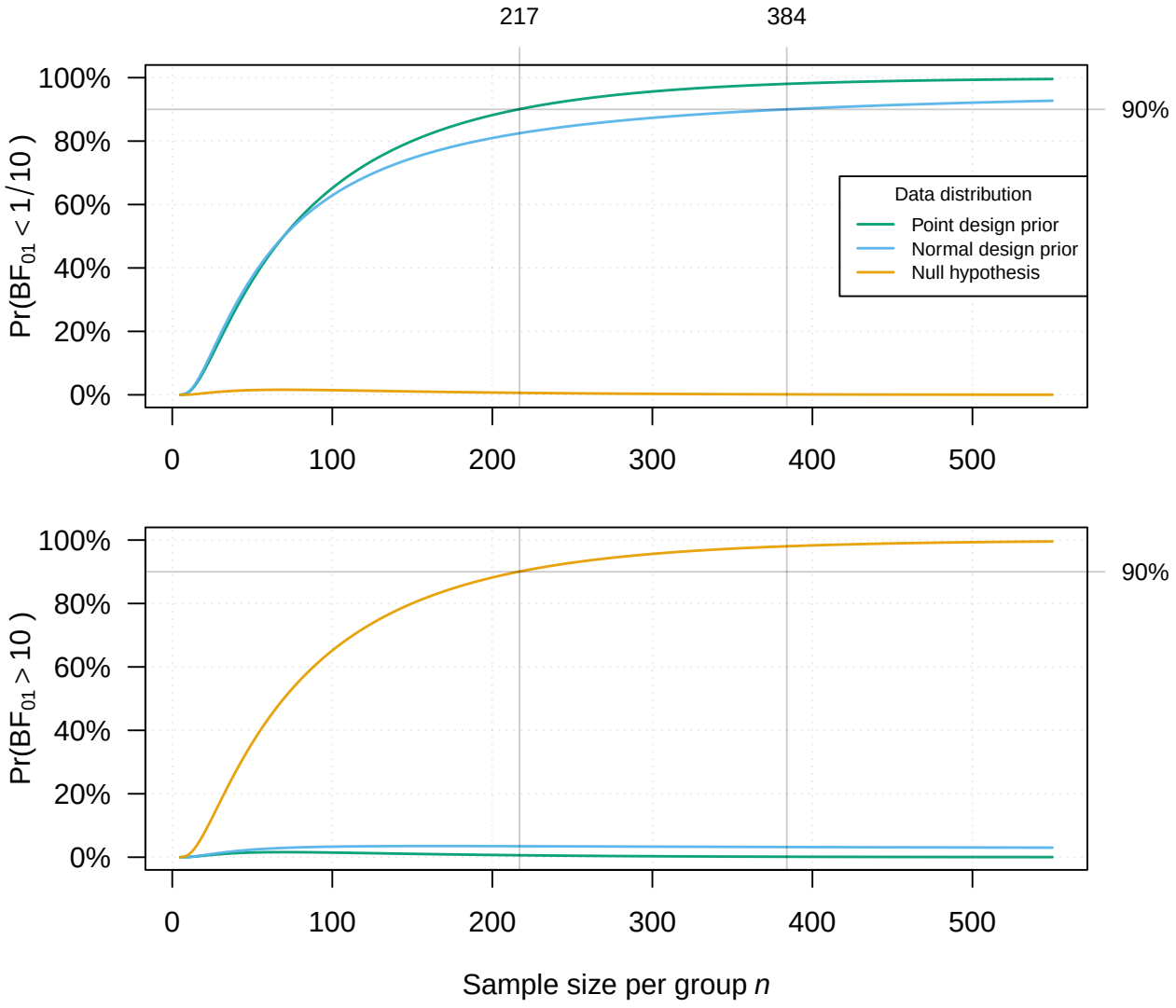


Figure 1: Power and sample size calculations for MIST trial (MIST Study Group, 1998). The Bayes factor assumed for the analysis contrasts $H_0 : \theta = 0$ to $H_1 : \theta = 1$. Curves are shown for data distributions induced by a point design prior at the alternative $\mu_d = \mu = 1$, a normal design prior centered at the alternative with standard deviation $\tau_d = 0.25$, and under the null hypothesis.

$1 - \beta = 90\%$ for finding evidence for the null hypothesis. This is due to the symmetric nature of the point versus point hypothesis Bayes factor considered in this example, as swapping the null θ_0 and alternative μ in formula (8) does not change the resulting sample size. Finally, looking at the green and blue curves we can see that the probability of misleading evidence in favor of the null when the data are generated from the design priors seems to be reasonably well controlled (below 5%) across the whole range of sample sizes considered.

5.2 Comparison to Monte Carlo simulation methods

Schönbrodt and Wagenmakers (2017) proposed a Monte Carlo simulation approach for power and sample size calculations for Bayes factor analyses, and they provide the R package `BFDA` (Schönbrodt

and Stefan, 2019) for this purpose. The idea is to simulate data sets under an assumed design prior and sample size, and then analyze each data set with a specified Bayes factor. This results in a distribution of Bayes factors from which the power can be calculated. The simulation is then repeated for other sample sizes until the desired power is achieved. This approach can be used in quite general settings, but it is also computationally intensive and comes with Monte Carlo error.

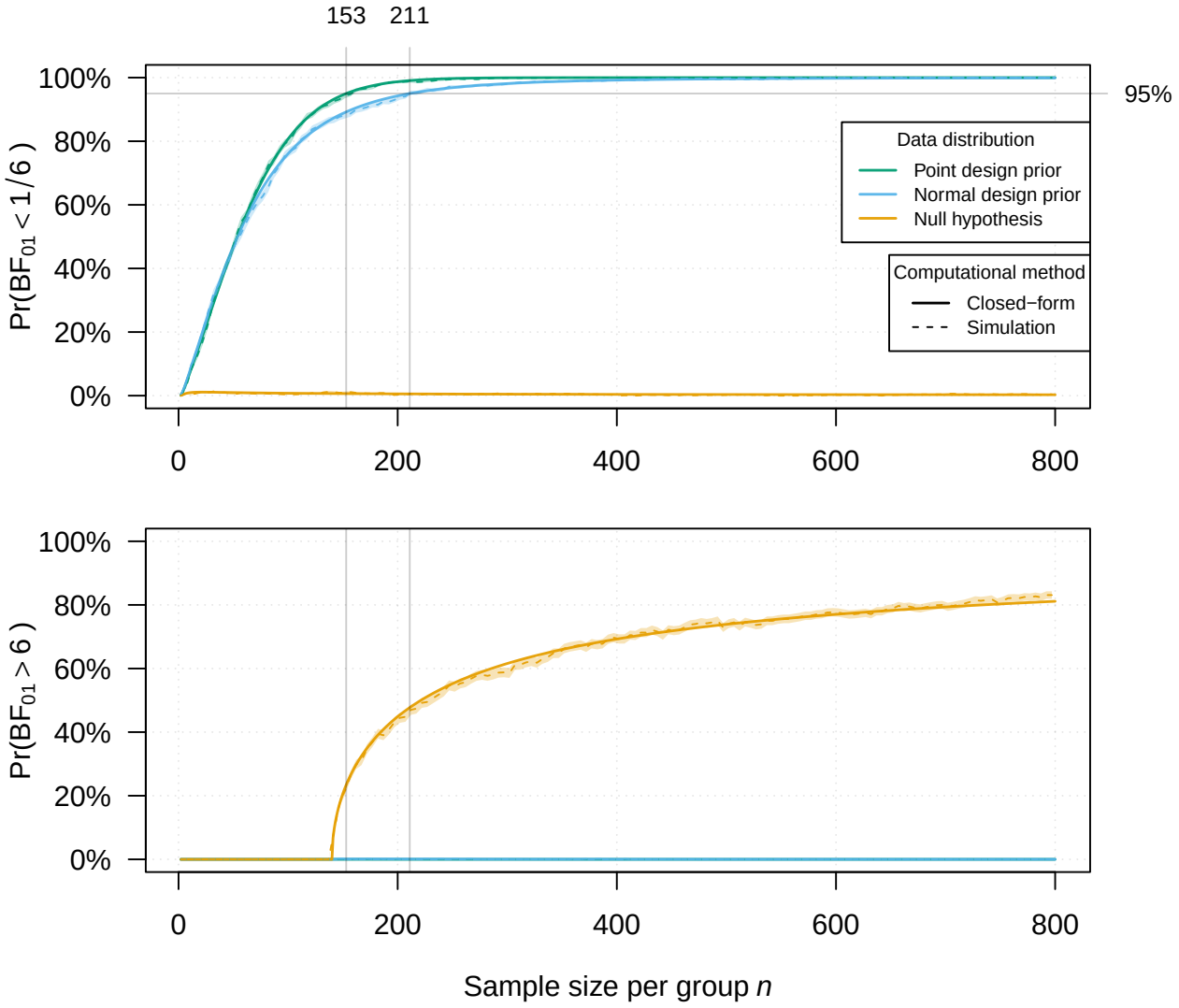


Figure 2: Power and sample size calculation for example adapted from Schönbrodt and Wagenmakers (2017, p. 133). The Bayes factor assumed for the analysis contrasts $H_0 : \theta = 0$ to $H_1 : \theta \neq 0$ with $\theta \mid H_1 \sim \text{N}(0, 1/2)$ prior assigned to the standardized mean difference θ under the alternative. The design prior is either set to a point prior at $\theta = 0.5$ or a normal prior at $\mu_d = 0.5$ with standard deviation $\tau_d = 0.1$. For comparison, 1000 Monte Carlo simulations are performed with the `BFDA` package (Schönbrodt and Stefan, 2019) to obtain a simulation-based power curve approximation (shown with one Monte Carlo standard error bands).

We will now look at an adaptation of an example examined in Schönbrodt and Wagenmakers (2017, p. 133). Suppose we want to test the null hypothesis that a standardized mean difference θ is zero ($H_0 : \theta = 0$) versus the alternative hypothesis that it is different from zero ($H_1 : \theta \neq 0$). We assume a $\theta \mid H_1 \sim \text{N}(0, 1/2)$ analysis prior for the standardized mean difference under the alternative,

similar to the Cauchy prior with scale $1/\sqrt{2}$ that was assumed by Schönbrodt and Wagenmakers (2017). As they did, we also investigate two design priors: either a point prior at $\theta = 0.5$, which is a convention for a 'medium' standardized mean difference (Cohen, 1992), or a normal prior at $\mu_d = 0.5$ with standard deviation $\tau_d = 0.1$ to incorporate parameter uncertainty. Figure 2 shows the resulting power curves computed from equation (5), along with the `BFDA` Monte Carlo simulation approximations (with Monte Carlo standard error bands) for comparison. The closed-form power curves were computed instantly while each simulated power curve took about 10 minutes to compute when parallelized across 10 cores on a modern laptop.

We see that closed-form and simulation curves closely align in all cases, though the former shows Monte Carlo error. Looking at the green curves in the top plot, we see that a sample size of $n = 153$ per group is sufficient to achieve a target power of 95% with a Bayes factor threshold of $k = 1/6$ under the point design prior. This sample size is slightly larger than the $n = 146$ that Schönbrodt and Wagenmakers (2017) obtained in their calculations, which instead assumed a Cauchy analysis prior. As in the previous example, incorporating parameter uncertainty via a normal design prior increases the required sample size, in this case to $n = 211$. Looking at the orange curve in the bottom plot, we see that these same sample sizes only have modest power of around 20% and 50%, respectively, for obtaining evidence in favor of the true null hypothesis (at a threshold of $k = 6$). To obtain a power of 80%, say, around $n = 800$ samples per group would be required. This illustrates the well-known fact that evidence accumulates slower under point null hypotheses compared to normal alternatives when the Bayes factor involves testing a point null against a normal alternative (Johnson and Rossell, 2010). Finally, looking at the orange curve in the top plot and the green/blue curves in the bottom plot, we see that the probability of misleading evidence in favor of the incorrect hypothesis appears to be adequately controlled, as these curves are virtually zero over the entire range of sample sizes.

6 Discussion

We presented methods for performing power and sample size calculations for settings where future data are analyzed with Bayes factor hypothesis tests. These methods rely on approximate normality of the data, which is a common assumption in the methodology of power and sample size calculations. We have synthesized and extended previous theoretical results on power functions for Bayes factors and implemented them in an easy-to-use R package `bfpwr`. We also derived novel sample size formulae that are-easy-to-use, help fostering intuition, and enable asymptotic analysis of power and sample size. Compared to commonly used simulation-based methods, our methods are less general. However, in the setting where they are applicable – which includes many common scenarios, such as testing mean differences – they are faster, deterministic, and require no simulation parameters to be specified. Therefore, we believe that the availability of such methods addresses an important practical need and can help researchers design efficient studies with minimal effort.

A clear limitation of our methodology is the asymptotic normality assumption. This assumption may be inappropriate for certain data or parameter types, and may lead to an underappreciation of uncertainty and consequently an underestimation of sample size. Simulation-based methods do not have this shortcoming, as they can be tailored to any data distribution and analysis method. Nevertheless, simulation methods may be intimidating or too advanced for research workers, in which

case we believe it is better to do an approximate calculation than no calculation at all. One avenue for future work might be to extend closed-form sample size calculations to more specific settings, such as, binary outcomes or continuous outcomes with unknown variances. Another limitation is the type of Bayes factor that we considered for the analysis, which is limited to univariate parameters with normal or point priors under the alternative. Our work could be extended to ANOVA or regression settings with multivariate parameters and/or to other prior distributions such as truncated normal distributions to incorporate directionality. Finally, we did not consider 'open-ended' sequential designs, where data are collected continuously until compelling evidence for one of the competing hypotheses is found (Wald, 1947). The sequential approach is particularly attractive for Bayes factor inference as design and analysis prior distributions can be updated based on the accumulating data. Researchers can then make informed decisions about whether or not it is worthwhile to continue collecting data or to stop (Stefan et al., 2024). For these purposes, it would be interesting to consider the Bayes factor indexed by the sample size as a stochastic process, and study its properties.

Acknowledgments

We thank František Bartoš for valuable comments on drafts of the manuscript. The acknowledgment of these individuals does not imply their endorsement of the paper.

Conflict of interest

We declare no conflict of interest.

Software and data

Code and data to reproduce our analyses are openly available at `https://github.com/SamCH93/bfpwr`. A snapshot of the repository at the time of writing is available at `https://doi.org/10.5281/zenodo.XXXXXX`. We used the statistical programming language R version 4.4.0 (2024-04-24) for analyses (R Core Team, 2024) along with the `lamW` (Adler, 2015), `xtable` (Dahl et al., 2019), and `knitr` (Xie, 2024) packages.

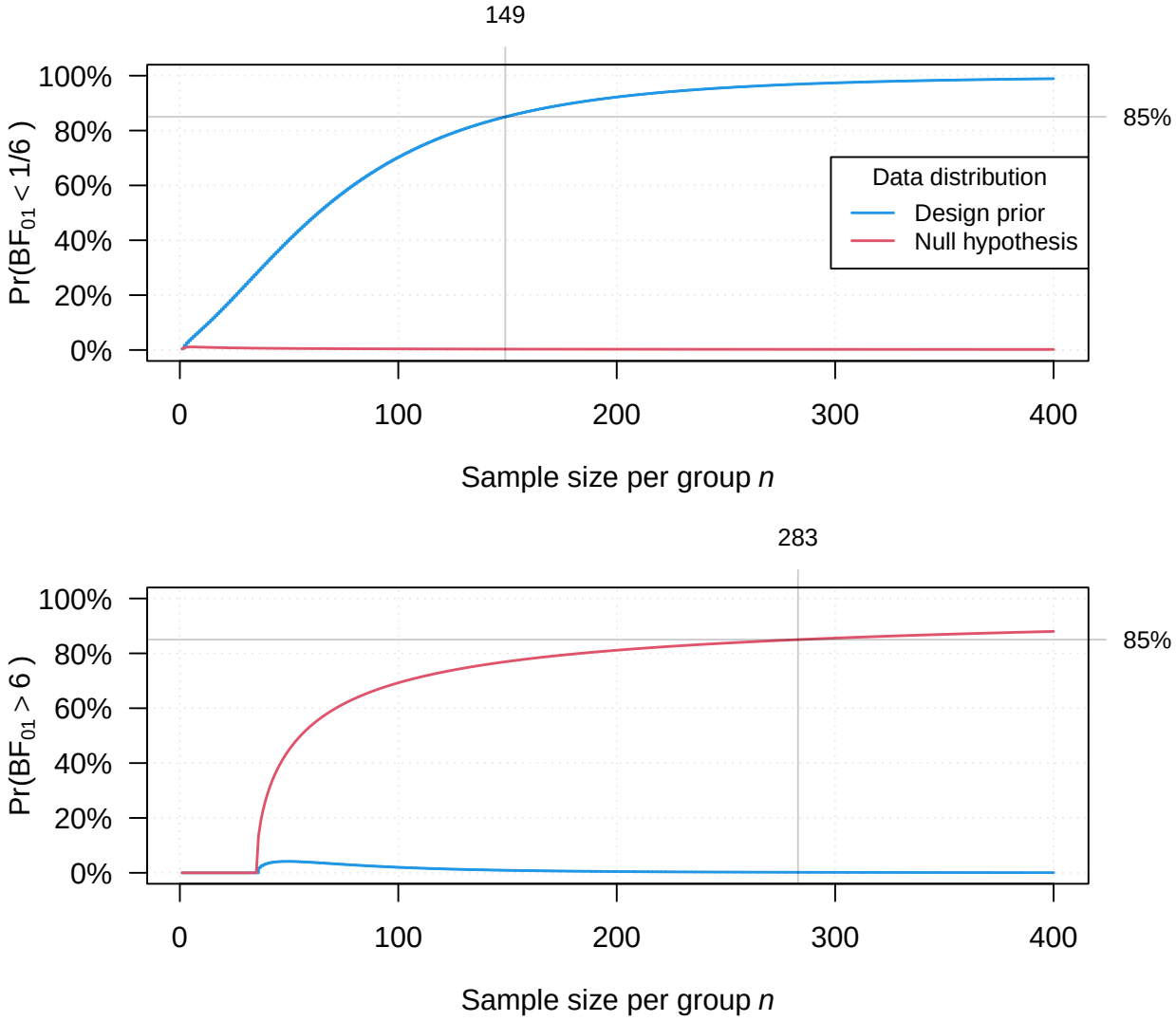
Appendix A The bfpwr R package

Our R package can be installed by running `install.packages("bfpwr")` in an R session, the development version can be installed from GitHub (`https://github.com/SamCH93/bfpwr`). The workhorse function of our R package is `powerbf01`. It is inspired by the `power.t.test` function from the `stats` package, with which many user will be familiar. As `power.t.test`, the function `powerbf01` assumes that the data are continuous and that the parameter of interest is either a mean or a mean difference. The functions `pbf01` and `nbf01` are more general and can be used for any approximately normally distributed parameter estimate with approximate variance $\text{Var}(\hat{\theta}) = \sigma^2_{\hat{\theta}}/n$, though users have to

specify the unit variance $\sigma^2_{\tilde{\theta}}$ themselves. The following code chunk illustrate how `powerbf01` can be used.

```
## install from CRAN or GitHub (the latter requires "remotes" package)
## install.packages("bfpwr") # not yet on CRAN
## remotes::install_github(repo = "SamCH93/bfpwr", subdir = "package")

## load package
library(bfpwr)

## BF parameters
k <- 1/6 # BF threshold
null <- 0 # null value
sd <- 1 # standard deviation of one sample
pm <- null # analysis prior centered around null value
psd <- sqrt(2) # unit information sd for a standardized mean difference
type <- "two.sample" # two-sample test

## design prior
dpm <- 0.5 # design prior mean equal to large SMD effect size
dpsd <- 0.1 # design prior sd to incorporate parameter uncertainty

## determine sample size to achieve 85% power
power <- 0.85
ssd <- powerbf01(k = k, power = power, sd = sd, null = null, pm = pm, psd = psd,
                 dpm = dpm, dpsd = dpsd, type = type)
ssd

##
##      Two-sample Bayes factor power calculation
##
##                          n = 148.5498
##                      power = 0.85
##                         sd = 1
##                       null = 0
##       analysis prior mean = 0
##         analysis prior sd = 1.414214
##         design prior mean = 0.5
##           design prior sd = 0.1
##            BF threshold k = 1/6
##
## NOTE: BF oriented in favor of H0 (BF < 1 indicates evidence for H1 over H0)
```

```
##       n is number of *samples per group*
##       sd is standard deviation of one sample (assumed equal in both groups)

## plot power curve
plot(ssd, nlim = c(1, 400))
```



Appendix B Distribution of the Bayes factor

The Bayes factor (2) with $\tau^2 = 0$ can be rewritten as

$$\text{BF}_{01} = \exp\left[\frac{n}{\sigma^2_{\hat{\theta}}}\left\{\hat{\theta}(\theta_0 - \mu) - \frac{\theta_0^2 - \mu^2}{2}\right\}\right]. \tag{11}$$

Suppose that compelling evidence for H_1 is achieved when $\text{BF}_{01} \leq k < 1$. In this case, $\text{BF}_{01} \leq k$ can be rewritten as

$$\hat{\theta}(\theta_0 - \mu) \leq \frac{\sigma_{\hat{\theta}}^2 \log k}{n} + \frac{\theta_0^2 - \mu^2}{2}.$$

Dividing by $(\theta_0 - \mu)$ changes the inequality if $\mu > \theta_0$. We then have that under a normal distribution $\hat{\theta} \mid n, \mu_d, \tau_d^2 \sim \text{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$, the probability of compelling evidence is given by (3).

The Bayes factor (2) with $\tau^2 > 0$ can be rewritten as

$$\text{BF}_{01} = \sqrt{1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}} \exp\left(-\frac{1}{2}\left[\frac{\{\hat{\theta} - \theta_0 - \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}(\theta_0 - \mu)\}^2}{\frac{\sigma_{\hat{\theta}}^2}{n}(1 + \frac{\sigma_{\hat{\theta}}^2}{n\tau^2})} - \frac{(\theta_0 - \mu)^2}{\tau^2}\right]\right). \tag{12}$$

Suppose that compelling evidence for H_1 is achieved when $\text{BF}_{01} \leq k$, which can be rearranged to

$$\left\{\hat{\theta} - \theta_0 - \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}(\theta_0 - \mu)\right\}^2 \geq \left\{\log\left(1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}\right) + \frac{(\theta_0 - \mu)^2}{\tau^2} - \log k^2\right\}\left(1 + \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}\right)\frac{\sigma_{\hat{\theta}}^2}{n}.$$

Therefore, under a normal distribution $\hat{\theta} \mid n, \mu_d, \tau_d^2 \sim \text{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$, the probability of compelling evidence is given by (5).

Appendix C Limiting power of Bayes factor with normal alternative

We have that

$$\lim_{n\to\infty} M = \frac{\mu_d - \theta_0}{\tau_d}$$

and

$$\lim_{n\to\infty} X = \lim_{n\to\infty}\left[\left\{\log\left(1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}\right) + \frac{(\theta_0 - \mu)^2}{\tau^2} - \log k^2\right\}\frac{\sigma_{\hat{\theta}}^2}{n\tau_d^2 + \sigma_{\hat{\theta}}^2}\right].$$

Thus, when $\tau_d^2 \downarrow 0$ and $\mu_d \neq \theta_0$, both M and X diverge but the M term diverges faster than the X term. When $\tau_d^2 > 0$, the M term approaches a constant while the X term approaches zero. Consequently, in both cases it holds that

$$\lim_{n\to\infty} \Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau^2 > 0) = \lim_{n\to\infty}\left\{\Phi(-\sqrt{X} - M) + \Phi(-\sqrt{X} + M)\right\} = 1.$$

Appendix D Sample size for Bayes factor with local normal prior

Equating the power function (9) to $1 - \beta$ and rearranging, we have that

$$\begin{aligned} z^2_{(1-\beta)/2} &= \left\{ \log\left(1 + \frac{n\tau^2}{\sigma^2_{\hat{\theta}}}\right) - \log k^2 \right\} \frac{\sigma^2_{\hat{\theta}}}{n\tau^2} \\ &\approx \left\{ \log\left(\frac{n\tau^2}{\sigma^2_{\hat{\theta}}}\right) - \log k^2 \right\} \frac{\sigma^2_{\hat{\theta}}}{n\tau^2} \\ &= \log\left(\frac{n\tau^2}{\sigma^2_{\hat{\theta}} k^2}\right) \frac{\sigma^2_{\hat{\theta}}}{n\tau^2} \end{aligned}$$

Multiplying by $-k^2$ and rewriting the second factor on the right-hand-side as exponential leads to

$$-k^2 z^2_{(1-\beta)/2} = -\log\left(\frac{n\tau^2}{\sigma^2_{\hat{\theta}} k^2}\right) \exp\left\{ -\log\left(\frac{n\tau^2}{\sigma^2_{\hat{\theta}} k^2}\right) \right\}.$$

Hence, we can apply the Lambert W function to obtain

$$-\log\left(\frac{n\tau^2}{\sigma^2_{\hat{\theta}} k^2}\right) = \mathrm{W}\left(-k^2 z^2_{(1-\beta)/2}\right)$$

from which we get the sample size

$$n = \frac{\sigma^2_{\hat{\theta}}}{\tau^2} k^2 \exp\left\{ -\mathrm{W}\left(-k^2 z^2_{(1-\beta)/2}\right) \right\}.$$

For arguments $y \in (-1/e, 0)$, the Lambert W function has two branches. The sample size is obtained from the branch commonly denoted as $\mathrm{W}_{-1}(\cdot)$ which satisfies $\mathrm{W}(x) < -1$ for $y \in (-1/e, 0)$ (Corless et al., 1996). This is because this branch always leads to larger samples sizes than the other and guarantees that unit information sample sizes are always larger than one.

References

Adler, A. (2015). *lamW: Lambert-W Function*. URL https://CRAN.R-project.org/package=lamW. R package version 2.2.3.

Anderson, S. F. and Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. doi:10.1037/met0000520. Advance online publication.

Bartoš, F. and Wagenmakers, E. (2023). A general approximation to nested Bayes factors with informed priors. *Stat*, 12(1). doi:10.1002/sta4.600.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. doi:10.1214/12-aos1013.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3):317–335. doi:10.1214/ss/1177013238.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, 21(17):2563–2599. doi:10.1002/sim.1216.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159. doi:10.1037/0033-2909.112.1.155.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359. doi:10.1007/bf02124750.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML*. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-4.

Dawid, P. A. (2011). Posterior model probabilities. In Bandyopadhyay, P. S. and Forster, M. R., editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 607–630. North-Holland, Amsterdam.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, 124(1):121–144. doi:10.1016/s0378-3758(03)00198-8.

De Santis, F. (2007). Alternative Bayes factors: Sample size determination and discriminatory power assessment. *TEST*, 16(3):504–522. doi:10.1007/s11749-006-0017-7.

Gelfand, A. E. and Wang, F. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2):193–208. doi:10.1214/ss/1030550861.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005. doi:10.7326/0003-4819-130-12-199906150-00019.

Grieve, A. P. (2022). *Hybrid frequentist/Bayesian power and Bayesian power in planning clinical trials*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, London.

Held, L. and Ott, M. (2018). On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1):393–419. doi:10.1146/annurev-statistics-031017-100307.

Jeffreys, H. (1936). Further significance tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(3):416–445. doi:10.1017/s0305004100019125.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford, first edition.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170. doi:10.1111/j.1467-9868.2009.00730.x.

Julious, S. A. (2023). *Sample Sizes for Clinical Trials*. Chapman and Hall/CRC. doi:10.1201/9780429503658.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:10.1080/01621459.1995.10476572.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934. doi:10.1080/01621459.1995.10476592.

Kieser, M. (2020). *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*. Springer International Publishing. doi:10.1007/978-3-030-49528-2.

Ly, A. and Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST*, 31:1121–1142. doi:10.1007/s11749-022-00819-w.

Matthews, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman and Hall/CRC, New York. doi:10.1201/9781420011302.

Micheloud, C. and Held, L. (2022). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:10.1214/21-sts828.

MIST Study Group (1998). Randomised trial of efficacy and safety of inhaled zanamivir in treatment of influenza A and B virus infections. *The Lancet*, 352(9144):1877–1881. doi:10.1016/s0140-6736(98)10190-3.

O'Hagan, A., Stevens, J. W., and Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201. doi:10.1002/pst.175.

O'Hagan, A., Stevens, J. W., and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine*, 20(5):733–753. doi:10.1002/sim.861.

Pawel, S., Consonni, G., and Held, L. (2023). Bayesian approaches to designing replication studies. *Psychological Methods*. doi:10.1037/met0000604.

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

Pawel, S., Ly, A., and Wagenmakers, E.-J. (2024). Evidential calibration of confidence intervals. *The American Statistician*, 78(1):1–11. doi:10.1080/00031305.2023.2216239.

Perneger, T. V. (2021). How to use likelihood ratios to interpret evidence from randomized trials. *Journal of Clinical Epidemiology*, 136:235–242. doi:10.1016/j.jclinepi.2021.04.010.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Royall, R. (1997). *Statistical Evidence: A likelihood paradigm*. Chapman & Hall, London.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142. doi:10.3758/s13423-017-1230-y.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2):322–339. doi:10.1037/met0000061.

Schönbrodt, F. D. and Stefan, A. M. (2019). *BFDA: An R package for Bayes factor design analysis (version 0.5.0)*. URL https://github.com/nicebread/BFDA.

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester. doi:10.1002/0470092602.

Stefan, A. M., Gronau, Q. F., and Wagenmakers, E.-J. (2024). Interim design analysis using Bayes factor forecasts. *Psychological Methods*. doi:10.1037/met0000641.

Strug, L. J. (2018). The evidential statistical paradigm in genetics. *Genetic Epidemiology*, 42(7):590–607. doi:10.1002/gepi.22151.

Strug, L. J., Rohde, C. A., and Corey, P. N. (2007). An introduction to evidential sample size calculations. *The American Statistician*, 61(3):207–212. doi:10.1198/000313007x222488.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104. doi:10.1080/00107510802066753.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804. doi:10.3758/bf03194105.

Wagenmakers, E.-J. (2022). Approximate objective Bayes factors from *P*-values and sample size: The $3p\sqrt{n}$ rule. doi:10.31234/osf.io/egydq.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):185–191. doi:10.1111/1467-9884.00075.

Xie, Y. (2024). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. URL https://yihui.org/knitr/. R package version 1.46.

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2024-05-30 16:11:36.126434 Europe/Zurich

sessionInfo()

## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] BayesRep_0.42.2 lamW_2.2.3      xtable_1.8-4    bfpwr_0.1
## [5] knitr_1.46
##
## loaded via a namespace (and not attached):
## [1] compiler_4.4.0     tools_4.4.0        Rcpp_1.0.12        highr_0.10
## [5] xfun_0.43          RcppParallel_5.1.7 evaluate_0.23
```