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

Preprint version November 13, 2024

Abstract

Determining an appropriate sample size is a critical element of study design, and the method used to determine it should be consistent with the planned analysis. When the planned analysis involves Bayes factor hypothesis testing, the sample size is usually desired to ensure a sufficiently high probability of obtaining a Bayes factor indicating compelling evidence for a hypothesis, given that the hypothesis is true. In practice, Bayes factor sample size determination is typically performed using computationally intensive Monte Carlo simulation. Here, we summarize alternative approaches that enable sample size determination without simulation. We show how, under approximate normality assumptions, sample sizes can be determined numerically, and provide the R package `bfpwr` for this purpose. Additionally, we identify conditions under which sample sizes can even be determined in closed-form, resulting in novel, easy-to-use formulas that also help foster intuition, enable asymptotic analysis, and can also be used for hybrid Bayesian/likelihoodist design. Furthermore, we show how power and sample size can be computed without simulation for more complex analysis priors, such as Jeffreys-Zellner-Siow priors or non-local normal moment priors. Case studies from medicine and psychology illustrate how researchers can use our methods to design informative yet cost-efficient studies.
Keywords: Bayesian hypothesis testing, design prior, evidence, likelihood ratio, 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 waste samples which 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 goals of the analysis (Anderson and Kelley, 2024), 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 (with two groups of equal size) for continuous outcome data, based on a frequentist hypothesis test of a mean difference, 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 is 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, 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), such as, adaptations 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) enables 'quick-and-dirty' calculations that are often accurate enough for practical purposes. It also helps fostering intuition and is therefore useful, for example, in teaching of statistics. Finally, the formula is helpful for theoretical derivations such as 'how much do we need to increase the sample size if we reduce $\alpha = 0.05$ down to $\alpha = 0.005$?' (Benjamin et al., 2017).

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 (Morey et al., 2016; Heck et al., 2023), or physics (Trotta, 2008). Bayes factor tests are conceptually different from frequentist tests in several ways. For example, they can 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, 2018; Schönbrodt et al., 2017; Pawel et al., 2023; Stefan et al., 2024). However, to our knowledge, there are no simple formulas 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, 2018; 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 used in the analysis) and the design prior (the prior distribution for the parameter used in the design). A distinction must be made between point priors and normal priors, both in the design and in the analysis. Point analysis priors lead to Bayes factors reducing to likelihood ratios, the analysis thereby corresponding with a frequentist 'likelihoodist' analysis (Edwards, 1971; Royall, 1997; Blume, 2002; Strug, 2018; Cahusac, 2020), while point design priors lead to 'conditional power', which corresponds to traditional frequentist power. In contrast, normal analysis and design priors can account for parameter uncertainty, producing Bayes factors that differ from likelihood ratios and 'predictive power' that differs from frequentist power (see e.g., O'Hagan et al., 2005; Micheloud and Held, 2022; Grieve, 2022, for conditional/predictive power related to posterior tail probability analyses).

Based on the point/normal prior distinction, we find that closed-form sample sizes are available

for Bayes factors with point analysis priors (i.e., likelihood ratios) along with point or normal design priors, and for Bayes factors with local normal analysis and design priors (i.e., normal priors centered on the null value). Table 1 provides an overview of our results. In addition to our novel formulas, we summarize sample size determination for Bayes factors with normal priors based on numerical root-finding (which has been done before, e.g., in Weiss, 1997; Pawel et al., 2023), and show how it can be extended to more advanced prior distributions, such as normal moment priors and Jeffreys-Zellner-Siow priors. While root-finding approaches also do not produce closed-form sample sizes, computations are deterministic and usually faster than with simulation approaches. To facilitate reuse of our results, all methods are made available through our R package `bfpwr`.

Table 1: Availability of closed-form sample size formulas for Bayes factor hypothesis test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with either a point prior or a normal analysis prior assigned to θ under H_1, and data in the form of a normally distributed parameter estimate $\hat{\theta} \mid \theta \sim \mathrm{N}(\theta, \sigma^2_{\hat{\theta}}/n)$. In all cases, the power can be computed in closed-form.

	Analysis prior	
Design prior	Point prior (likelihood ratio)	Normal prior (Bayes factor)
Point prior (conditional power)	✓ Equation (9)	✗ Unavailable
Normal prior (predictive power)	✓ Equation (8)	✓ Equation (12) for local normal priors

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 various design priors (Section 3). Combining these results, we derive several formulas for the sample size under different constellations of design and analysis priors (Section 4). Examples from medicine and psychology then illustrate how our formulas can be used by researchers in practice (Section 5). Section 6 illustrates possible extensions of our framework to other popular types of analysis priors, such as informed t priors (Gronau et al., 2020) and non-local normal moment priors (Johnson and Rossell, 2010). The paper ends with a closing discussion of our results and final remarks on limitations and extensions (Section 7). Our R package `bfpwr` that implements the developed methods is illustrated in Appendix A.

2 Bayes factor analysis

To derive a sample size formula, 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 a future data set with effective sample size n. Suppose that the estimate's standard error is of the form $\sigma_{\hat{\theta}}/\sqrt{n}$ where $\sigma^2_{\hat{\theta}}$ is the variance of one effective observation. For example, if the data are normally distributed with known variance σ^2 and the parameter of interest is their mean θ estimated with the sample mean $\hat{\theta}$ (the 'one-sample' case), we have that $\mathrm{Var}(\hat{\theta}) = \sigma^2/n$ so the unit variance is $\sigma^2_{\hat{\theta}} = \sigma^2$ and the effective sample size n is the number of observations. If there were another group of n normally distributed observations with a potentially different mean but the same known variance σ^2, and we were interested in the mean difference θ estimated with the empirical mean difference $\hat{\theta}$ (the 'two-sample' case), we would have that $\mathrm{Var}(\hat{\theta}) = (2\sigma^2)/n$ so the unit variance would be $\sigma^2_{\hat{\theta}} = 2\sigma^2$

and the effective sample size would be the number of observations per group n. For many commonly used estimators, it is reasonable to assume that the estimate (but not necessarily the underlying data) is approximately normally distributed around θ with variance equal to the squared standard error, i.e., $\hat{\theta} \mid \theta \sim \mathrm{N}(\theta, \sigma_{\hat{\theta}}^2/n)$. Table 2 shows common types of parameter estimates and the resulting interpretation of the effective sample size n and the unit variance $\sigma_{\hat{\theta}}^2$. It is important to note that these are only approximations and they may be inadequate in certain situations, such as small sample sizes or when the assumed variance σ^2 is strongly misspecified (Spiegelhalter et al., 2004). The frequentist sample size formula (1) can also be cast in this framework; it assumes continuous outcome data and a mean difference parameter (second row in Table 2). As such, the formula could be generalized to other settings by replacing $2\sigma^2$ in the numerator with other unit variances from Table 2, which would in turn change the interpretation of n and $\hat{\theta}$.

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

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

Assume now a point null hypothesis that postulates that θ equals a certain null value $H_0\colon \theta = \theta_0$ and an alternative hypothesis that postulates 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 standard deviation τ determines its uncertainty. A point alternative at μ may be obtained by letting the standard deviation of the prior go to zero. Whenever we write $\tau = 0$, we informally refer to a point prior at μ, since for all calculations in this paper this notation leads to the same results as a more formal treatment of point priors. The Bayes factor is then given by the updating factor of the prior to posterior odds of H_0 versus H_1, i.e.,

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

A Bayes factor less than one ($\mathrm{BF}_{01} < 1$) indicates evidence for the alternative hypothesis H_1, while a Bayes factor greater than one ($\mathrm{BF}_{01} > 1$) indicates evidence for the null hypothesis H_0. The larger the deviation of the Bayes factor from one, the stronger the evidence. Conventional thresholds for

substantial and strong evidence for the alternative (null) hypothesis are $1/3$ and $1/10$ (3 and 10), respectively. A Bayes factor $1/3 < \text{BF}_{01} < 3$ is typically interpreted as absence of evidence for either hypothesis, calling for more data to discern the more appropriate hypothesis (Jeffreys, 1939; Held and Ott, 2018).

The Bayes factor (2) is implemented in our package in the function `bf01`. It has already appeared in the literature in one form or another (e.g., in Weiss, 1997; De Santis, 2004; Spiegelhalter et al., 2004; Dienes, 2014; 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. 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 – either in favor of the alternative H_1 over the null hypothesis H_0 with a Bayes factor (2) smaller than some threshold $k < 1$ (e.g., $k = 1/3$ or $k = 1/10$) or in favor of H_0 over H_1 with a Bayes factor greater than $k > 1$ (e.g., $k = 3$ or $k = 10$). 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 in the design of the study (O'Hagan and Stevens, 2001; O'Hagan et al., 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. Point priors, and as such classical sample size determination with an 'assumed parameter', then represent a special case of the normal prior where the standard deviation becomes infinitesimally small ($\tau_d = 0$). In addition, specifying a point design prior at the null value ($\mu_d = \theta_0$ and $\tau_d = 0$) allows us to compute the probability of finding compelling evidence for the true null hypothesis H_0, as well as the probability of finding misleading evidence for the alternative when the null hypothesis is true (the 'type I error rate'). Presenting the latter probability can be useful for demonstrating that the chosen design is appropriately 'calibrated' (Dawid, 1982; Rubin, 1984; Little, 2006; Grieve, 2016).

Such a normal design prior induces a predictive distribution $\hat{\theta} \mid n, \mu_d, \tau_d \sim \mathrm{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$ for the future parameter estimate $\hat{\theta}$, which is again a normal distribution centered around the design prior mean μ_d but with a variance given by the sum of the squared standard error $\sigma_{\hat{\theta}}^2/n$ and the design prior variance τ_d^2. Under this distribution, the distribution of the Bayes factor with normal analysis prior (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 $\tau = 0$ (point analysis prior under the alternative) and $\tau > 0$ (normal analysis prior under the alternative) need to be distinguished, as the resulting Bayes factor distributions take a different form, and only the latter has been considered previously in the Bayes factor literature.

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

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

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

$$Z = \frac{1}{\sqrt{\tau_d^2 + \sigma_{\hat{\theta}}^2/n}} \left\{ \frac{\sigma_{\hat{\theta}}^2 \log k}{n(\theta_0 - \mu)} + \frac{\theta_0 + \mu}{2} - \mu_d \right\}, \tag{4}$$

see Appendix B for details. We may also want to compute the probability that the Bayes factor is greater than k (e.g., to determine the probability of compelling evidence for a true null hypothesis), which is simply one minus the probability (3).

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 analysis prior, 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

$$\lim_{n \to \infty} \Pr(\mathrm{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau = 0) = \begin{cases} 1 - \Phi(Z_{\lim}) & \text{if } \mu - \theta_0 > 0 \\ \Phi(Z_{\lim}) & \text{if } \mu - \theta_0 < 0 \end{cases} \tag{5}$$

with $Z_{\lim} = (\theta_0 + \mu - 2\mu_d)/(2\tau_d)$, the limit of (4) for increasing n. When the design prior is also a point prior ($\tau_d = 0$), $Z_{\lim}$ diverges and the limiting power (5) 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 ($\mu_d = (\theta_0 + \mu)/2$), the limiting power approaches a half. On the other hand, for a normal design prior ($\tau_d > 0$), the limiting power is bounded by a value in between (and not including) zero and one given by (5). 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 hybrid Bayesian/frequentist power (Spiegelhalter et al., 2004; Micheloud and Held, 2022; Grieve, 2022).

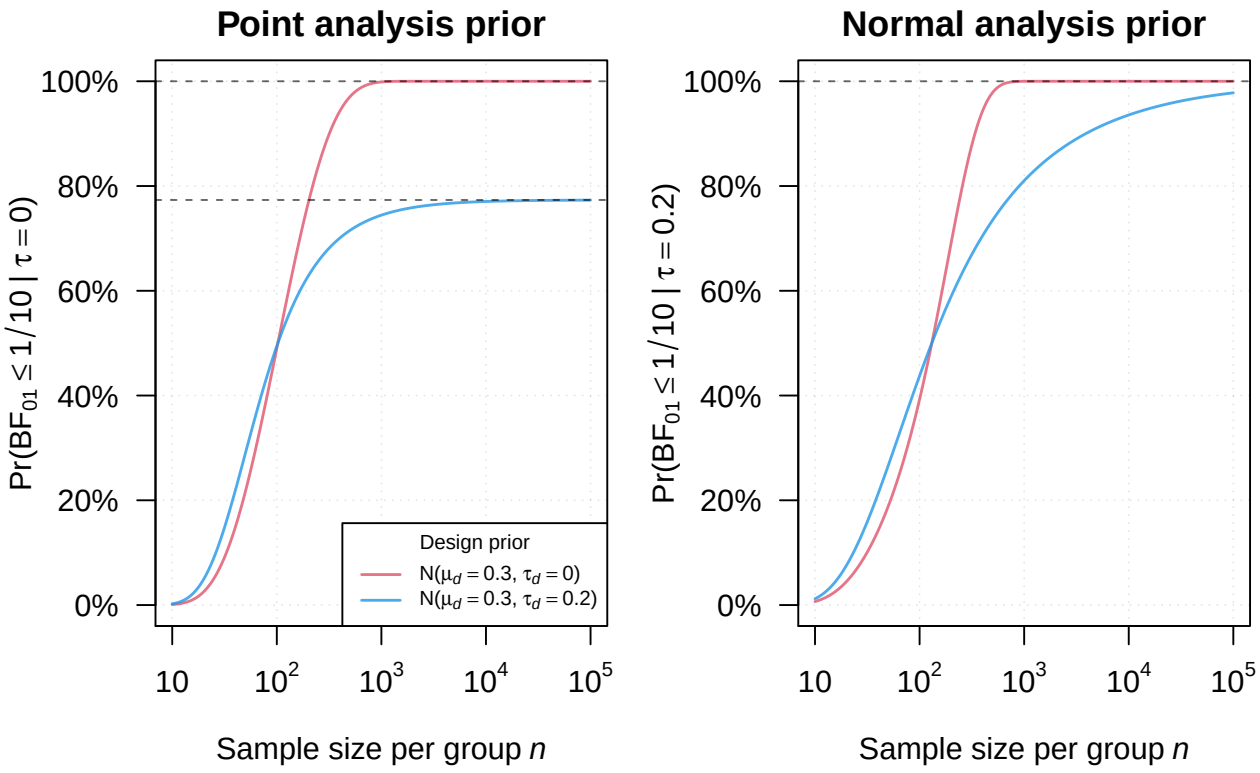


Figure 1: Examples of Bayes factor power curves computed with equation (3) (left) and (6) (right). Both Bayes factors contrast $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ for a standardized mean difference parameter θ with unit variance $\sigma^2_{\hat{\theta}} = 2$ with either a point analysis prior ($\tau = 0$, left plot) or a normal analysis prior ($\tau = 0.2$, right plot) both with $\mu = 0.3$. The power is computed under either a point (red curve) or normal design prior (blue curve), both at the same location as the analysis prior. The dashed lines depict the corresponding limiting power values.

The left plot in Figure 1 shows examples of two power functions related to the Bayes factor for a standardized mean difference parameter with point analysis prior. The power is computed under either a point (red curve) or a normal design prior (blue curve). We see that increasing the sample size n increases the power in both cases. However, as known from (5), the power under the normal design prior is bounded by $1 - \Phi(Z_{\text{lim}}) = 77.3\%$ while the power under the point design prior can be increased up to 100%. Finally, the curves cross at 50%, the intuition being that with the normal design prior there is more uncertainty, and hence the power is always closer to 50%.

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

$$\Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau > 0) = \Phi(-\sqrt{X} - M) + \Phi(-\sqrt{X} + M) \tag{6}$$

with

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

and

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

see Appendix B for details. Again, to compute the probability of a Bayes factor in favor of the null hypothesis, we have to take one minus the probability (6) along with a level $k > 1$.

Unlike the power function based on the Bayes factor with point analysis prior (3), the power function based on the Bayes factor with normal analysis prior (6) can be increased arbitrarily close to one by increasing the sample size n (see Appendix C). That is, we have

$$\lim_{n \to \infty} \Pr(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d, \tau > 0) = 1 \tag{7}$$

regardless of whether the design prior is a point prior ($\tau_d = 0$) or a normal prior ($\tau_d > 0$), provided that the design prior is not equal to the point null hypothesis itself. This is expected because Bayes factors contrasting 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).

The right plot in Figure 1 shows example power curves for the Bayes factor related to a standardized mean difference parameter with normal analysis prior, and computed under either point (red curve) or normal design prior (blue curve). We see that in both cases an increase in sample size also increases the power. As expected from (7), the power can increase to 100% in both cases, although it approaches the limit much slower under the normal than under the point design prior because there is more uncertainty. Finally, as with the point analysis prior Bayes factor, the curves cross at 50%.

4 Sample size determination

Both power functions (3) and (6) 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 (as pioneered by Weiss, 1997). It is important to emphasize again that one can also compute a power curve in favor of the null hypothesis. That is, one can look at one minus the power functions (3) and (6) along with a Bayes factor threshold $k > 1$ and fixing the design prior to the null hypothesis ($\mu_d = \theta_0$, $\tau_d = 0$). In this way, sample sizes can be determined that ensure a desired probability of compelling evidence for both the null and the alternative.

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 analysis priors ($\tau = 0$) and normal analysis priors ($\tau > 0$). We start again with the former.

4.1 Bayes factor with point analysis prior

Assuming that the alternative μ is larger than the null θ_0 and setting the power function (3) equal to a target power $1 - \beta$, we obtain a quadratic equation in the sample size n. Its solution can be expressed as

$$n = \left[\left\{ z_{1-\beta} + \sqrt{z_{1-\beta}^2 - \frac{\Delta_{\mu_d} \log k^2}{\Delta_\mu} + \left(\frac{\tau_d \log k^2}{\Delta_\mu} \right)^2} \right\}^2 - \left(\frac{\tau_d \log k^2}{\Delta_\mu} \right)^2 \right] \times \frac{\sigma_{\hat{\theta}}^2}{\Delta_{\mu_d}^2 - 4 z_{1-\beta}^2 \tau_d^2} \tag{8}$$

where $\Delta_{\mu_d} = 2\mu_d - \mu - \theta_0$ is the 'generalized effect size', which reduces to the ordinary effect size $\Delta_\mu = \mu - \theta_0$ when the design prior mean is set to the parameter value under the alternative ($\mu_d = \mu$). From looking at (8) one can recognize that a valid sample size can only exist if the denominator in the right factor is positive, as sample sizes cannot be negative. This condition is equivalent to the target power $1 - \beta$ being lower than the limiting power (5). Of note, if for a given design prior the limiting power (5) is higher than 50%, replacing the first plus in (8) by a minus gives the sample size that leads to a target power of β instead of $1 - \beta$. It is also worth noting that the sample size formula (8) will usually produce non-integer values and hence needs to be rounded to the next larger integer in order to be an evaluable sample size in practice (an actual number of participants, animals, etc.).

To better understand the sample size formula (8), we will now investigate it closer for two special cases. First, suppose that the design prior is a point prior ($\tau_d = 0$) at μ_d, not necessarily the same as the alternative μ. This leads to (8) reducing to

$$n = \frac{\sigma_{\hat{\theta}}^2 \left\{ z_{1-\beta} + \sqrt{z_{1-\beta}^2 - (\Delta_{\mu_d} / \Delta_\mu) \log k^2} \right\}^2}{\Delta_{\mu_d}^2}. \tag{9}$$

Assuming that the tested parameter is a mean difference with unit variance $\sigma_{\hat{\theta}}^2 = 2\sigma^2$ (see Table 2, second row), we can see that (9) represents a modification of the frequentist sample size formula (1): The effect size $\Delta_\mu = \mu - \theta_0$ in the denominator of the frequentist sample size is replaced by the generalized effect size Δ_{μ_d} that takes into account the mean of the design prior μ_d, but reduces to the effect size when the design prior mean equals parameter under the alternative ($\mu_d = \mu$). Moreover, the quantile $z_{1-\alpha/2}$ in the frequentist sample size is replaced by $\sqrt{(z_{1-\beta}^2 - (\Delta_{\mu_d} / \Delta_\mu) \log k^2)}$, reflecting the fact that we are interested in a Bayes factor hypothesis test with evidence threshold k instead of a frequentist test with level α.

Second, assume that the design prior is also equal to the alternative ($\mu_d = \mu$), so that the formula (9) further reduces to

$$n = \frac{\sigma_{\hat{\theta}}^2 \left\{ z_{1-\beta} + \sqrt{z_{1-\beta}^2 - \log k^2} \right\}^2}{\Delta_\mu^2}. \tag{10}$$

The same formula (10) was also found by Strug et al. (2007) for 'evidential' sample size calculations, but is unfortunately not well known. Not surprisingly, the two formulas coincide, since Bayes factors and likelihood ratios – the measure of evidence used in evidential/likelihoodist statistics (see e.g.,

Edwards, 1971; Royall, 1997; Blume, 2002; Strug, 2018; Cahusac, 2020) – are equivalent when the Bayes factor involves only point hypotheses. Our more general formulas (8) and (9) thus enable 'hybrid Bayesian/likelihoodist' design that assumes a likelihoodist analysis but can incorporate prior knowledge and uncertainty via Bayesian design prior, similar to how Bayesian design priors can be used to incorporate parameter uncertainty in the design of frequentist hypothesis tests (see e.g., Grieve, 2022, for an overview of hybrid Bayesian/frequentist design approaches).

Table 3: Sample size per group n to obtain a Bayes factor $\text{BF}_{01} \leq k$ with at least a power of $1 - \beta$. The parameter of interest is a standardized mean difference and the analysis and design prior assume both an effect size of one ($\Delta_\mu = 1$) so that equation (10) can be used to compute the sample size in closed-form.

	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

To illustrate formula (10), we now assume that $\hat{\theta}$ is a standardized mean difference so that the unit variance is $\sigma^2_{\hat{\theta}} = 2$ and n can be interpreted as the sample size per group (see the third row in Table 2). Table 3 shows the sample size (10) based on an assumed effect size of one ($\Delta_\mu = 1$). We see, for instance, that a sample size of $n = 20$ per group is required to achieve $1 - \beta = 80\%$ power for a Bayes factor threshold $k = 1/10$. If the assumed effect size was smaller, the required sample size would become larger. For example, for a half as large effect size, the sample sizes from Table 3 quadruple, e.g., requiring a sample size of $n = 80$ per group to have $1 - \beta = 80\%$ power for a Bayes factor threshold of $k = 1/10$.

The left plot in Figure 2 shows sample size calculations for a Bayes factor at threshold $k = 1/10$ with the same point analysis and design prior as in Table 3 (red curve), but additionally illustrates the sample size formula (8) that also incorporates parameter uncertainty via a normal design prior (blue curve). We can see that for a target power above 50%, larger sample sizes are required under the normal design prior than under the point design prior, while it is reversed for a target power below 50%. Furthermore, we see that as the target power approaches its theoretical upper bound (5), the sample size goes to infinity.

4.2 Bayes factor with normal analysis prior

Finding a sample size formula becomes more difficult when we move from point to normal analysis priors. The technical reason is that when we set the power function (6) 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

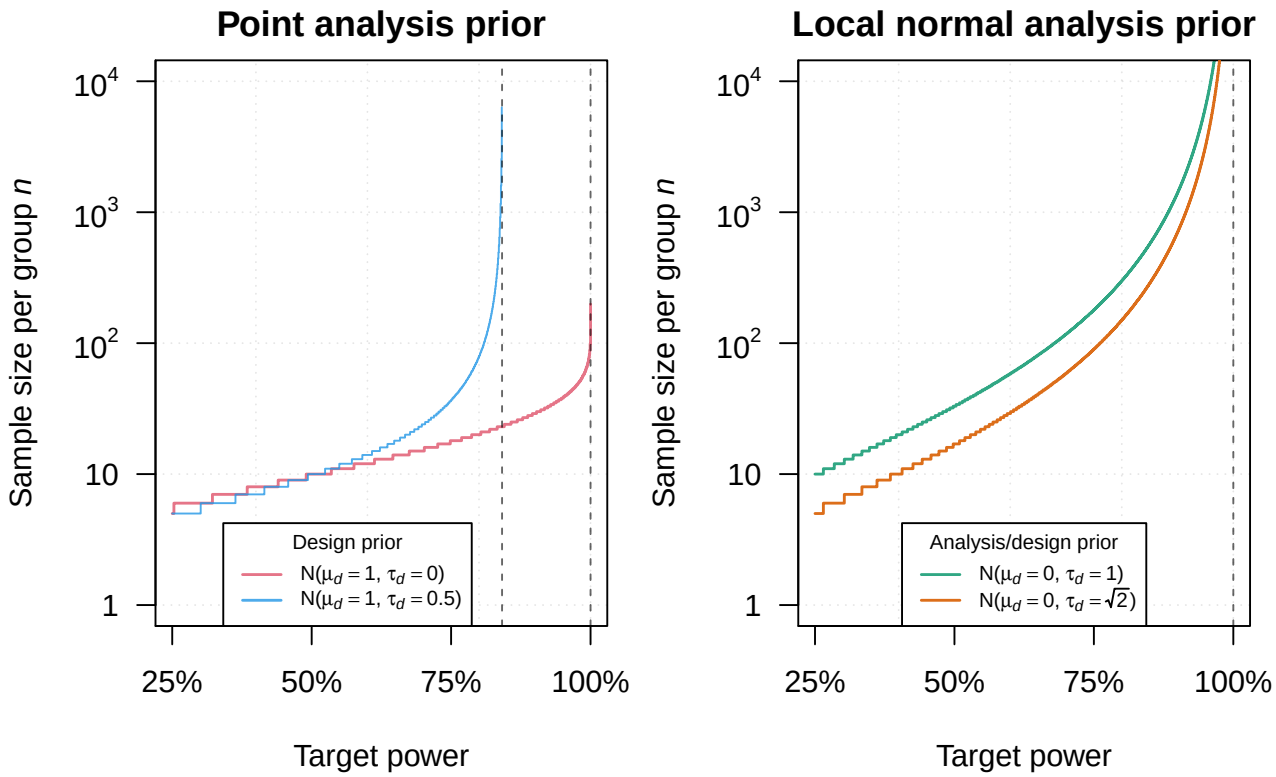


Figure 2: Examples of Bayes factor sample size calculations with equation (8) (left) and equation (12) (right). Both Bayes factors contrast $H_0 \colon \theta = 0$ against $H_1 \colon \theta \neq 0$ for a standardized mean difference parameter θ with unit variance $\sigma^2_{\hat{\theta}} = 2$ with either a point analysis prior ($\mu = 1$, $\tau = 0$, left plot) or a local normal analysis prior ($\mu = 0$, right plot), and at a threshold of $k = 1/10$. The dashed lines represent the upper bounds for the power that are achievable with a finite sample size.

transcendental equation that cannot be solved in terms of elementary functions. We were unable to find a closed-form solution in the general case. However, as we will show now, solutions can, under certain conditions, be expressed in terms of the Lambert W function (Corless et al., 1996). The Lambert W function is the function $\mathrm{W}(\cdot)$ that satisfies $\mathrm{W}(x)\exp\{\mathrm{W}(x)\} = x$, and it is therefore sometimes also called 'product logarithm'. It has many fundamental applications and has also previously appeared in the context of Bayes factor hypothesis testing (Pawel and Held, 2022; Wagenmakers, 2022; Held et al., 2022; Pawel et al., 2024).

Suppose now that the design and analysis prior 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 (6) reduces to

$$\Pr\{\mathrm{BF}_{01} \leq k \mid n, \tau_d, \mu_d = \mu = \theta_0\} = 2\Phi(-\sqrt{X}). \tag{11}$$

Further, assume that the variance of the design prior corresponds to the variance of the analysis prior

($\tau_d = \tau$). We then have that

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

Setting the power function (11) equal to a target power of $1 - \beta$ and assuming that $\log\{1 + (n\tau^2)/\sigma^2_{\hat{\theta}}\} \approx \log\{(n\tau^2)/\sigma^2_{\hat{\theta}}\}$, we obtain the following approximate sample size formula

$$n = \frac{\sigma^2_{\hat{\theta}}}{\tau^2} \underbrace{k^2 \exp\left\{-\mathrm{W}_{-1}(-k^2 z^2_{(1-\beta)/2})\right\}}_{=n_{k,\beta}} \tag{12}$$

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

We can see that the sample size (12) 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 with the frequentist sample size (1), smaller unit variances $\sigma^2_{\hat{\theta}}$ reduce the sample size (12). The prior variance τ^2 determines how large parameters are expected under the alternative hypothesis, and as such, larger prior variances lead to a reduction of sample size similar to how larger effect sizes lead to a reduction of sample size in the frequentist formula (1). Finally, the formula (12) allows us to study the potential existence of a 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 $-k^2 z^2_{(1-\beta)/2} \geq -1/e$ 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 $-1 \times z^2_{0.25} = -0.45 < -1/e = -0.37$.

Table 4: Required unit information sample size $n_{k,\beta}$ computed with equation (12) to obtain a Bayes factor $\mathrm{BF}_{01} \leq k$ 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 4 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 3 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 local normal analysis prior from Table 4 represents a parameter distribution that is centered around the null value while the point analysis prior from Table 3 represents a mean difference of one standard deviation away from the null. The former with unit information variance represents a more pessimistic assumption about the parameter than the latter. To incorporate more optimistic beliefs into the calculations we may increase the standard deviation τ of the distribution as this encodes the assumption of potentially larger parameters. For example, doubling τ leads to a four-fold decrease of the sample size (12). This is also illustrated in the right plot of Figure 2 where the prior with doubled variance (orange curve) leads to a two-fold decrease in sample size over the prior with variance of one (green curve).

The formula (12) is, to our knowledge, the first closed-form sample size formula for Bayes factor analysis with normal analysis priors. While interesting from a theoretical point of view, its practical use is perhaps more limited than the sample size formula for Bayes factors with point analysis priors (8). 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, given our limited mathematical abilities, 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. 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 mirtazapine in dementia

Banerjee et al. (2021) conducted a randomized controlled clinical trial to assess the effect of the antidepressant mirtazapine on agitated behavior in patients with dementia. The primary outcome of the trial was the Cohen-Mansfield Agitation Inventory (CMAI) score at 12 weeks after treatment start. The CMAI score is commonly used to quantify agitation in dementia and it can take values from 29 to 203. The treatment effect was quantified as the difference in mean CMAI between the mirtazapine and placebo groups θ. The null hypothesis was defined as no mean difference between mirtazapine and placebo ($H_0 \colon \theta = 0$), while the alternative hypothesis was formulated as a mean decrease of six CMAI points ($H_1 \colon \theta = -6$). With a target power of $1 - \beta = 80\%$, a two-sided level of $\alpha = 0.05$, and an assumed standard deviation of $\sigma = 15$ CMAI points, formula (1) gives a required per-group sample size of $n = 99$, which the investigators adjusted further upwards to 111 to account for attrition. Contrary to their expectations, however, the trial did not find a statistically significant effect of mirtazapine over placebo (estimated mean difference of $\hat{\theta} = -1.74$ with 95% confidence interval

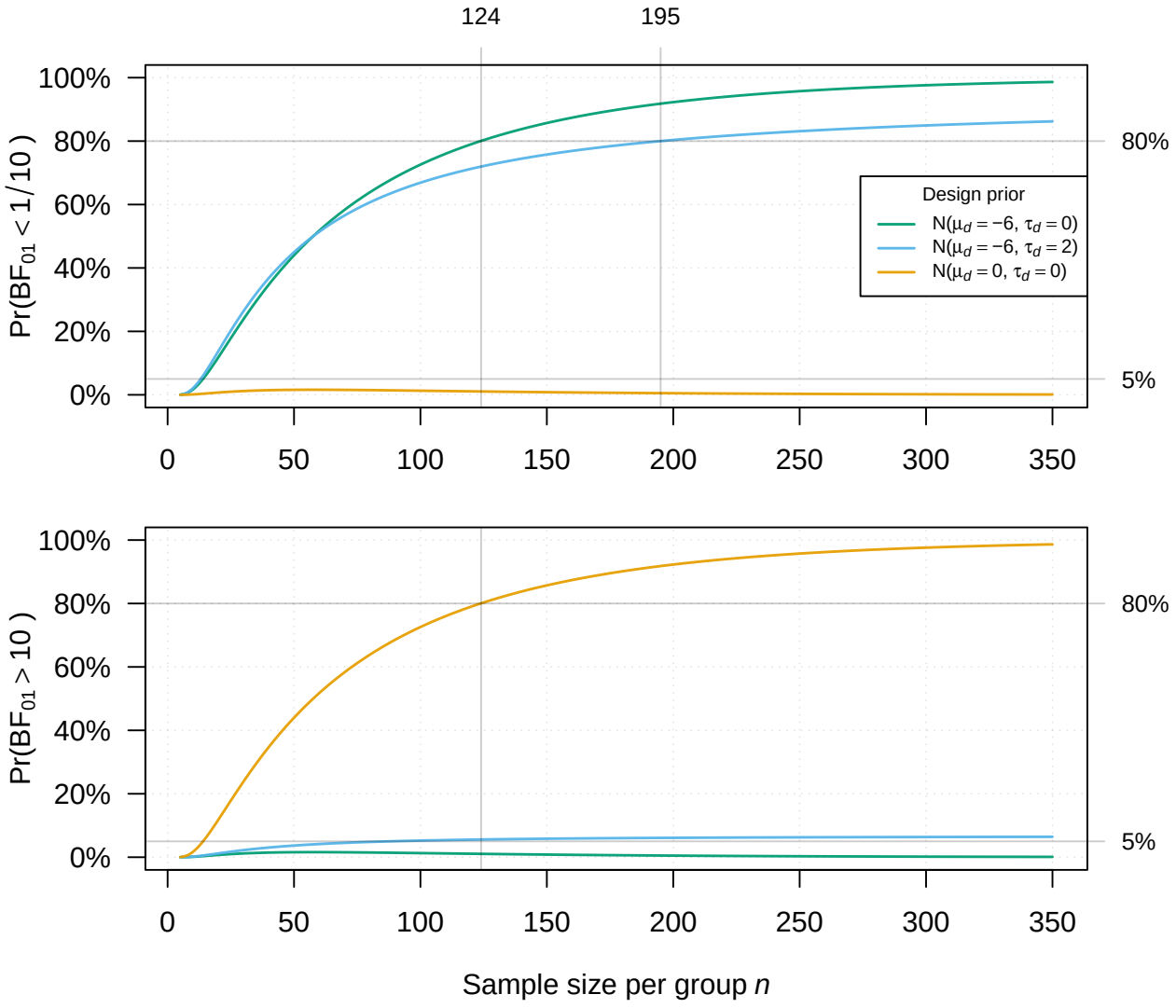


Figure 3: Power and sample size calculations for randomized clinical trial assessing the effect of mirtazapine on agitated behavior in patients with dementia (Banerjee et al., 2021). The Bayes factor assumed for the analysis contrasts $H_0 \colon \theta = 0$ to $H_1 \colon \theta = -6$ related to the mean difference θ in CMAI scores between mirtazapine and placebo. The power is computed under either a point prior at the assumed effect under the alternative (green), a normal prior that incorporates additional parameter uncertainty (blue), or under the null hypothesis (orange).

from -7.17 to 3.69 and two-sided p-value of $p = 0.53$ for the null hypothesis of no difference).

While this study was designed and analyzed using frequentist methodology, we will now examine what the analysis, power, and sample size calculations might have looked like if the planned analysis were performed using Bayes factors. To do this, we assume a point analysis prior equal to the alternative hypothesis specified by the trial investigators ($\mu = -6$ and $\tau = 0$), so that the Bayes factor corresponds to a likelihood ratio. Based on the estimated mean difference $\hat{\theta} = -1.74$ and its standard error $\sigma_{\hat{\theta}}/\sqrt{n} = 2.77$ (recalculated from the confidence interval), we can apply equation (2) and obtain a Bayes factor $\text{BF}_{01} = 2.7$. This Bayes factor indicates only anecdotal evidence in favor of the null hypothesis of no effect of mirtazapine over the specified alternative.

Taking a step back and assuming that no data have yet been collected, we can use equations (3) and (8) to calculate power and sample size. Figure 3 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 design prior the power was computed.

Focusing on the top plot, we can see from the green curve that at least $n = 124$ observations per group are required to ensure a target power of $1 - \beta = 80\%$ assuming that the same point prior is used in the design as for the analysis (i.e., a point prior at $\mu = -6$). This number increases to $n = 195$ when we move to a design prior that incorporates parameter uncertainty (blue curve), i.e., a prior that is still centered around $\mu = -6$ but with a standard deviation of $\tau_d = 2$. 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 (the type I error rate) is very low and appears to be reasonably controlled by conventional standards (i.e., below 5%) for each of the two samples sizes. This curve could be presented to a trial regulator to demonstrate that the design is adequately calibrated.

Focusing now on the bottom plot, we can see that the sample size $n = 124$ based on the point prior also ensures a power of $1 - \beta = 80\%$ 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 (10) 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 non-null design priors seems to be reasonably well controlled (below the conventional 5%) for the point design prior (green), while it is slightly inflated for the normal design prior (blue) for sample sizes larger than 88.

5.2 Comparison to Monte Carlo simulation methods

Schönbrodt and Wagenmakers (2018) proposed a Monte Carlo simulation approach for power and sample size calculations for Bayes factor analyses, and 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 can be computationally intensive and comes with Monte Carlo error.

We will now look at an adaptation of an example examined in Schönbrodt and Wagenmakers (2018, p. 133). Suppose we want to test the null hypothesis that a standardized mean difference θ is zero ($H_0 \colon \theta = 0$) versus the alternative hypothesis that it is different from zero ($H_1 \colon \theta \neq 0$). 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. As they did, we also investigate two design priors: either a point prior at $\mu_d = 0.5$, which is a convention for a 'medium' standardized mean difference in psychology (Cohen, 1992), or a normal prior at $\mu_d = 0.5$ with standard deviation $\tau_d = 0.1$ to incorporate parameter uncertainty. Figure 4 shows the resulting power curves computed from equation (6), along with the `BFDA` Monte Carlo simulation

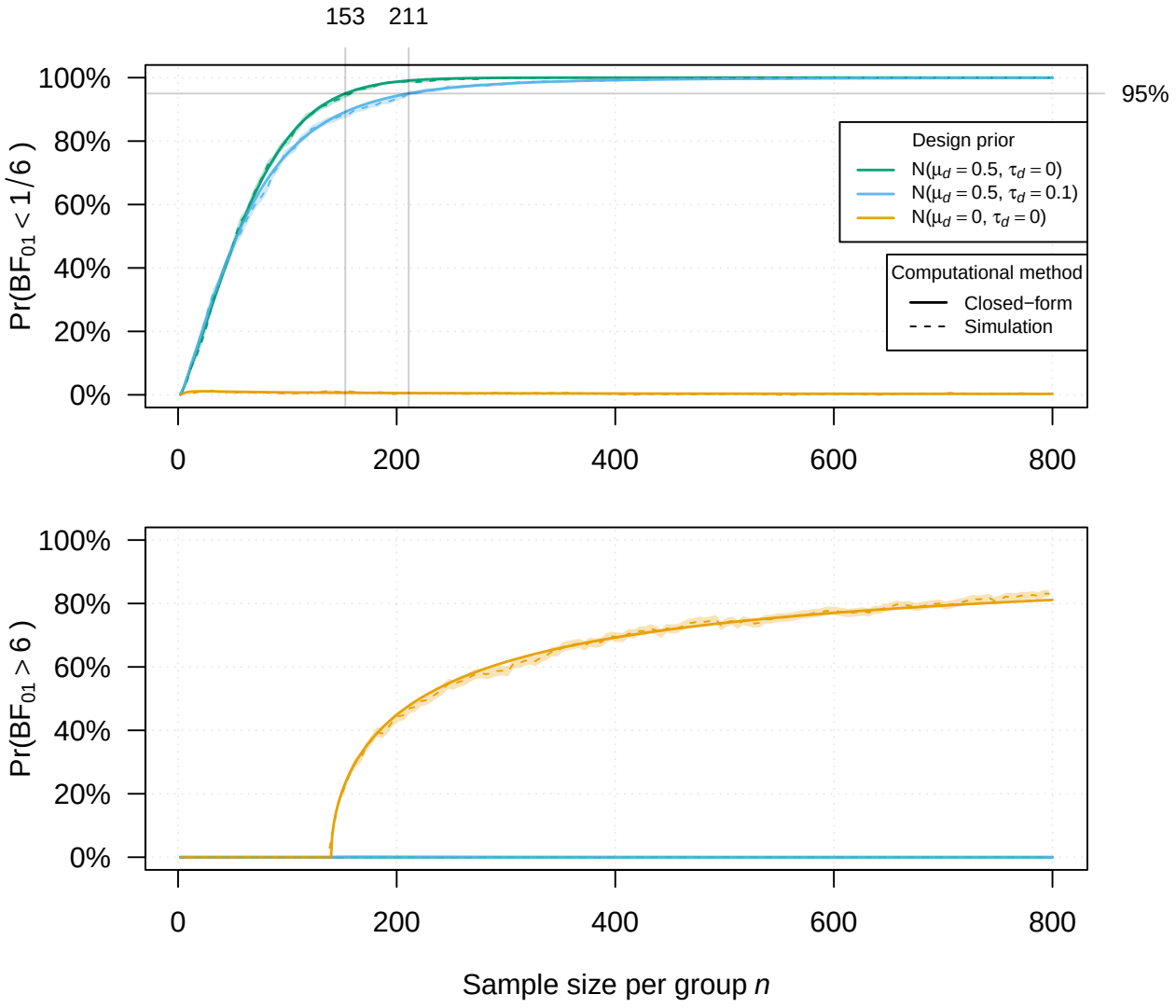


Figure 4: Power and sample size calculation for standardized mean difference example. The Bayes factor assumed for the analysis contrasts $H_0 : \theta = 0$ to $H_1 : \theta \neq 0$ with $\theta \mid H_1 \sim \text{N}(\mu = 0, \tau^2 = 1/2)$ prior assigned to the standardized mean difference θ under the alternative. The power is computed under either a point prior at a medium effect size (green), a normal prior that incorporates additional parameter uncertainty (blue), or under the null hypothesis (orange). For comparison, 1000 Monte Carlo simulations are performed with the `BFDA` package to obtain a simulation-based power curve approximation (dashed line with one Monte Carlo standard error band).

approximations (with Monte Carlo standard error bands) for comparison.

We see that closed-form and simulation curves closely align in all cases. The latter shows a maximum Monte Carlo error of 0.016. This Monte Carlo error may or may not be negligible for practical purposes, as it may lead to slight over- or underestimation of the sample size. For example, in animal studies, an overestimation of the sample size by even one animal may be highly problematic and thus a higher number of simulations may be preferred, whereas in other studies this would not be a problem at all.

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 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 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 95%, a sample size of $n = 6691$ per group would be required (not shown in Figure 4). This illustrates the well-known fact that evidence accumulates slower for the null than the alternative when the Bayes factor involves testing a point null against a composite alternative with normal analysis prior (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 constant at zero over the entire range of sample sizes.

Appendix E shows a more systematic comparison with Monte Carlo simulation, including many other design and analysis prior conditions. To verify that the methods work as intended under a given condition, a sample size was first calculated using closed-form (if available) or root-finding methods to ensure a desired target power for a specified Bayes factor threshold. Subsequently, data were simulated based on this sample size and the design prior. Bayes factors were then calculated from the simulated data, and power was estimated empirically from the proportion of Bayes factors below the specified threshold. These simulation-based power estimates were spread closely around the specified target power in all conditions.

6 Extensions

The types of Bayes factors that we considered so far are limited to data in the form of asymptotically normally distributed parameter estimates, and normal or point priors in the analysis. Researchers may also want to use different data models or prior distributions in the analysis. In the following, we outline two extensions that modify the Bayes factor used for the analysis, while still retaining the normal likelihood and point/normal design prior used for the design.

6.1 Informed Bayesian *t*-test

Gronau et al. (2020) proposed a Bayes factor for testing a standardized mean difference parameter based on normally distributed data with unknown variance – the same situation where a classical t-test would be used. The Bayes factor is given by

$$\text{BF}_{01} = \frac{\text{T}_\nu(t \mid 0, 1)}{\int_{-\infty}^{+\infty} \text{NCT}_\nu(t \mid \theta\sqrt{n})\, \text{T}_\kappa(\theta \mid \mu, \tau)_{[a,b]}\, \text{d}\theta} \tag{13}$$

with t the t-test statistic, n the effective sample size (the actual number of observations/pairs for one-sample/paired t-tests, or $n = 1/(1/n_1 + 1/n_2)$ for two-sample t-tests), $\text{T}_\nu(\cdot \mid \mu, \tau)$ the location-scale t density with degrees of freedom ν, location μ, and scale τ, and $\text{NCT}_\nu(\cdot \mid \lambda)$ the non-central t density with degrees of freedom ν and non-centrality parameter λ (Johnson et al., 1995, chapters

28 and 31). The subscript $[a, b]$ denotes truncation of a distribution to the interval from a to b, i.e., $f(x)_{[a,b]} = \{f(x)1_{[a,b]}(x)\}/\{F(b) - F(a)\}$. For example, for a one-sided test in positive direction, we have $a = 0$ and $b = +\infty$, while for a two-sided test we have $a = -\infty$ and $b = +\infty$. When the prior under the alternative is set to a (scaled) Cauchy distribution ($\kappa = 1$, $\mu = 0$, $a = -\infty$, $b = +\infty$), the Bayes factor reduces to the 'Jeffreys-Zellner-Siow' (JZS) Bayes factor (Jeffreys, 1939; Zellner and Siow, 1980), which is often used as a 'default' Bayes factor in the social sciences (Rouder et al., 2009). Setting other values for these parameters allows data analysts to incorporate directionality, such as a one-sided JZS Bayes factor (Wetzels et al., 2009), or prior knowledge about the standardized mean difference, potentially improving the efficiency of the test (Stefan et al., 2019). The Bayes factor (13) is also implemented in our package in the function `tbf01`.

A complication in power and sample size calculations based on the Bayes factor (13) is that it is not available in closed-form but requires numerical evaluation of the integral in the denominator. Perhaps for this reason, so far power and sample size have been computed using simulation. To compute the power for a given sample size without simulation, we propose to use the following two-step procedure instead:

1. For a given sample size n and analysis prior (specified with the parameters μ, τ, κ, a, b), use numerical root-finding to determine the 'success region' S in terms of t-statistics where the Bayes factor is equal or below the threshold k, i.e., $S = \{t : \text{BF}_{01} \leq k\}$. For a two-sided test, S is typically a region specified by two critical values $S = (-\infty, t_{\text{crit}-}] \cup [t_{\text{crit}+}, +\infty)$.

2. For a given design prior, compute the probability of obtaining a t-statistic included in S. For example, based on a normal design prior $\theta \sim \text{N}(\mu_d, \tau_d^2)$ and assuming a two-sample test with equally sized groups and that the data variance σ^2 is known, we have that approximately $t \mid n, \mu_d, \tau_d \sim \text{N}\{\mu_d\sqrt{n/2}, 1 + (n\tau^2)/2\}$, hence the power can be computed by

$$\Pr\left(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d\right) = \Phi\left(\frac{t_{\text{crit}-} - \mu_d\sqrt{n/2}}{\sqrt{1 + (n\tau_d^2)/2}}\right) + \Phi\left(\frac{\mu_d\sqrt{n/2} - t_{\text{crit}+}}{\sqrt{1 + (n\tau_d^2)/2}}\right).$$

To calculate a sample size that ensures a certain target power, we can again use numerical root-finding. To illustrate this, we now revisit the example from Schönbrodt and Wagenmakers (2018) which was already re-analyzed in Section 5.2 with a normal analysis prior. We now specify the analysis parameters $\mu = 0, \tau = 1/\sqrt{2}, \kappa = 1, a = 0, b = +\infty$, i.e., a one-sided JZS Bayes factor with scale $1/\sqrt{2}$ (Wetzels et al., 2009), as in the original analysis. Using the previously described iterative approach, we calculate that a sample size of $n = 143$ is required under a point design prior at $\mu_d = 0.5$ to achieve a target power of 95%. This differs only slightly from the simulation-based $n = 146$ reported by Schönbrodt and Wagenmakers, possibly due to Monte Carlo error.

6.2 Normal moment priors

Non-local priors are prior distributions that have zero probability density/mass at the null value. They were introduced to allow evidence for the null to accumulate more quickly if it is indeed true (Johnson and Rossell, 2010). A convenient class of non-local priors are normal moment priors, which

have a density of the form $\text{NM}(\theta \mid \theta_0, \tau) = \text{N}(\theta \mid \theta_0, \tau^2) \times (\theta - \theta_0)^2/\tau^2$ with location θ_0 and spread τ. Figure 5 shows three examples of normal moment priors. We see that the distributions have a density of zero at the null value $\theta_0 = 0$, encoding the assumption that the null is the least likely parameter value under the alternative. Normal moment priors have two modes at $\pm\tau\sqrt{2}$, which gives a convenient way to elicit a prior, as τ may, for example, be set so that the modes equals two parameter values deemed most plausible under the alternative.

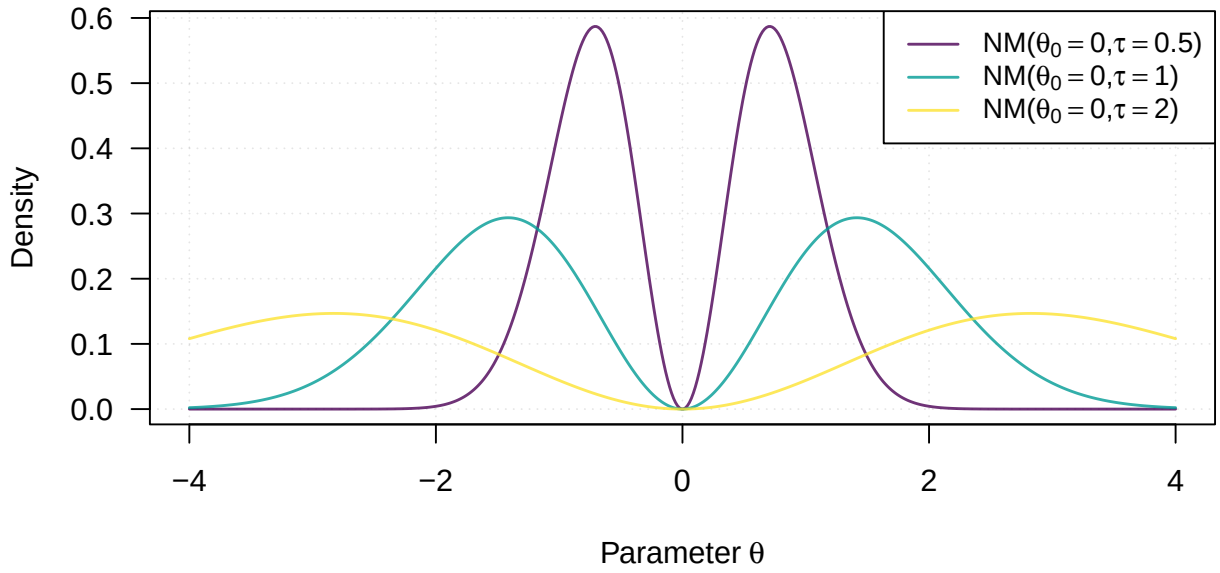


Figure 5: Illustration of a normal moment prior distribution for different spread parameters τ.

We will now extend power and sample size calculation to Bayes factors with non-local moment analysis priors and normal design priors. The Bayes factor based on a normally distributed parameter estimate $\hat{\theta} \mid \theta \sim \text{N}(\theta, \sigma_{\hat{\theta}}^2/n)$ that contrasts $H_0\colon \theta = \theta_0$ to $H_1\colon \theta \neq \theta_0$ with prior $\theta \mid H_1 \sim \text{NM}(\theta_0, \tau)$ assigned to θ under H_1 is given by

$$\text{BF}_{01} = \left(1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}\right)^{3/2} \exp\left[-\frac{1}{2}\frac{n(\hat{\theta} - \theta_0)^2}{\sigma_{\hat{\theta}}^2\{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)\}}\right] \left[1 + \frac{n(\hat{\theta} - \theta_0)^2}{\sigma_{\hat{\theta}}^2\{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)\}}\right]^{-1} \tag{14}$$

see e.g., Pramanik and Johnson (2024) or Pawel et al. (2024). Under a normal design prior $\theta \sim \text{N}(\mu_d, \tau_d^2)$, the power to obtain a Bayes factor below a threshold k can then be expressed in closed-form

$$\Pr\left(\text{BF}_{01} \leq k \mid n, \mu_d, \tau_d\right) = \Phi\left(-\sqrt{Y} - A\right) + \Phi\left(-\sqrt{Y} + A\right) \tag{15}$$

with

$$Y = \left(2\text{W}_0\left[\frac{\{1 + (n\tau^2)/\sigma_{\hat{\theta}}^2\}^{3/2}\sqrt{e}}{2k}\right] - 1\right)\left\{\frac{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)}{1 + (n\tau_d^2)/\sigma_{\hat{\theta}}^2}\right\} \quad \text{and} \quad A = \frac{\mu_d - \theta_0}{\sqrt{\tau_d^2 + \sigma_{\hat{\theta}}^2/n}}$$

with W_0 the principal branch of the Lambert W function, see Appendix F for details. Based on the power function (15), numerical root-finding can again be used to determine the sample size such that

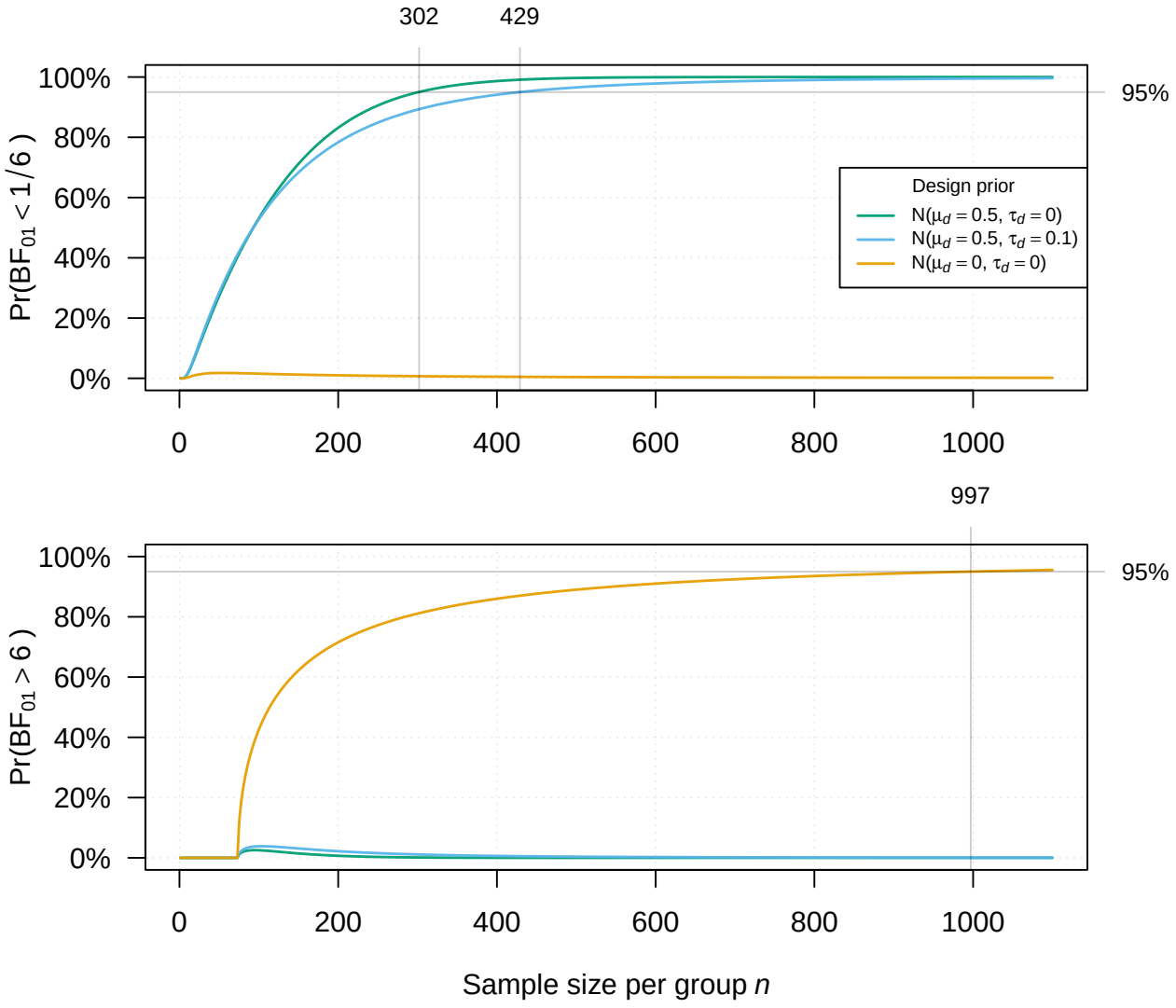


Figure 6: Power and sample size calculations for standardized mean difference example from Schönbrodt and Wagenmakers (2018, p. 133) based on normal moment prior Bayes factor. The normal moment analysis prior has modes at a medium effect size ($\tau = 0.5/\sqrt{2}$). The power is computed under either a point prior at a medium effect size (green), a normal prior that incorporates additional parameter uncertainty (blue), or under the null hypothesis (orange).

a specified target power is achieved. Methods for power and sample size calculations for normal moment prior Bayes factors are implemented in our R package (functions `nmbf01` and `powernmbf01`).

Figure 6 shows again power and sample size calculations for the example from Schönbrodt and Wagenmakers (2018) considered earlier. This time, however, we specify a normal moment analysis prior with modes at a medium standardized mean difference effect size of $\theta = \pm 0.5$, which translates to a prior spread parameter of $\tau = 0.5/\sqrt{2} = 0.35$. Using the same design priors as before, we can see that the sample sizes required to achieve a target power of 95% are larger than when using the normal analysis prior as in Section 5.2. For example, under the point design prior a sample size of $n = 302$ per group is required while only $n = 153$ were required with the normal analysis prior (see Figure 4). At the same time, the sample size required to find evidence in favor of the null hypothesis

with a target power of 95% (bottom plot) is drastically reduced ($n = 997$) compared to the normal analysis prior ($n = 6691$). This is expected since the evidence for a true null hypothesis accumulates faster with normal moment priors (Johnson and Rossell, 2010).

7 Discussion

We presented methods for performing power and sample size calculations for the situation when data are analyzed with Bayes factor hypothesis tests. These methods rely on the approximate normality of parameter estimates (but not of the underlying data), which is a common assumption underlying many methods for power and sample size calculation. We have synthesized and extended previous theoretical results on power functions for Bayes factors and implemented them in an R package `bfpwr`. We also derived novel sample size formulas that are easy-to-use, help fostering intuition, and enable understanding of theoretical properties such as asymptotic power or (non-)existence of sample size for a target power and design prior. 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.

An important issue in any Bayesian design and analysis is the choice of prior distributions. While there are differences between point and normal priors in terms of the closed-form availability of sample sizes (see Table 1), we believe that the choice between them should be guided only by prior knowledge and the questions researchers wish to answer. Point analysis priors seem more appropriate when researchers have precise hypotheses to test, they can also be motivated from a likelihood perspective, and they are easier to communicate to a non-statistical audience. In contrast, normal analysis priors can incorporate uncertainty and may therefore be more appropriate for exploratory settings where little is known about the phenomenon under investigation (Held and Ott, 2018). Similar considerations apply to the choice of the design prior. For example, a point design prior may be specified at a minimally relevant parameter value if such a value can be formulated. On the other hand, if such a value is difficult to formulate but data from a previous study are available, a normal design prior that takes the previous estimate and its standard error as the mean and standard deviation may be a reasonable way to incorporate prior knowledge and uncertainty.

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 power and sample size calculations to more specific settings not considered here, such as, binary outcomes.

Another limitation is the types of Bayes factors that we considered for the analysis, which is limited to univariate parameters with normal, normal moment, t, or point priors under the alternative.

However, after publishing an initial version of this article on the arXiv preprint server, Wong and Tendeiro (2024) have built on our results and extended the root-finding approach from Section 6 to more flexible design prior distributions and t likelihoods. Nevertheless, the approach could even be further extended to ANOVA or regression settings.

Furthermore, in the case of two-group parameters, we have assumed equal allocation of observations between the two groups, but unequal allocation may sometimes be desirable. For example, it might be easier to recruit patients for a study if they have a higher chance of receiving a new treatment than placebo, as participation will then be more attractive. In the case of (standardized) mean difference parameters, this is usually addressed by introducing an allocation ratio $r = n_2/n_1$ and changing the sample size per group n based on equal allocation to $n_1 = (2n)/(1+r)$ and $n_2 = (2n)/(1+1/r)$, see e.g., Kieser (2020, p. 21). While this adjustment has been developed for frequentist sample size determination, it is equally applicable to the Bayes factor methods considered here. However, for other parameter types, such as log hazard ratios, this is more complicated (Kieser, 2020, p.74-75).

We also did not consider sequential designs, where data are collected continuously until compelling evidence for one of the competing hypotheses is found (Wald, 1947), or the sample size is 'recalculated' after an interim analysis (see part III in Kieser, 2020). 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.

Finally, in situations where researchers have only a fixed sample size at their disposal, it may be interesting to use a 'reverse-Bayes' approach (Held et al., 2022) and determine the design prior required to achieve a desired target power. Researchers can then reason whether or not this prior is scientifically plausible and they should undertake the study based on their limited resources. This is similar to classical power analysis, where one can determine the value of the true parameter needed to achieve a given power for a given sample size, and then reason about whether or not this true parameter is realistic. In both cases, however, it is important that the focus is on reasoning about the plausibility of the design prior/parameter, and researchers should be careful not to give the impression that this choice has been made *a priori*.

Acknowledgments

We thank Angelika Stefan and František Bartoš for valuable comments on drafts of the manuscript. We thank Tsz Keung Wong for helpful suggestions related to our R package. We thank two anonymous reviewers and an associate editor for constructive comments. 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

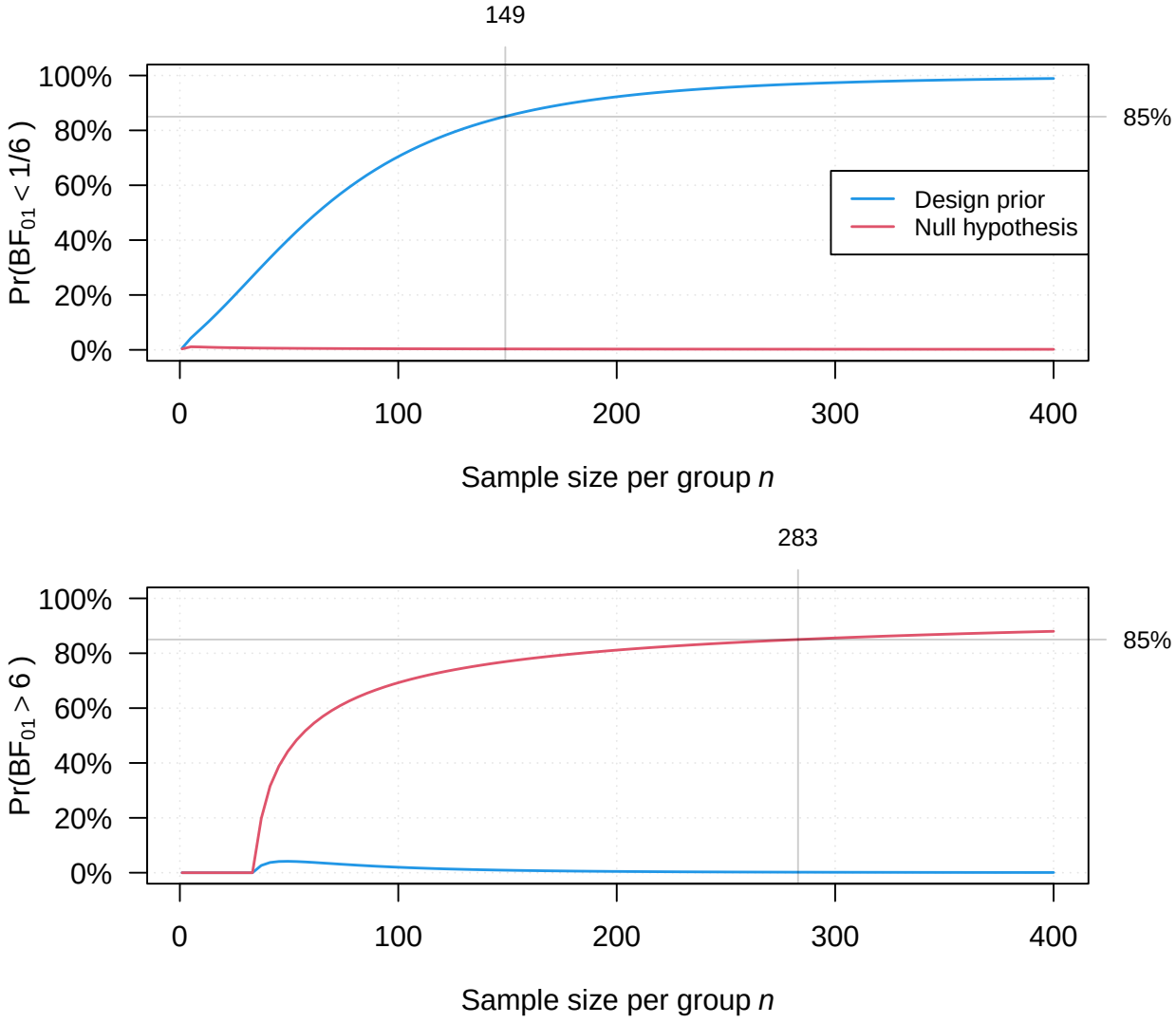
Code and data to reproduce our analyses are openly available at https://github.com/SamCH93/bfpwr. A snapshot of the repository at the time of writing is available at https://doi.org/10.5281/zenodo.12582277. We used the statistical programming language R version 4.4.1 (2024-06-14) for analyses (R Core Team, 2024) along with the `BFDA` (Schönbrodt and Stefan, 2019), `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 found on 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 (standardized) mean difference. The functions `pbf01` and `nbf01` are more general and can be used for any approximately normally distributed parameter estimate with approximate variance $\text{Var}(\hat{\theta}) = \sigma^2_{\hat{\theta}}/n$, although users have to specify the unit standard deviation $\sigma_{\hat{\theta}}$ themselves. Similar functions exist also for t-test Bayes factors (`powertbf01`) and normal moment prior Bayes factors (`powernmbf01`). The following code chunk illustrate how `powerbf01` can be used.

```
library(bfpwr)

## BF parameters
k <- 1/6 # set BF_01 threshold to 1/6
null <- 0 # set null value to 0
sd <- 1 # set standard deviation of one observation to 1
pm <- 0 # set analysis prior mean to 0
psd <- sqrt(2) # set analysis prior SD set to sqrt(2)
type <- "two.sample" # set test to two-sample

## design prior
dpm <- 0.5 # set design prior mean equal to medium SMD effect size
dpsd <- 0.1 # set positive design prior SD to incorporate parameter uncertainty

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

##
```

```
##      Two-sample z-test 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 *observations per group*
##       sd is standard deviation of one observation (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 point analysis prior ($\tau = 0$) can be rewritten as

$$\text{BF}_{01} = \exp\left[\frac{n}{\sigma_{\hat{\theta}}^2}\left\{\hat{\theta}(\theta_0 - \mu) - \frac{\theta_0^2 - \mu^2}{2}\right\}\right]. \tag{16}$$

Suppose that compelling evidence for H_1 is achieved when $\text{BF}_{01} \leq k$. 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 $\hat{\theta} \mid n, \mu_d, \tau_d \sim \text{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$, the probability of compelling evidence is given by (3).

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

$$\text{BF}_{01} = \sqrt{1 + \frac{n\tau^2}{\sigma_{\hat{\theta}}^2}}\, \exp\left(-\frac{1}{2}\left[\frac{\{\hat{\theta} - \theta_0 - \frac{\sigma_{\hat{\theta}}^2}{n\tau^2}(\theta_0 - \mu)\}^2}{\frac{\sigma_{\hat{\theta}}^2}{n}(1 + \frac{\sigma_{\hat{\theta}}^2}{n\tau^2})} - \frac{(\theta_0 - \mu)^2}{\tau^2}\right]\right). \tag{17}$$

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 $\hat{\theta} \mid n, \mu_d, \tau_d \sim \text{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$, the probability of compelling evidence is given by (6).

Appendix C Limiting power of Bayes factor with normal analysis prior

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

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

Appendix D Sample size for Bayes factor with local normal prior

Equating the power function (11) to $1 - \beta$ and applying algebraic manipulations, 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 obtain 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 sample sizes than the other and guarantees that unit information sample sizes are always larger than one.

Appendix E Simulation-based evaluation of the bfpwr package

Figure 7 shows a simulation-based evaluation of the power and sample size calculation methods for the Bayesian z-test as implemented in our `bfpwr` R package. The values for the design and analysis prior means were chosen to represent conventions for no (0), small (0.2), medium (0.5), and large (0.8) standardized mean differences (Cohen, 1992). The null and alternative hypotheses were defined as $H_0 \colon \theta = 0$ against $H_1 \colon \theta \neq 0$. The standard deviations were chosen to include point and normal priors. For each combination of analysis/design prior mean/standard deviation, the sample size to obtain a Bayes factor equal or below $k = 1/10$ with a target power of 80% was computed (shown at the top of each plot). This sample size along with the design prior was subsequently used to simulate 50'000 standardized mean difference parameter estimates based on which 50'000 Bayes factors were computed. The power was then estimated from the proportion of Bayes factors equal or below the level $k = 1/10$. Note that for certain design/analysis prior combinations, it is impossible to achieve the target power with a finite sample size. In this case an "x" is shown in the plot.

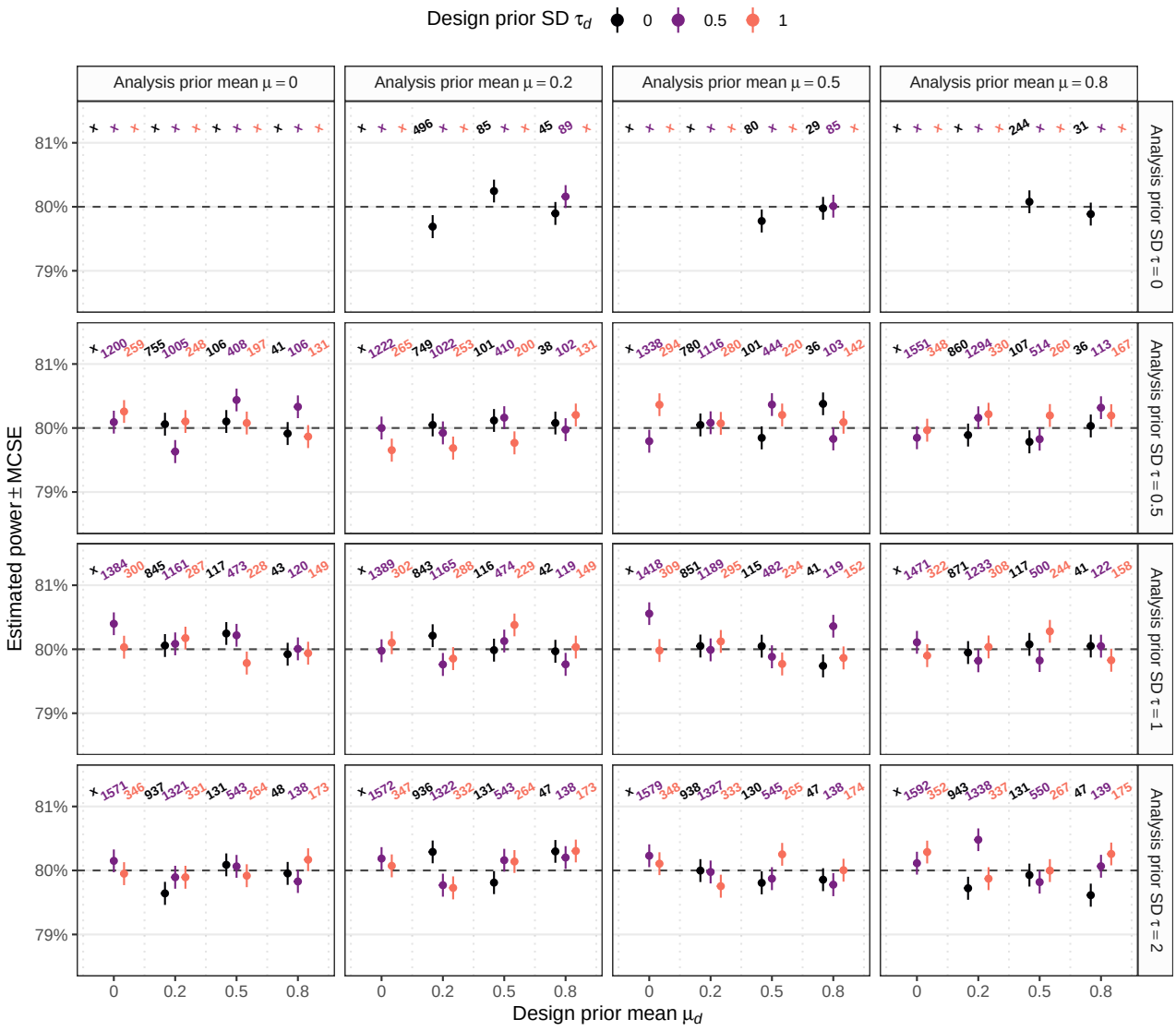


Figure 7: Simulation-based evaluation of power and sample size calculations related to the Bayesian z-test as implemented in the `bfpwr` package. The top of each plot shows the sample size to obtain a Bayes factor equal or below $k = 1/10$ with a target power of 80% for the corresponding combination of analysis and design prior (an "x" is shown if the target power is impossible to achieve for a given condition). 50'000 Bayes factors were then simulated based on this sample size, from which then the power was empirically estimated.

We can see that in all conditions, the simulation-based estimate of the power closely spreads around the target power of 80%. The maximally observed discrepancy is 0.56% while the median discrepancy is 0.14%. This suggests that the power and sample size calculation methods work as intended.

Appendix F Power with normal moment prior

Setting the Bayes factor (14) to less or equal than k and applying algebraic manipulations, we can bring the inequality into the form

$$\exp\left[1 + \frac{n(\hat{\theta} - \theta_0)^2}{\sigma_{\hat{\theta}}^2\{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)\}}\right]\left[1 + \frac{n(\hat{\theta} - \theta_0)^2}{\sigma_{\hat{\theta}}^2\{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)\}}\right] \geq \frac{\{1 + (n\tau^2)/\sigma_{\hat{\theta}}^2\}\sqrt{e}}{2k}.$$

Applying the Lambert W function on both sides, leads to

$$1 + \frac{n(\hat{\theta} - \theta_0)^2}{\sigma_{\hat{\theta}}^2\{1 + \sigma_{\hat{\theta}}^2/(n\tau^2)\}} \geq \mathrm{W}_0\left[\frac{\{1 + (n\tau^2)/\sigma_{\hat{\theta}}^2\}\sqrt{e}}{2k}\right]. \tag{18}$$

Since the argument of the Lambert W function is real and always non-negative, only the principal branch W_0 can satisfy the inequality. Assuming a $\hat{\theta} \mid n, \mu_d, \tau_d \sim \mathrm{N}(\mu_d, \tau_d^2 + \sigma_{\hat{\theta}}^2/n)$ distribution induced by a normal design prior, we can rearrange the inequality (18) and obtain the power function (15).

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. (2024). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*, 29(5):844–867. doi:10.1037/met0000520.

Banerjee, S., High, J., Stirling, S., Shepstone, L., Swart, A. M., Telling, T., Henderson, C., Ballard, C., Bentham, P., Burns, A., Farina, N., Fox, C., Francis, P., Howard, R., Knapp, M., Leroi, I., Livingston, G., Nilforooshan, R., Nurock, S., O'Brien, J., Price, A., Thomas, A. J., and Tabet, N. (2021). Study of mirtazapine for agitated behaviours in dementia (SYMBAD): a randomised, double-blind, placebo-controlled trial. *The Lancet*, 398(10310):1487–1497. doi:10.1016/s0140-6736(21)01210-1.

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.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., et al. (2017). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10. doi:10.1038/s41562-017-0189-z.

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.

Cahusac, P. M. B. (2020). *Evidence-Based Statistics: An Introduction to the Evidential Approach – from Likelihood Principle to Statistical Practice*. Wiley, Hoboken (N.J.). doi:10.1002/9781119549833.

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, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610. doi:10.1080/01621459.1982.10477856.

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.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781. doi:10.3389/fpsyg.2014.00781.

Edwards, A. W. F. (1971). *Likelihood*. Cambridge University Press, London.

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. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharmaceutical Statistics*, 15(2):96–108. doi:10.1002/pst.1736.

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

Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, 74(2):137–143. doi:10.1080/00031305.2018.1562983.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., Mulder, J., Palfi, B., Schönbrodt, F. D., Tendeiro, J. N., van den Bergh, D., Van Lissa, C. J., van Ravenzwaaij, D., Vanpaemel, W., Wagenmakers, E.-J., Williams, D. R., Zondervan-Zwijnenburg, M., and Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, 28(3):558–579. doi:10.1037/met0000454.

Held, L., Matthews, R., Ott, M., and Pawel, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, 13(3):295–314. doi:10.1002/jrsm.1538.

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, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol. 2*. Wiley, New York, second 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, New York, second edition. 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.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60(3):213–223. doi:10.1198/000313006x117837.

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.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18. doi:10.1016/j.jmp.2015.11.001.

O'Hagan, A. and Stevens, J. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230. doi:10.1177/02729890122062514.

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.

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

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.

Pramanik, S. and Johnson, V. E. (2024). Efficient alternatives for Bayesian hypothesis tests in psychology. *Psychological Methods*, 29(2):243–261. doi:10.1037/met0000482.

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/.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237. doi:10.3758/pbr.16.2.225.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172. doi:10.1214/aos/1176346785.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2018). 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., Schönbrodt, F. D., and Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, 51(3):1042–1058. doi:10.3758/s13428-018-01189-8.

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

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

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.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, 16(4):752–760. doi:10.3758/pbr.16.4.752.

Wong, T. K. and Tendeiro, J. (2024). On a generalizable approach for sample size determination in Bayesian *t* tests. doi:10.31234/osf.io/6h4d5. Preprint.

Xie, Y. (2024). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. URL https://yihui.org/knitr/. R package version 1.48.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1):585–603. doi:10.1007/bf02888369.

Computational details

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

## 2024-11-22 10:30:43.142925 Europe/Zurich

sessionInfo()

## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 24.04.1 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0
##
## 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] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] BFDA_0.5.0        ggplot2_3.5.1     dplyr_1.1.4       doParallel_1.0.17
##  [5] iterators_1.0.14  foreach_1.5.2     BayesRep_0.42.2   lamW_2.2.4
##  [9] xtable_1.8-4      bfpwr_0.1.3       knitr_1.48
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4          generics_0.1.3     stringi_1.8.4      lattice_0.22-5
##  [5] digest_0.6.35       magrittr_2.0.3     evaluate_0.24.0    grid_4.4.1
##  [9] fastmap_1.2.0       sfsmisc_1.1-18     plyr_1.8.9         Matrix_1.7-1
## [13] hypergeo_1.2-13     deSolve_1.40       gridExtra_2.3      promises_1.3.0
## [17] mgcv_1.9-1          doRNG_1.8.6        fansi_1.0.6        viridisLite_0.4.2
```

```
## [21] scales_1.3.0       truncnorm_1.0-9    abtest_1.0.1       codetools_0.2-19
## [25] cli_3.6.3          shiny_1.8.1.1      rlang_1.1.4        contfrac_1.1-12
## [29] munsell_0.5.1      splines_4.4.1      withr_3.0.1        plotrix_3.8-4
## [33] tools_4.4.1        reshape2_1.4.4     colorspace_2.1-1   httpuv_1.6.15
## [37] rngtools_1.5.2     mime_0.12          vctrs_0.6.5        R6_2.5.1
## [41] qgam_1.3.4         lifecycle_1.0.4    stringr_1.5.1      MASS_7.3-60.2
## [45] pkgconfig_2.0.3    later_1.3.2        RcppParallel_5.1.9 pillar_1.9.0
## [49] gtable_0.3.5       glue_1.7.0         Rcpp_1.0.13        TeachingDemos_2.13
## [53] highr_0.11         xfun_0.49          tibble_3.2.1       tidyselect_1.2.1
## [57] farver_2.1.2       htmltools_0.5.8.1  nlme_3.1-165       labeling_0.4.3
## [61] elliptic_1.4-0     compiler_4.4.1
```