BayesPower: A General Application of Power and Sample Size Calculation for the Bayes Factor

Tsz Keung Wong
Tilburg University
t.k.wong3004@gmail.com

Samuel Pawel
University of Zurich

Jorge Tendeiro
Hiroshima University

Abstract

BayesPower is an R package with a user-friendly Shiny interface for conducting sample size determination, power calculation, and Bayes factor calculation in the context of Bayesian hypothesis testing without the use of simulation. The app supports a wide range of commonly encountered statistical tests in the social, behavioral, and biomedical sciences, including standardized mean differences, correlations, linear regression and ANOVA, as well as tests for one and two proportions. In addition to traditional point-null hypothesis versus composite alternative hypothesis, **BayesPower** supports sample size planning for interval hypothesis Bayes factors (i.e., equivalence testing). Moreover, the Shiny app provides command-line-based R code for the reproducibility of results. For most testing problems, the sample size and power are returned almost instantly, enabling researchers to rapidly design informative and efficient studies.

Keywords: Bayes factor, power analysis, design analysis, sample size determination.

1. Introduction

The application of Bayes factors in data analysis has been facilitated by the development of accessible software, such as **BayesFactor** (Morey and Rouder 2024), **BFpack** (Mulder, Williams, Gu, Tomarken, Böing-Messing, Olsson-Collentine, Meijerink, Menke, van Aert, Fox, Hoijtink, Rosseel, Wagenmakers, and van Lissa 2021), and **JASP** (JASP Team 2025). However, to obtain informative Bayes factor inferences, experiments need to be well designed, requiring large enough samples to ensure adequate statistical power. Statistical software for power and sample size calculation in Bayes factor analysis is mainly based on Monte Carlo simulation methodology, such as **BFDA** (Stefan, Gronau, Schönbrodt, and Wagenmakers 2019). Simulation methods are associated with Monte Carlo error, can take a long time to run, and can be challenging to use for inexperienced users. Recently, an alternative approach has been

implemented in the **bfpwr** package by Pawel and Held (2025) and extended by Kelter and Pawel (2025). Following Weiss (1997) and De Santis (2004), this approach uses root-finding algorithms for sample size determination in conjunction with an analytic solution for power calculation, given a normal or a normal-moment prior for z-tests and a beta prior in binomial settings. In the settings where it is applicable, this numerical approach represents a notable improvement, as it substantially reduces computational time while avoiding Monte Carlo error. Concurrently, Wong and Tendeiro (2025) employed numerical integration for power calculation in t-tests, enabling sample size determination for Bayes factors even when the marginal likelihood functions lack a closed-form solution. Building on these recent advances, the **BayesPower** package was developed for sample size determination in common testing problems using Bayes factors.

BayesPower supports sample size determination for a variety of testing problems, including standardized mean differences, correlation, linear regression and ANOVA, as well as one- and two-proportion tests—commonly encountered in the social, behavioral, and biomedical sciences. The Bayes factors implemented in the package are primarily based on common test statistics, where the corresponding likelihood functions depends on the data only through a one-dimensional statistic (e.g., a t-statistic), ensuring computational efficiency. The procedure for sample size determination for these Bayes factors relies on numerical integration and root-finding algorithms. An exact method has also been developed and implemented to handle Bayes factors involving a likelihood function based on data in the form of a two-dimensional statistic. These implementations include Bayes factors with default and normal-moment priors for the t-test (Gronau, Ly, and Wagenmakers 2020; Pramanik and Johnson 2024); default stretched beta (Ly, Verhagen, and Wagenmakers 2016), scaled beta, and normal-moment priors for correlation; effect size and moment priors for regression and ANOVA (Klauer, Meyer-Grant, and Kellen 2024); beta and moment priors for one-proportion tests; and independent beta priors for two-proportion tests.

Most of the existing Bayes factors discussed above primarily concern testing a point-null hypothesis against a composite alternative. Nonetheless, the point-null hypothesis (e.g., an effect size of 0) is often argued to be never true (Cohen 1994; Meehl 1978). For this reason, **BayesPower** supports sample size determination for testing either a point-null hypothesis or an interval-null hypothesis (i.e., equivalence testing[1]) against a (truncated) composite alternative—except in the case of two-proportion tests, where equivalence testing is not currently supported. More specifically, the non-overlapping hypotheses Bayes factor by Morey and Rouder (2011) is implemented and extended to scenarios beyond the t-test. As the Bayes factors above are not collectively available in any existing package, **BayesPower** also allows users to compute them directly based either on test statistics or on frequencies, hence without the need of raw data. Thus, in addition to power analysis, our application allows users to calculate Bayes factors directly from reported statistical results.

The targeted audience of **BayesPower** is quantitative researchers who are somewhat familiar with Bayesian testing and wish to conduct power analysis and sample size calculations

[1]In the frequentist equivalence test, the null hypothesis is typically defined as a non-negligible effect, while the alternative hypothesis states that the effect is negligible. In contrast, we adopt the original formulation of Morey and Rouder (2011). Although the direction of the hypotheses is reversed, this does not affect the interpretation of the Bayes factor in comparison to the frequentist equivalence test. However, the interpretation of the false positive rate differs: under our specification, it represents the probability of obtaining misleading evidence in favor of the alternative hypothesis of a non-negligible effect, given that the null hypothesis of a negligible effect is true.

to ensure efficient use of limited resources and maximize the informativeness of their empirical studies. For this reason, a user-friendly Shiny app was developed and is distributed via an R package (R Core Team 2025), rather than as a traditional command-line-based R package, in order to reduce the programming burden on users and improve computational efficiency. Nonetheless, the Shiny application provides the results of sample size determination accompanied by the corresponding command-line-based R code for reproducibility. **BayesPower** is available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/package=BayesPower. The main contribution of the present paper is to report the development of the general application for sample size determination without relying on simulation for Bayesian testing, with results returned almost instantaneously for most testing problems.

The structure of the paper is as follows. Section 2 briefly introduces the Bayes factor and a general method for sample size determination. Section 3 provides technical details on the likelihood and the prior distributions for each test. Section 4 introduces the user interface and functionality of the Shiny app. Section 5 presents five examples demonstrating its use. The paper ends with some concluding remarks in Section 6.

2. The general procedure for sample size determination

2.1. Bayes Factor

The Bayes factor (Jeffreys 1961; Kass and Raftery 1995) is a relative measure of evidence for an alternative hypothesis $\mathcal{H}_1$ relative to a null hypothesis $\mathcal{H}_0$ (or vice versa when the subscript is "01"), defined as

$$\text{BF}_{10} = \frac{p(\mathcal{D} \mid \mathcal{H}_1)}{p(\mathcal{D} \mid \mathcal{H}_0)}. \tag{1}$$

Here, $p(\mathcal{D} \mid \mathcal{H}_i)$ denotes the marginal likelihood of the data $\mathcal{D}$ under hypothesis $\mathcal{H}_i$ $(i = 0, 1)$. Thus, the Bayes factor is the ratio of the marginal likelihood of the data under one hypothesis to that under another. The Bayes factor indicates the relative predictability of the observed data among two hypotheses. For instance, a BF_{10} of 7 suggests that $\mathcal{H}_1$ predicts the data 7 times better than $\mathcal{H}_0$. In other words, the data are 7 times more likely to be observed under $\mathcal{H}_1$ than $\mathcal{H}_0$. The 'factor' in the Bayes factor is due to it also being the multiplicative term transforming the prior odds $\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}$ into the corresponding posterior odds $\frac{p(\mathcal{H}_1|\mathcal{D})}{p(\mathcal{H}_0|\mathcal{D})}$ given the data (in words: prior odds times Bayes factor equals posterior odds).

Each marginal likelihood in Equation 1 is obtained by

$$p(\mathcal{D} \mid \mathcal{H}_i) = \int_{\Theta_i} p(\mathcal{D} \mid \theta)\, \pi(\theta \mid \mathcal{H}_i)\, \mathrm{d}\theta. \tag{2}$$

The first term inside the integral is the likelihood of the data given a set of parameters θ, weighted by the prior density function of θ under $\mathcal{H}_i$ over the parameter space Θ_i (the second term under the integral). In the case of a point prior, the marginal likelihood function reduces to the ordinary likelihood of the data evaluated at that parameter value. Depending on the testing problems at hand, the data $\mathcal{D}$ and the prior $\pi(\theta|\mathcal{H}_i)$ can be multidimensional. Note that Equation 2 is the density function of the so-called *prior predictive distribution*, when regarded for any potential (observable) data $\mathcal{D}$. The prior predictive distribution is

the distribution of the observable data conditional on the prior distribution $\pi(\theta)$, which encapsulates both parameter uncertainty and sampling variability.

2.2. Method for sample size determination

The purpose of sample size determination is to enable informative inferences from a study (Schönbrodt and Wagenmakers 2018), ensuring that the study has a sufficiently high probability of yielding compelling evidence. In the context of Bayes factor hypothesis testing, such evidence is typically indicated by the Bayes factor exceeding a certain threshold k (say, 3, 10, or 20). For example, the goal may be to find the minimal sample size that assures that BF_{10} is larger than k with a prespecified (high) probability if $\mathcal{H}_1$ is true, thus indicative of strong relative evidence in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$. Likewise, if $\mathcal{H}_0$ is true, we may want to determine the minimal sample size that allows finding decisive evidence for $\mathcal{H}_0$ over $\mathcal{H}_1$ by ensuring that $\text{BF}_{01} = 1/\text{BF}_{10}$ is larger than k with a prespecified probability. This search is operationalized by exploring the relationship between three ingredients: the Bayes factor threshold k, the sample size, and the subset of the sample space for which the data produce a Bayes factor that exceeds the threshold k. In this section, a general three-step procedure for Bayes factor power and sample size calculation is elaborated. Note that more detailed information on prior distributions and likelihood functions for different testing problems will be given in Section 3.

First step: the specification of analysis priors

An analysis prior under either hypothesis $\pi_a(\theta \mid \mathcal{H}_i)$ needs to be specified in order to determine the subset of the sample space leading to a compelling Bayes factor (i.e., reaching the threshold). The analysis prior consists of a prior density function (e.g., normal distribution) over a parameter space Θ_i for θ. In our implementation, a point-null hypothesis is automatically chosen for the analysis prior under the null hypothesis. Meanwhile, for interval hypothesis Bayes factors, the prior density functions are the same for both hypotheses except for the parameter space under either hypothesis, Θ_i. The parameter spaces Θ_0 and Θ_1 form a partition of the entire parameter space of θ across both hypotheses for equivalence testing.

Then, the chosen analysis priors under $\mathcal{H}_1$ and $\mathcal{H}_0$ are used for the calculation of the Bayes factor in Equation 1. By relating the computed Bayes factor to the desired threshold k, we are then able to find the subset Ω_k of the sample space Ω that yields compelling evidence, i.e., $\Omega_k = \{\mathcal{D} \in \Omega : \text{BF} \geq k\}$. The subset Ω_{k10} represents the range of $\mathcal{D}$ for which $\text{BF}_{10} \geq k$, while the subset Ω_{k01} corresponds to the range of $\mathcal{D}$ for which $\text{BF}_{01} \geq k$.

When $\mathcal{D}$ can be summarized by a one-dimensional and continuous statistic (e.g., t-statistic), a root-finding algorithm is used to find the sample subset Ω_{k10} for which the associated BF_{10} is at least equal to k (and similarly for Ω_{k01}). For discrete data, the sample subset Ω_k is determined by considering the associated Bayes factor values that are the closest to being greater than or equal to k (being exactly equal to k might not be possible due to the discreteness of the data). As for $\mathcal{D}$ when the data summary statistics is two-dimensional, Bayes factors are calculated across the entire sample space Ω via a grid search. For instance, in the case of testing two proportions, a grid with $(n_1+1)(n_2+1)$ Bayes factors is calculated. This grid is then used to identify the desired sample subsets Ω_k in much the same vein as described for the discrete data case above.

Second step: the specification of design priors

A design prior under either hypothesis $\pi_d(\theta \mid \mathcal{H}_i)$ must be specified, along with the parameter space Θ for θ, to calculate the probability of obtaining a compelling Bayes factor that exceeds a specified threshold k, i.e., $\Pr(\text{BF} > k \mid \mathcal{H}_i)$. The probability is obtained by:

$$\Pr(\text{BF} > k \mid \mathcal{H}_i) = \int_{\Theta_i} \left(\int_{\Omega_k} p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D} \right) \pi_d(\theta \mid \mathcal{H}_i) \, \mathrm{d}\theta. \tag{3}$$

As for discrete $\mathcal{D}$ when testing two proportions, the equation is:

$$\Pr(\text{BF} > k \mid \mathcal{H}_i) = \int_{\Theta_i} \left(\sum_{\Omega_k} p(\mathcal{D} \mid \theta) \right) \pi_d(\theta \mid \mathcal{H}_i) \, \mathrm{d}\theta. \tag{4}$$

The inner integral in Equation 3 and the summation in Equation 4 represent the probability of obtaining a compelling Bayes factor being greater than k under a hypothesis $\mathcal{H}_i$ given the subset Ω_k from step 1. This probability is then marginalized by the outer integral with respect to the design prior. Consequentially, the true positive rate (i.e., power) and false positive rate are obtained for a given sample size by plugging in Ω_{k10} or Ω_{k10}. More concretely,

$$\text{True positive rate} = \Pr(\text{BF}_{10} > k \mid \mathcal{H}_1) = \int_{\Theta_1} \left(\int_{\Omega_{k10}} p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D} \right) \pi_d(\theta \mid \mathcal{H}_1) \, \mathrm{d}\theta$$

$$\text{False positive rate} = \Pr(\text{BF}_{10} > k \mid \mathcal{H}_0) = \int_{\Theta_0} \left(\int_{\Omega_{k10}} p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D} \right) \pi_d(\theta \mid \mathcal{H}_0) \, \mathrm{d}\theta.$$

As for true negative and false negative rates,

$$\text{True negative rate} = \Pr(\text{BF}_{01} > k \mid \mathcal{H}_0) = \int_{\Theta_0} \left(\int_{\Omega_{k01}} p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D} \right) \pi_d(\theta \mid \mathcal{H}_0) \, \mathrm{d}\theta$$

$$\text{False negative rate} = \Pr(\text{BF}_{01} > k \mid \mathcal{H}_1) = \int_{\Theta_1} \left(\int_{\Omega_{k01}} p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D} \right) \pi_d(\theta \mid \mathcal{H}_1) \, \mathrm{d}\theta.$$

In some cases, analytical solutions are available for the computation above, as shown in Pawel and Held (2025) for the z-test. Otherwise, the R function `stats::integrate` can be used for numerical integration as in Wong and Tendeiro (2025) for the t-test.

The design prior and the analysis prior serve different purposes. The analysis prior in the previous step is used to calculate Bayes factors with the observed data, while the design prior in the current step is used to assess the probability of compelling or misleading evidence before data collection (O'Hagan and Stevens 2001; Schönbrodt and Wagenmakers 2018). Researchers may use conventional default analysis priors, together with more subjective design priors based on their knowledge (Pawel and Held 2025). Alternatively, the same prior can be specified for both design and analysis, as both reflect prior beliefs about the parameter of interest (Stefan *et al.* 2019).

In our implementation, the specification of the design and analysis priors can be different under $\mathcal{H}_1$ except for the parameter space, which is the same for both priors. This assumption avoids logical inconsistencies—such as specifying a one-sided hypothesis (e.g., $\delta > 0$) for the analysis prior while using a two-sided prior ($\delta \neq 0$) or an oppositely directed one-sided prior

($\delta < 0$) for the design prior. Moreover, the analysis and the design priors under $\mathcal{H}_0$ are always the same.

Third step: determining the sample size

The required sample size for given values of the false positive rate and power is determined using a root-finding algorithm—the `stats::uniroot` function in R. The sample size is iteratively adjusted until the condition $\Pr(\text{BF}_{10} > k \mid \mathcal{H}_1) \geq$ targeted power is satisfied for the power, based on Equation 3 or Equation 4. The function is used again for ensuring that $\Pr(\text{BF}_{10} > k \mid \mathcal{H}_0) \leq$ targeted false positive rate. Note that, unlike frequentist tests, Bayesian tests rarely achieve the exact nominal values of the specified false positive rate and power for a given sample size. In our implementation, the determined sample size ensures that the resulting false positive rate does not exceed the specified false positive rate, and that the statistical power meets or exceeds the predefined threshold. Moreover, in our implementation, users can determine the required sample size based on the desired true negative and false negative rates. An exception is testing two proportions, where the user can determine the sample size only based on either the desired power or the true negative rate, in order to ensure computational efficiency.

3. Details about the implemented statistical tests

Details about the likelihood functions and the prior distributions for applying the procedure outlined in the previous section are given in this section. Researchers primarily interested in applying **BayesPower** may skip this section and proceed directly to section 4.

3.1. Standardized mean difference

The Bayes factor for t-tests by Gronau *et al.* (2020) is implemented in our Shiny app. The marginal likelihood function is:

$$p(t \mid \mathcal{H}_i) = \int_{\Theta_i} T_v(t \mid \sqrt{n_\delta}\delta)\, \pi(\delta \mid \mathcal{H}_i)\, \mathrm{d}\delta. \tag{5}$$

T_v denotes the noncentral t-distribution with v degrees of freedom and noncentrality parameter $\sqrt{n_\delta}\delta$. The effective sample size is n_δ, in which n_δ is obtained by the number of (paired) samples n for the one-sample/paired t-test, and $(\frac{1}{n_1} + \frac{1}{n_2})^{-1}$ for the independent samples t-test. The implemented priors are the scaled t-distribution, normal distribution, and the normal-moment distribution (Pramanik and Johnson 2024).

3.2. Pearson's correlation

The exact density function of the sample correlation r is

$$p(r \mid \rho, v) = \frac{v-1}{\sqrt{2\pi}} \frac{\Gamma(v)}{\Gamma(v+\frac{1}{2})} (1-\rho^2)^{\frac{1}{2}v} (1-r^2)^{\frac{1}{2}(v-3)} (1-\rho r)^{-v+\frac{1}{2}} {}_2F_1\left(\tfrac{1}{2}, \tfrac{1}{2}, v+\tfrac{1}{2}, \frac{1+\rho r}{2}\right) \tag{6}$$

with the Gaussian hypergeometric function ${}_2F_1(.,.,.,.)$, true correlation ρ, degrees of freedom

$v = n - 1$, and n the number of observations (Hotelling 1953).[2] The prior density under consideration is given by the four-parameter beta distribution:

$$\pi(\rho \mid \alpha, \beta, a, b) = \frac{(\rho - a)^{\alpha-1}(b - \rho)^{\beta-1}}{(b - a)^{\alpha+\beta-1}B(\alpha, \beta)}, \quad \text{for } \rho \in [a, b] \tag{7}$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters, and the distribution is defined over the interval $[a, b]$ as the parameter space, which serves to scale the standard beta distribution. When $\alpha = \beta = \frac{1}{\kappa}$ with $\rho \in [-1, 1]$, the four-parameter distribution becomes the default stretched beta distribution by Ly *et al.* (2016):

$$\pi(\rho, \kappa) = \frac{2^{\frac{\kappa-2}{\kappa}}}{B\left(\frac{1}{\kappa}, \frac{1}{\kappa}\right)} \left(1 - \rho^2\right)^{\frac{1-\kappa}{\kappa}}. \tag{8}$$

Moreover, the normal-moment prior is implemented as well (Pramanik and Johnson 2024). Note that when substituting the equations above in Equation (3), the `integrate` function in R occasionally leads to error or divergence because the Γ terms in Equation 6 become ∞ and ${}_2F_1(.,.,.,.)$ returns "NaN" for large sample sizes. For this reason, the exact density of r is only used for the calculation of the Bayes factor, leading to the same values as Ly *et al.* (2016) with the default beta prior. Meanwhile, the inner integral of Equation (3) is approximated by the normal distribution with Fisher's Z transformation, which performed well in our simulation. Thus, the simulated probabilities of compelling and misleading outcomes closely align with the probabilities obtained through our numerical method. The simulation can be found in our Github repository (https://github.com/tkWong3004/BayesPower/tree/main/BF_Simulation).

3.3. Linear regression and ANOVA

Unlike the default Bayesian tests for linear regression and ANOVA in **BayesFactor** by Rouder, Morey, Speckman, and Province (2012), Klauer *et al.* (2024) reparametrize the testing problem in terms of the F-statistic and the effect size measure Cohen's f^2 (sometimes also called λ^2). Cohen's f^2 is the proportion of variance being explained by additional predictors in the full model compared to the reduced model, which is given by Cohen (1988) as:

$$f^2 = \frac{R^2_{\text{full}} - R^2_{\text{reduced}}}{1 - R^2_{\text{full}}}.$$

Subsequently, the marginal likelihood is derived as:

$$p(F \mid \mathcal{H}_i) = \int_{\Theta_i} f_F(F, q, M - q, M\lambda^2)\, \pi(\lambda^2 \mid \mathcal{H}_i)\, d\lambda^2. \tag{9}$$

The likelihood of an F-value is given by the noncentral F-distribution, see Equation A4 in Klauer *et al.* (2024):

$$f_F(F;\, \text{df}_1, \text{df}_2, \text{ncp})$$

where F denotes the observed F-statistic, $\text{df}_1 = q = k - p$, and $\text{df}_2 = M - q = N - k$. Here, N is the number of observations, k is the number of predictors in the full model, p is the

[2] It is worth noting that Hotelling (1953) uses $v = n - 1$ for the degrees of freedom instead of $v = n - 2$ as in Ly *et al.* (2016). However, the resulting Bayes factor based on density function by Hotelling (1953) aligns with the one in Ly *et al.* (2016) and **BayesFactor**.

number of predictors in the reduced model, and $M = N - p$ is the effective sample size. In other words, df_1 is the number of additional parameters in the full model compared to the reduced model and df_2 is the number of residual degrees of freedom in the full model.

Linear regression and ANOVA models are special cases of the normal linear model. Therefore, the same Bayes factor with the F-likelihood can be used for the two testing problems. The detail for the specification of k and p for different models comparison can be found in Table 1.

Table 1: Specification of p and k for different model comparisons (m denotes the number of levels of a factor). The reduced model has p parameters, and the full model has k parameters.

Model comparison	**Reduced model (p)**	**Full model (k)**
Intercept-only vs. one-factor	1	$p + (m_1 - 1)$
Intercept-only vs. two-factor	1	$p + (m_1 - 1) + (m_2 - 1)$
One-factor vs. two-factor	$1 + (m_1 - 1)$	$p + (m_2 - 1)$
Without interaction vs. with interaction	$1 + (m_1 - 1) + (m_2 - 1)$	$p + (m_1 - 1)(m_2 - 1)$

As for the prior density function $\pi(\lambda^2)$, Klauer *et al.* (2024) derived the prior density functions for the effect size prior in Equation (10) and the moment prior in Equation (11) as follows:

$$\pi(\lambda^2) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{q}{2}\right)}(\nu r^2)^{\nu/2}(\lambda^2)^{q/2-1}\left(\lambda^2 + f^2 + \nu r^2\right)^{-(\nu+q)/2} \times {}_2F_1\left(\frac{\nu+q}{4}, \frac{1}{4}(2+\nu+q); \frac{q}{2}; \frac{4f^2\lambda^2}{\left(\lambda^2 + f^2 + \nu r^2\right)^2}\right) \tag{10}$$

and

$$\pi(\lambda^2) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}\frac{2(\nu-2)}{q(\nu+q-2)f^2}\left((\xi\nu)^{-1}\lambda^2\right)^{q/2}\left(1 + (\xi\nu)^{-1}\lambda^2\right)^{-(\nu+q)/2} \tag{11}$$

where

$$\xi = \frac{(\nu+q-2)f^2}{2\nu}.$$

Both priors are mixtures of distributions, including: a normal distribution on the outcome variable, a uniform distribution on the regression coefficients of the predictors common to both models, a Jeffrey's prior on the variance, and a scaled and shifted multivariate t distribution on the regression weights for the additional predictors in the full model. For the effect size prior, the hyperparameter f^2 influences the location of the multivariate t distribution, which has ν degrees of freedom and is scaled by the hyperparameter r. In the case of the moment prior, the multivariate t distribution is centered at zero and scaled by the hyperparameter ξ.

3.4. One proportion

The number of successes y for a given sample size n, conditional on a probability of success θ, follows a binomial distribution (i.e., $y \mid \theta \sim \text{Bin}(n, \theta)$). Thus, the marginal likelihood is

obtained as:

$$p(y \mid \mathcal{H}_i) = \int_{\Theta_i} \text{Bin}(y \mid n, \theta)\, \pi(\theta \mid \mathcal{H}_i)\, \mathrm{d}\theta. \tag{12}$$

When a beta prior is specified, as in Kelter and Pawel (2025), the marginal likelihood (12) is available in closed-form as a beta-binomial distribution. In our application, beta prior and normal-moment prior are implemented.

3.5. Two proportions with independent beta priors

Similarly, let y_1 and y_2 denote the number of successes in group 1 and group 2, respectively, with sample sizes n_1 and n_2 and success probabilities θ_1 and θ_2. The number of successes are assumed to follow independent binomial distributions:

$$y_1 \mid \theta_1 \sim \text{Bin}(n_1, \theta_1), \qquad y_2 \mid \theta_2 \sim \text{Bin}(n_2, \theta_2).$$

Subsequently, the marginal likelihood is:

$$p(y_1, y_2 \mid \mathcal{H}_i) = \int_0^1 \int_0^1 \text{Bin}(y_1 \mid \theta_1)\, \text{Bin}(y_2 \mid \theta_2)\, \pi(\theta_1 \mid \mathcal{H}_i)\, \pi(\theta_2 \mid \mathcal{H}_i)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2. \tag{13}$$

In our package, the independent beta approach is implemented to test the equality of two proportions. Formally, the hypotheses are specified as

$$\mathcal{H}_0 \colon \theta_1 = \theta_2 = \theta, \qquad \mathcal{H}_1 \colon \theta_1 \neq \theta_2.$$

Under the null hypothesis, the two probabilities of success are assumed to be equal and are assigned a beta prior distribution on θ (i.e., $\theta \sim \text{Beta}(a_0, b_0)$). The analytical solution for the marginal likelihood is derived as (see Appendix A):

$$p(y_1, y_2 \mid \mathcal{H}_0) = \frac{\binom{n_1}{y_1}\binom{n_2}{y_2}\text{B}(a_0 + y_1 + y_2, b_0 + n_1 + n_2 - y_1 - y_2)}{\text{B}(a_0, b_0)}$$

where $\text{B}(z_1, z_2)$ is the beta function.

Under the alternative hypothesis, each probability of success per group is assigned a beta prior distribution (i.e., $\theta_1 \sim \text{Beta}(a_1, b_1)$ and $\theta_2 \sim \text{Beta}(a_2, b_2)$). The marginal likelihood is derived as (see Appendix A):

$$p(y_1, y_2 \mid \mathcal{H}_1) = \frac{\binom{n_1}{y_1}\text{B}(a_1 + y_1, b_1 + n_1 - y_1)}{\text{B}(a_1, b_1)} \, \frac{\binom{n_2}{y_2}\text{B}(a_2 + y_2, b_2 + n_2 - y_2)}{\text{B}(a_2, b_2)}.$$

Taking the ratio of the two marginal likelihoods produces the Bayes factor (Equation 1). The equations are used for the calculation of the probability of compelling and misleading evidence as well (Equation 4).

4. BayesPower

BayesPower can be downloaded and opened using the following commands in R:

```
R>install.packages("BayesPower")
R>BayesPower::BayesPower_BayesFactor()
```

The source code for the package and the simulation scripts used to validate our method are available on our GitHub repository: https://github.com/tkWong3004/BayesPower. The package is also archived on Zenodo with the DOI https://doi.org/10.5281/zenodo.17405100.

Figure 1 shows a screenshot of the application. Users can select the testing problem via the navigation bar at the top. **BayesPower** supports the following statistical tests:

- Standardized mean difference: one-sample/paired t-test and independent samples t-test
- Pearson's correlation
- Linear regression/ANOVA
- Proportion: one proportion and two proportion tests

The sidebar panel on the left allows users to choose the application mode under "Select Mode", and to specify the relevant inputs. The results of the power analysis are displayed in the main panel on the right.

The "Sample size determination" mode, as its name suggests, is used to determine the required sample size given the specified conditions. In this mode, users first specify the parameter space Θ under both the null hypothesis $\mathcal{H}_0$ and alternative hypothesis $\mathcal{H}_1$. Next, the analysis prior is defined under "Analysis Prior Distribution". If the design prior and analysis prior differ, additional hyperparameters of the design prior (e.g., location and scale) must be specified. Then, users set the minimal acceptable true positive (or true negative) rate and the maximal probability of false positive (or false negative). After defining the bounds of compelling evidence and clicking the "Run" button, the results will be displayed in the main panel. The top left panel displays the chosen analysis and design priors, whereas the top right panel shows the probability of obtaining compelling evidence under the two hypotheses and the required sample size. Once the results are displayed in the main panel, users can download them as a HTML file with a timestamp by clicking the "Download result as HTML" button at the bottom of the sidebar panel. The R code for reproducing the results is provided as well. Users are strongly advised to always start with the Shiny app first before using command-line based functions in R as the package was mainly developed for the usage of the Shiny app.

Additional plots—such as the power curve and the relationship between Bayes factors and data—can be enabled by checking the corresponding checkboxes. These plots are intended for advanced users and are computationally intensive, potentially taking a few seconds to generate. Note that the procedure is particularly computationally intensive when the testing problem involves two proportions.

For the power curve, the left panel displays the true positive rate (black line) and the false positive rate (gray line) across varying sample sizes, whereas the right panel presents the true negative rate (black line) and the false negative rate (gray line).

For the plots illustrating the relationship between Bayes factors and the data, the y-axis represents the natural logarithm of the Bayes factor, and the x-axis corresponds to the observed test statistic (i.e., the t-value and the f-value). The left panel shows the relationship when the Bayes factor indicates evidence in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$, with the title indicating the minimal value of the test statistic that yields a compelling Bayes factor, $\text{BF}_{10} > k$. Similarly, the right panel illustrates the relationship when the Bayes factor supports $\mathcal{H}_0$ over $\mathcal{H}_1$, with the title indicating the minimal value of the test statistic for which $\text{BF}_{01} > k$. These calculations are

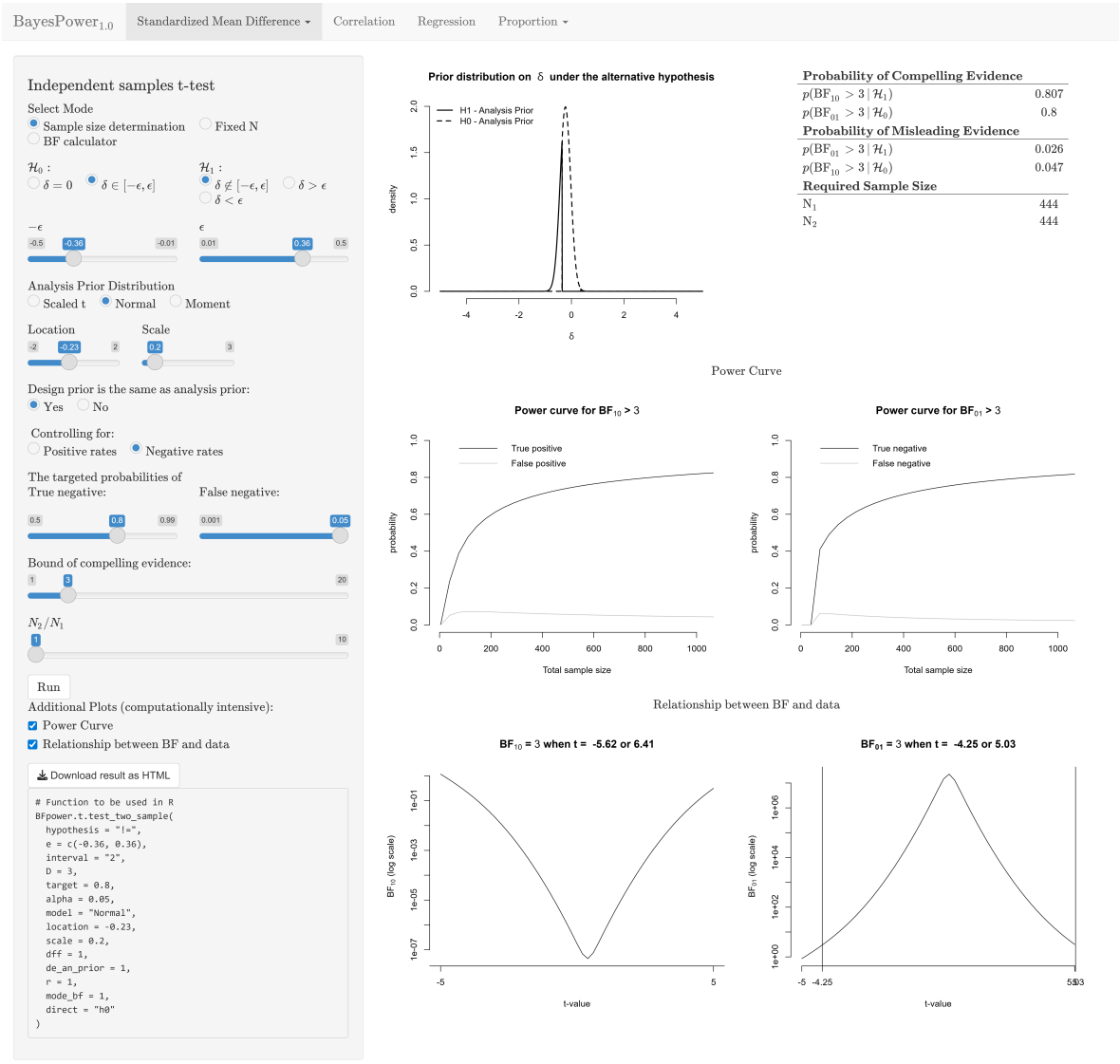

Figure 1: **BayesPower** (*Standardized Mean Difference* tab).

based on the analysis priors under both hypotheses, the specified threshold for compelling evidence, and the sample sizes provided in the table in the top-right corner. In the case of testing two proportions, a heatmap is provided over the grid of outcomes for groups 1 and 2, indicating where the evidence is compelling for either $\mathcal{H}_1$ or $\mathcal{H}_0$.

The "Fixed N" mode is used to calculate the probabilities of compelling and misleading evidence of the test for a given sample size. Similar to the "Sample Size Determination" mode, users must specify the parameter space as well as the analysis and design priors. However, instead of specifying the desired probability of compelling and misleading evidence, users provide the sample size directly. The generated table provides the estimated true/false

positive rates and true/false negative rates.

The "BF calculator" mode is used to compute the Bayes factor based on the parameter space under $\mathcal{H}_0$, $\mathcal{H}_1$, analysis prior, observed data (in the form of summary statistics), and sample size. The resulting Bayes factor is displayed at the bottom of the sidebar panel.

5. Application

We present five applications to illustrate the use of **BayesPower** in different testing problems. Readers are encouraged to explore the app using these applications as running examples.

5.1. Standardized mean difference

We made use of an example from Moon and Roeder (2014), which is also discussed in Lakens, Scheel, and Isager (2018) in the context of equivalence testing. Moon and Roeder (2014) investigated whether Asian American women would perform better than a control group on a mathematics exam when primed with their Asian identity. The descriptive statistics for the primed sample were $\overline{x}_1 = .46$, $s_1 = .17$, and $n_1 = 53$, and for the control group $\overline{x}_2 = .50$, $s_2 = .18$, and $n_2 = 48$. The null hypothesis posits a negligible effect, whereas the alternative hypothesis posits that the difference between the experimental group and the control group is non-negligible:

$$\mathcal{H}_0 : \delta \in [-0.36, 0.36] \quad \text{versus} \quad \mathcal{H}_1 : \delta \notin [-0.36, 0.36],$$

where the parameter space in the brackets indicates an interval hypothesis for equivalence test. Following Lakens *et al.* (2018), an equivalence bound of 0.0625 in raw score units was specified. When expressed in terms of standardized mean difference, the equivalence bounds correspond to $\frac{0.0625}{s_p} \approx 0.36$. An independent samples t-test was conducted (equal variances assumed), yielding $t(99) = -1.148$, $p = 0.254$ and $d = -0.23$. Moreover, the result of equivalence test in Lakens *et al.* (2018) is not statistically equivalent and not statistically different. However, using our application, the interval Bayes factor was computed as $\text{BF}_{01} = 9.05$ with the analysis prior being a two-sided scaled t-distribution with location 0, scaling parameter 0.707, and 1 degree of freedom. Therefore, the Bayes factor suggest that the observed data are 9.05 times more likely under the null hypothesis $\mathcal{H}_0$ of negligible effect than under the alternative hypothesis $\mathcal{H}_1$ of non-negligible effect.

Suppose a possible future study is conducted and the **BayesPower** application would be used for sample size determination. Based on the existing study with $d = -0.23$, the parameter space is specified as follows:

$$\mathcal{H}_0 : \delta \in [-0.36, 0.36] \quad \text{versus} \quad \mathcal{H}_1 : \delta \notin [-0.36, 0.36]$$

The analysis and design priors are specified as a two-sided normal prior with a location of -0.23, a scaling parameter of .2. The standard error is given by $s_d = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}}$, an equation commonly used in meta-analysis to obtain the standard error of the standardized mean difference effect size estimate $\sigma_{\hat{\theta}}$ (Cooper, Hedges, and Valentine 2019). Given a bound for compelling evidence of 3, a targeted true negative value of 0.8, the greatest acceptable false negative rate of 0.05 and equal sample size per group, the required sample size is 444

per group, see Figure 1. The actual true negative rate and false negative rate are .8 and .047, respectively.

5.2. Pearson's correlation

Ly *et al.* (2016) conducted a Bayesian analysis using the data from Stulp, Buunk, Verhulst, and Pollet (2013) to investigate whether height is associated with receiving a higher popular vote in the U.S. presidential election. The correlation is 0.393 with $n = 46$, between the relative heights of U.S. presidents (compared to their most successful opponents) and the ratio of popular votes—that is, the percentage of popular votes for the president divided by the sum of the percentages of popular votes for both the president and their most popular opponent (Stulp *et al.* 2013, p. 162). Ly *et al.* (2016) reported $\text{BF}_{10} = 6.33$ with a two-sided default beta prior and $k = 1$, **BayesPower** and **BayesFactor** returned the values of Bayes factor within rounding error.

As for sample size determination in a possible future study, the hypotheses are:

$$\mathcal{H}_0 : \rho = 0 \quad \text{versus} \quad \mathcal{H}_1 : \rho \sim \text{Beta}(\alpha = 1, \beta = 1), \quad \rho \in [0, 1].$$

The one-sided test is motivated by the expectation discussed in Stulp *et al.* (2013), which states: "According to conventional wisdom, U.S. presidential elections are often won by the taller of the two candidates" (Stulp *et al.* 2013, p. 159). For this reason, a right-sided beta prior with $\alpha = 1$ and $\beta = 1$ is specified as an analysis prior. However, a point design prior with $\rho_1 = 0.3$ is selected, following Cohen's benchmark for a "medium effect size". Thus, the design prior under the alternative hypothesis assumed that the true correlation is 0.3. Based on this, the required sample size is determined to be 104 in order to achieve a targeted power of at least 0.8 with a maximum acceptable false positive rate of 0.05 and bound of compelling evidence $k = 3$, see Figure 2. Therefore, we would need to wait approximately 230 years to gather enough data—assuming there will eventually be an 104th U.S. president.

5.3. Regression/ANOVA

Mulder *et al.* (2021) conducted a Bayesian ANOVA on a dataset from Janiszewski and Uy (2008) investigating the anchoring effect. A 2×2 factorial experiment was conducted, in which participants were asked to estimate the price of a television given an anchor price. In the first condition, participants were presented with either a round anchor (e.g., \$5,000) or a precise anchor (e.g., \$4,988 or \$5,012). The second condition manipulated whether participants were told that the actual price was close to the anchor price. The outcome variable was the standardized difference between the participants' guess on the price and the anchor price.

For illustrative purposes, suppose the hypothesis of interest is whether there is an interaction effect between the two factors. To test this, a two-factor model without interaction terms is compared to a two-factor model with an interaction term. Thus,

$$\mathcal{H}_0 : \lambda^2 = 0 \quad \text{versus} \quad \mathcal{H}_1 : \lambda^2 > 0$$

where λ^2 is the standardized noncentrality parameter of the F-distribution when comparing a full with a reduced model. An effect size prior of $f^2 = 0.01$ (representing a small effect size) is specified, with $r = 0.18$ and $v = 3$ following the recommendations by Klauer *et al.* (2024) as the analysis prior. Meanwhile, a point design prior located at 0.01 is specified. The

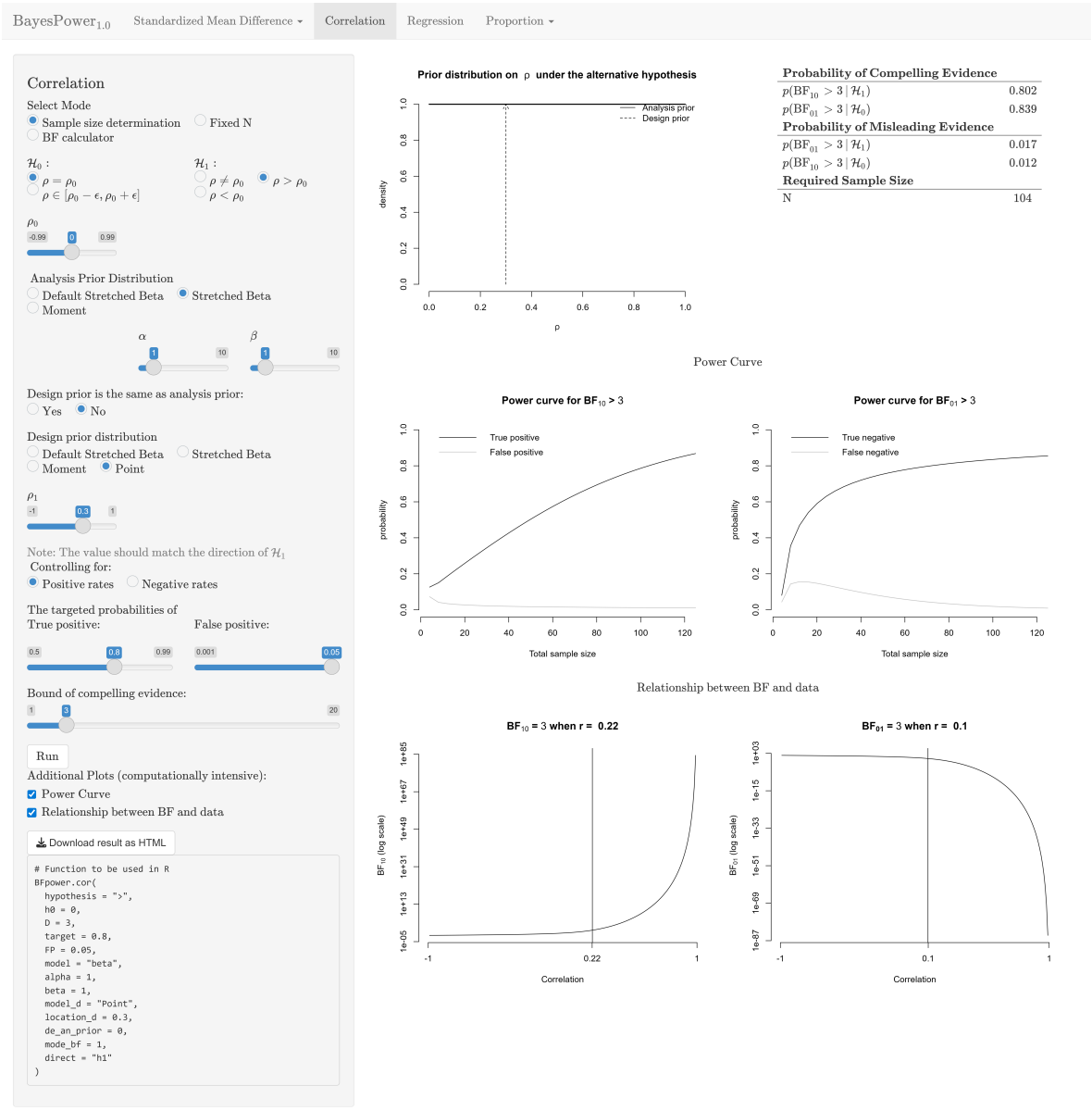

Figure 2: **BayesPower** (*Correlation* tab).

required sample size is determined to be 1155 in order to achieve a target power of 0.8 with $\alpha = 0.05$ with the bound of evidence being 3. The actual power and false positive rate are 0.8 and 0.024, respectively, see Figure 3.

5.4. One proportion

In Van Doorn, Matzke, and Wagenmakers (2019), a beer tasting experiment was conducted at the University of Amsterdam. Participants were given two small cups filled with Weihenstephaner Hefeweissbier—one containing alcohol and the other alcohol-free—and were asked

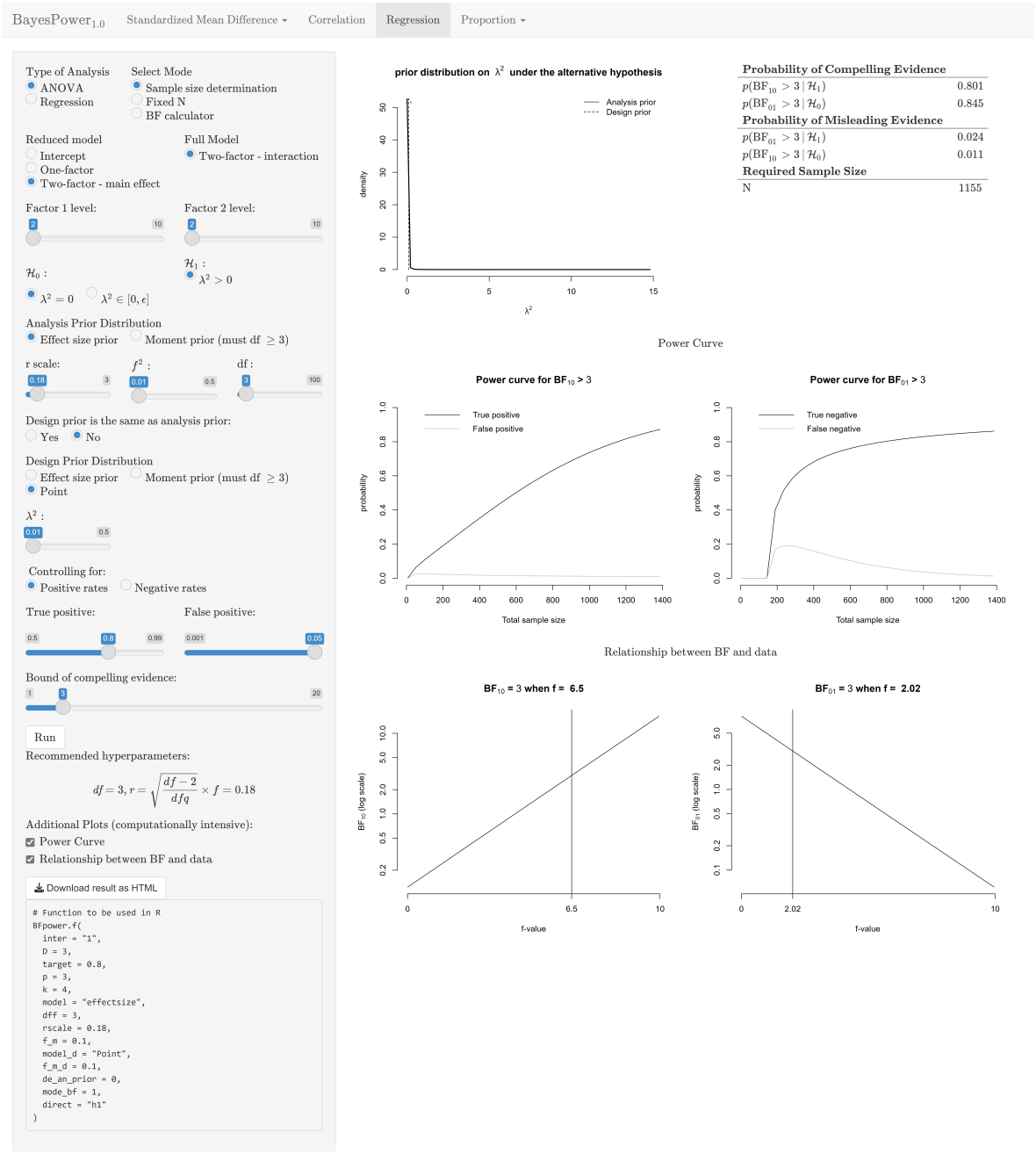

Figure 3: **BayesPower** (*Regression* tab).

to identify which one contained alcohol. The hypotheses for the experiment are:

$$\mathcal{H}_0 : \theta = 0.5 \quad \text{versus} \quad \mathcal{H}_1 : \theta \sim \text{Beta}(\alpha = 1, \beta = 1), \quad \theta \in [0.5, 1].$$

The null hypothesis states that participants are guessing at random, implying no ability to distinguish between the beers. The alternative hypothesis assumes that participants possess some ability to discriminate between alcoholic and non-alcoholic beer. A total of 42 out of 52

participants correctly distinguished the beer that contained alcohol. **JASP** and **BayesPower** return $BF_{10} = 10742.5$, indicating decisive evidence for discriminative ability over random guessing.

Suppose a sample size determination had been conducted prior to the experiment, specifying a right-sided analysis and a design prior of $\text{Beta}(\alpha = 1, \beta = 1)$, targeting a power of 0.8 and a false positive rate of 0.05. Under these specifications, a sample size of 140 would be required, resulting in an actual power of 0.808 and an effective α level of 0.011.

5.5. Two proportions

Magee, Von Dadelszen, Rey, Ross, Asztalos, Murphy, Menzies, Sanchez, Singer, Gafni, Gruslin, Helewa, Hutton, Lee, Lee, Logan, Ganzevoort, Welch, Thornton, and Moutquin (2015) conducted a randomized controlled trial to examine the impact of less-tight versus tight control of hypertension on pregnancy complications. Hypertension was defined as a diastolic blood pressure (DBP) of 90 mmHg or higher. Women with hypertension were randomized to either a *less-tight control* group, where the target DBP was 100 mmHg, or a *tight control* group, where the target DBP was 85 mmHg. The outcome variable, pregnancy complications, was a composite of pregnancy loss or the need of high-level neonatal care. Among 493 pregnancies assigned to the less-tight control group, 155 (31.4%) resulted in pregnancy loss or required high-level neonatal care. In comparison, 150 out of 488 pregnancies (30.7%) in the tight control group experienced the same outcome.

The hypotheses are as follows:

$$\begin{aligned} \mathcal{H}_0: &\quad p_1 = p_2 = p_0 \sim \text{Beta}(\alpha_0 = 1, \beta_0 = 1) \\ \mathcal{H}_1: &\quad p_1 \sim \text{Beta}(\alpha_1 = 1, \beta_1 = 1), \quad p_2 \sim \text{Beta}(\alpha_2 = 1, \beta_2 = 1) \end{aligned}$$

where p denotes the proportion of pregnancy complications. The calculated Bayes factors using **BayesPower** with the analysis priors above and **BayesFactor** are the same with $BF_{01} = 13.15$.

As for the sample size determination for a possible future study using the existing data, the analysis priors under each hypothesis are specified based on the observed data:

$$\begin{aligned} \mathcal{H}_0: &\quad p_1 = p_2 = p_0 \sim \text{Beta}(\alpha_0 = 1, \beta_0 = 1) \\ \mathcal{H}_1: &\quad p_1 \sim \text{Beta}(\alpha_1 = 156, \beta_1 = 339), \quad p_2 \sim \text{Beta}(\alpha_2 = 151, \beta_2 = 339). \end{aligned}$$

We further assume that the analysis and design priors are the same. The required sample size per group is determined to be 37 for achieving a power of 0.8 with the bound of compelling evidence being 3. The actual power is 0.82 with a false positive rate of .15 due to the discreteness of the data.

6. Discussion

To facilitate sample size determination in Bayesian hypothesis testing, a Shiny app has been developed for common testing scenarios and is distributed via an R package. A key advantage of **BayesPower** is the use of numerical methods instead of simulation, which ensures computational efficiency and reproducibility. This approach allows results to be returned almost

instantaneously for Bayes factors involving likelihood functions with one outcome variable, such as in t-tests, correlation tests, regression/ANOVA, and binomial tests. As for testing two proportions, the computational time ranges from a few seconds to 10 seconds or more depending on the true required sample size.

As for the limitations, **BayesPower** can only determine sample sizes up to 10,000 for t-tests, regression/ANOVA, and binomial tests, and up to 5,000 for correlation and two-proportion tests, due to computational efficiency considerations. Moreover, in the two-proportion setting, the application does not permit simultaneous control of both positive rates or negative rates, also for efficiency reasons. Consequently, the determined sample size may yield a higher false positive or false negative rate than the conventional levels of .05 or .2, even when the targeted power (or true negative rate) is achieved as in Section 5.5. In this example, a sample size of 5,150 per group is required for achieving the conventional false positive rate of 0.05 with the actual power of 0.95, which are obtained under the "Fixed N" mode. Users are therefore advised to review the output carefully to ensure that error rate is acceptable for their specific purposes.

Another limitation of the **BayesPower** package is its focus on fixed-N designs, where the sample size is determined before data collection. For sequential designs, readers are referred to the **BFDA** package (Stefan *et al.* 2019). Future work may focus on generalizing the methods in **BayesPower** to sequential settings.

For any problem and suggestion concerning **BayesPower**, users are encouraged to report it on https://github.com/tkWong3004/BayesPower/issues.

Author's Note

The order of the second and third authors was determined by the following code:

```
R> set.seed(1517192324)
R> sample(c("Samuel", "Jorge"), size = 1)
```

The name returned by the function was designated as the second author, while the other became the third. The seed was based on the outcome of the 'Mark Six' lottery organized by the Hong Kong Jockey Club during the Mid-Autumn Festival on October 6, 2025 (Result ID 25/107). In this lottery, six main balls and one extra ball are drawn from a pool of 49 without replacement. To construct the seed, the six main balls are first arranged in ascending order, after which the extra (seventh) ball is appended at the end. This method was agreed upon one week before the lottery draw. However, on the day the results were announced, it was discovered that the seed could not be set using the predetermined method because its value was too large. To resolve this issue, numbers were removed from the rightmost lottery balls one at a time until a valid seed was obtained. Ultimately, only the first five main balls were used to determine the final seed. This deviation from the "preregistration" was agreed among the authors.

A. Two-sample binomial Bayes factor

Let $Y_1 \mid \theta_1 \sim \text{Bin}(n_1, \theta_1)$ and $Y_2 \mid \theta_2 \sim \text{Bin}(n_2, \theta_2)$. Consider the null hypothesis $\mathcal{H}_0 \colon \theta_1 = \theta_2 = \theta$ versus the alternative hypothesis $\mathcal{H}_1 \colon \theta_1 \neq \theta_2$ with a beta prior $\theta \mid \mathcal{H}_0 \sim \text{Beta}(a_0, b_0)$

assigned to the common probability θ under $\mathcal{H}_0$, and two independent beta priors $\theta_1 \mid \mathcal{H}_1 \sim \text{Beta}(a_1, b_1)$ and $\theta_2 \mid \mathcal{H}_1 \sim \text{Beta}(a_2, b_2)$ assigned to the group-specific probabilities θ_1 and θ_2 under $\mathcal{H}_1$.

The marginal likelihood of two observed number of success y_1 and y_2 under the null hypothesis is then given by

$$\begin{aligned}
p(y_1, y_2 \mid \mathcal{H}_0) &= \int_0^1 p(y_1 \mid \theta)\, p(y_2 \mid \theta)\, \pi(\theta \mid \mathcal{H}_0)\, \mathrm{d}\theta \\
&= \int_0^1 \binom{n_1}{y_1} \theta^{y_1} (1-\theta)^{n_1 - y_1} \binom{n_2}{y_2} \theta^{y_2} (1-\theta)^{n_2 - y_2}\, \theta^{a_0 - 1} (1-\theta)^{b_0 - 1} \frac{1}{\mathrm{B}(a_0, b_0)}\, \mathrm{d}\theta \\
&= \frac{\binom{n_1}{y_1}\binom{n_2}{y_2}}{\mathrm{B}(a_0, b_0)} \int_0^1 \theta^{y_1 + y_2 + a_0 - 1} (1-\theta)^{n_1 + n_2 + b_0 - y_1 - y_2 - 1}\, \mathrm{d}\theta \\
&= \frac{\binom{n_1}{y_1}\binom{n_2}{y_2} \mathrm{B}(a_0 + y_1 + y_2, b_0 + n_1 + n_2 - y_1 - y_2)}{\mathrm{B}(a_0, b_0)},
\end{aligned}$$

where $\mathrm{B}(z_1, z_2) = \int_0^1 t^{z_1 - 1} (1-t)^{z_2 - 1}\, \mathrm{d}t$ is the beta function. In a similar way, the marginal likelihood of y_1 and y_2 under the alternative hypothesis can be derived to be

$$\begin{aligned}
p(y_1, y_2 \mid \mathcal{H}_1) &= \int_0^1 \int_0^1 p(y_1 \mid \theta_1)\, p(y_2 \mid \theta_2)\, \pi(\theta_1 \mid \mathcal{H}_1)\, \pi(\theta_2 \mid \mathcal{H}_1) \mathrm{d}\theta_1\, \mathrm{d}\theta_2 \\
&= \frac{\binom{n_1}{y_1} \mathrm{B}(a_1 + y_1, b_1 + n_1 - y_1)}{\mathrm{B}(a_1, b_1)} \frac{\binom{n_2}{y_2} \mathrm{B}(a_2 + y_2, b_2 + n_2 - y_2)}{\mathrm{B}(a_2, b_2)}.
\end{aligned}$$

Taking the ratio of the two marginal likelihoods produces the Bayes factor.

References

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. ISBN 9780121790608.

Cohen J (1994). "The earth is round (p < .05)." *American Psychologist*, **49**(12), 997–1003. doi:10.1037/0003-066x.49.12.997. URL https://doi.org/10.1037/0003-066x.49.12.997.

Cooper H, Hedges LV, Valentine JC (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. ISBN 9780871540058. doi:10.7758/9781610448864. URL http://dx.doi.org/10.7758/9781610448864.

De Santis F (2004). "Statistical evidence and sample size determination for Bayesian hypothesis testing." *Journal of Statistical Planning and Inference*, **124**(1), 121–144. ISSN 0378-3758. doi:https://doi.org/10.1016/S0378-3758(03)00198-8.

Gronau QF, Ly A, Wagenmakers EJ (2020). "Informed Bayesian t-Tests." *The American Statistician*, **74**(2), 137–143. doi:10.1080/00031305.2018.1562983.

Hotelling H (1953). "New Light on the Correlation Coefficient and its Transforms." *Journal of the Royal Statistical Society: Series B (Methodological)*, **15**(2), 193–225. doi:10.1111/j.2517-6161.1953.tb00135.x.

Janiszewski C, Uy D (2008). "Precision of the anchor influences the amount of adjustment." *Psychological Science*, **19**(2), 121–127.

JASP Team (2025). "JASP (Version 0.19.3)[Computer software]." URL https://jasp-stats.org/.

Jeffreys H (1961). *Theory of Probability*. third edition. Clarendon Press, Oxford.

Kass RE, Raftery AE (1995). "Bayes Factors." *Journal of the American Statistical Association*, **90**(430), 773–795. doi:10.1080/01621459.1995.10476572.

Kelter R, Pawel S (2025). "Bayesian Power and Sample Size Calculations for Bayes Factors in the Binomial Setting." 2502.02914, URL https://arxiv.org/abs/2502.02914.

Klauer KC, Meyer-Grant CG, Kellen D (2024). "On Bayes factors for hypothesis tests." *Psychonomic Bulletin & Review*, pp. 1–25.

Lakens D, Scheel AM, Isager PM (2018). "Equivalence Testing for Psychological Research: A tutorial." *Advances in Methods and Practices in Psychological Science*, **1**(2), 259–269. doi:10.1177/2515245918770963. URL https://doi.org/10.1177/2515245918770963.

Ly A, Verhagen J, Wagenmakers EJ (2016). "Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology." *Journal of Mathematical Psychology*, **72**, 19–32. ISSN 0022-2496. doi:https://doi.org/10.1016/j.jmp.2015.06.004. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments.

Magee LA, Von Dadelszen P, Rey E, Ross S, Asztalos E, Murphy KE, Menzies J, Sanchez J, Singer J, Gafni A, Gruslin A, Helewa M, Hutton E, Lee SK, Lee T, Logan AG, Ganzevoort W, Welch R, Thornton JG, Moutquin JM (2015). "Less-Tight versus Tight Control of Hypertension in Pregnancy." *New England Journal of Medicine*, **372**(5), 407–417. doi:10.1056/nejmoa1404595.

Meehl PE (1978). "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology." *Journal of Consulting and Clinical Psychology*, **46**(4), 806–834. doi:10.1037/0022-006x.46.4.806. URL https://doi.org/10.1037/0022-006x.46.4.806.

Moon A, Roeder SS (2014). "A Secondary Replication Attempt of Stereotype Susceptibility A Secondary Replication Attempt of Stereotype Susceptibility." doi:https://doi.org/10.1027/1864-9335/a000193.

Morey RD, Rouder JN (2011). "Bayes factor approaches for testing interval null hypotheses." *Psychological methods*, **16**(4), 406.

Morey RD, Rouder JN (2024). *BayesFactor: Computation of Bayes Factors for Common Designs*. doi:10.32614/CRAN.package.BayesFactor. R package version 0.9.12-4.7, URL https://CRAN.R-project.org/package=BayesFactor.

Mulder J, Williams DR, Gu X, Tomarken A, Böing-Messing F, Olsson-Collentine A, Meijerink M, Menke J, van Aert R, Fox JP, Hoijtink H, Rosseel Y, Wagenmakers EJ, van Lissa C

(2021). "BFpack: Flexible Bayes Factor Testing of Scientific Theories in R." *Journal of Statistical Software*, **100**(18), 1–63. doi:10.18637/jss.v100.i18.

O'Hagan A, Stevens JW (2001). "Bayesian Assessment of sample size for Clinical Trials of Cost-Effectiveness." *Medical Decision Making*, **21**(3), 219–230. doi:10.1177/0272989x0102100307.

Pawel S, Held L (2025). "Closed-Form Power and Sample Size Calculations for Bayes Factors." *The American Statistician*, **0**(0), 1–15. doi:10.1080/00031305.2025.2467919.

Pramanik S, Johnson VE (2024). "Efficient alternatives for Bayesian hypothesis tests in psychology." *Psychological methods*, **29**(2), 243.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rouder JN, Morey RD, Speckman PL, Province JM (2012). "Default Bayes factors for ANOVA designs." *Journal of Mathematical Psychology*, **56**(5), 356–374. ISSN 0022-2496. doi:https://doi.org/10.1016/j.jmp.2012.08.001.

Schönbrodt FD, Wagenmakers EJ (2018). "Bayes factor design analysis: Planning for compelling evidence." *Psychonomic Bulletin & Review*, **25**(1), 128–142. doi:10.3758/s13423-017-1230-y.

Stefan AM, Gronau QF, Schönbrodt FD, Wagenmakers EJ (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. URL https://doi.org/10.3758/s13428-018-01189-8.

Stulp G, Buunk AP, Verhulst S, Pollet TV (2013). "Tall claims? Sense and nonsense about the importance of height of US presidents." *The Leadership Quarterly*, **24**(1), 159–171. ISSN 1048-9843. doi:https://doi.org/10.1016/j.leaqua.2012.09.002.

Van Doorn J, Matzke D, Wagenmakers EJ (2019). "An In-Class Demonstration of Bayesian inference." *Psychology Learning and Teaching*, **19**(1), 36–45. doi:10.1177/1475725719848574.

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:https://doi.org/10.1111/1467-9884.00075.

Wong TK, Tendeiro JN (2025). "On a generalizable approach for sample size determination in Bayesian t tests." *Behavior Research Methods*, **57**(5), 130. doi:10.3758/s13428-025-02654-x.

Affiliation:

Tsz Keung Wong
Department of Methodology and Statistics
Tilburg School of Social and Behavioral Sciences
Tilburg University
Warandelaan 2
5037AB Tilburg
The Netherlands
E-mail: t.k.wong3004@gmail.com