multibridge: An R Package To Evaluate Informed Hypotheses in Binomial and Multinomial

Models

Alexandra Sarafoglou, Frederik Aust, Maarten Marsman, Eric-Jan Wagenmakers, & Julia M.

Haaf

University of Amsterdam

Author Note

Abstract

The **multibridge** R package allows a Bayesian evaluation of informed hypotheses $\mathcal{H}_r$ applied to frequency data from an independent binomial or multinomial distribution. **multibridge** uses bridge sampling to efficiently compute Bayes factors for the following hypotheses concerning the latent category proportions $\boldsymbol{\theta}$: (a) hypotheses that postulate equality constraints (e.g., $\theta_1 = \theta_2 = \theta_3$); (b) hypotheses that postulate inequality constraints (e.g., $\theta_1 < \theta_2 < \theta_3$ or $\theta_1 > \theta_2 > \theta_3$); (c) hypotheses that postulate mixtures of inequality constraints and equality constraints (e.g., $\theta_1 < \theta_2 = \theta_3$); and (d) hypotheses that postulate mixtures of (a)–(c) (e.g., $\theta_1 < (\theta_2 = \theta_3), \theta_4$). Any informed hypothesis $\mathcal{H}_r$ may be compared against the encompassing hypothesis $\mathcal{H}_e$ that all category proportions vary freely, or against the null hypothesis $\mathcal{H}_0$ that all category proportions are equal. **multibridge** facilitates the fast and accurate comparison of large models with many constraints and models for which relatively little posterior mass falls in the restricted parameter space. This paper describes the underlying methodology and illustrates the use of **multibridge** through fully reproducible examples.

multibridge: An R Package To Evaluate Informed Hypotheses in Binomial and Multinomial Models

1 Introduction

The most common way to analyze categorical variables is to conduct either binomial tests, multinomial tests, or chi-square goodness of fit tests. These tests compare the encompassing hypothesis to a null hypothesis that all underlying category proportions are either exactly equal, or follow a specific distribution. Accordingly, these tests are suitable when theories predict either the invariance of all category proportions or specific values. For instance, chi-square goodness of fit tests are commonly used to test Benford's law, which predicts the distribution of leading digits in empirical datasets (Benford, 1938; Newcomb, 1881). Often, however, the predictions that researchers are interested in are of a different kind. Consider for instance the weak-order mixture model of decision-making (Regenwetter & Davis-Stober, 2012). The theory predicts that individuals' choice preferences are weakly ordered at all times, that is, if they prefer choice A over B and B over C then they will also prefer A over C (Regenwetter, Dana, & Davis-Stober, 2011)—a well-constrained prediction of behavior. The theory is, however, silent about the exact values of each choice preference. Hence, the standard tests that compare $\mathcal{H}_e$ to $\mathcal{H}_0$ are unsuited to test the derived predictions. Instead, the predictions need to be translated into an informed hypothesis $\mathcal{H}_r$ that reflects the predicted ordinal relations among the parameters. Only then is it possible to adequately test whether the theory of weakly-ordered preference describes participants' choice behavior. Of course, researchers may be interested in more complex hypotheses, including ones that feature combinations of equality constraints, inequality constraints, and unconstrained category proportions. For instance, Nuijten, Hartgerink, Assen, Epskamp, and Wicherts (2016) hypothesized that articles published in social psychology journals would have higher error rates than articles published in other psychology journals. As in the previous example, the authors had no expectations about the exact error rate distribution across journals.

₅₈ Here, again, the standard tests are inadequate. Generally, by specifying informed hypotheses

₅₉ researchers and practitioners are able to "add theoretical expectations to the traditional

₆₀ alternative hypothesis" (Hoijtink, Klugkist, & Boelen, 2008, p. 2) and thus test hypotheses

₆₁ that relate more closely to their theories (Haaf, Klaassen, & Rouder, 2019; Rijkeboer & van

₆₂ den Hout, 2008).

In the Bayesian framework, researchers may test hypotheses of interest by means of
Bayes factors (Jeffreys, 1935; Kass & Raftery, 1995). Bayes factors quantify the extent to
which the data change the prior model odds to the posterior model odds, that is, the extent
to which one hypothesis outpredicts the other. Specifically, Bayes factors are the ratio of
marginal likelihoods of the respective hypotheses. For instance, the Bayes factor for the
informed hypothesis versus the encompassing hypothesis is defined as:

$$\text{BF}_{re} = \frac{\overbrace{p(\mathbf{x} \mid \mathcal{H}_r)}^{\substack{\text{Marginal likelihood} \\ \text{under } \mathcal{H}_r}}}{\underbrace{p(\mathbf{x} \mid \mathcal{H}_e)}_{\substack{\text{Marginal likelihood} \\ \text{under } \mathcal{H}_e}}},$$

₆₃ where the subscript r denotes the informed hypothesis and e denotes the encompassing

₆₄ hypothesis. Several available `R` packages compute Bayes factors for informed hypotheses. For

₆₅ instance, the package **multinomineq** (Heck & Davis-Stober, 2019) evaluates informed

₆₆ hypotheses for multinomial models as well as models that feature independent binomials.

₆₇ The package **BFpack** (Mulder et al., in press) evaluates informed hypotheses for statistical

₆₈ models such as univariate and multivariate normal linear models, generalized linear models,

₆₉ special cases of linear mixed models, survival models, and relational event models. The

₇₀ package **BAIN** (Gu, Hoijtink, Mulder, & Rosseel, 2019) evaluates informed hypotheses for

₇₁ structural equation models. Outside of `R`, the Fortran 90 program **BIEMS** (Mulder,

₇₂ Hoijtink, & de Leeuw, 2012) evaluates informed hypotheses for multivariate linear models

₇₃ such as MANOVA, repeated measures, and multivariate regression. All these packages rely

₇₄ on one of two implementations of the encompassing prior approach (Klugkist, Kato, &

Hoijtink, 2005; Sedransk, Monahan, & Chiu, 1985) to approximate order constrained Bayes

factors: the unconditional encompassing method (Hoijtink, 2011; Hoijtink et al., 2008;

Klugkist et al., 2005) and the conditional encompassing method (Gu, Mulder, Deković, &

Hoijtink, 2014; Laudy, 2006; Mulder, 2014, 2016; Mulder et al., 2009). Even though the

encompassing prior approach is currently the most common method to evaluate informed

hypotheses, it becomes increasingly unreliable and inefficient as the number of restrictions

increases or the parameter space of the restricted model decreases (Sarafoglou et al., in

press).

As alternative to the encompassing prior approach, Sarafoglou et al. (in press) recently

proposed a bridge sampling routine (Bennett, 1976; Meng & Wong, 1996) that computes

Bayes factors for informed hypotheses more reliably and efficiently. This routine is

implemented in **multibridge** (https://CRAN.R-project.org/package=multibridge) and is

suitable to evaluate inequality constraints for multinomial and binomial models. When an

informed hypothesis includes mixtures of equality and inequality constraints, the core

functions in **multibridge** split the hypothesis to compute Bayes factors separately for

equality constraints (for which the Bayes factor has an analytic solution) and inequality

constraints (for which the Bayes factor is estimated using bridge sampling). The core

functions of **multibridge**, that is `mult_bf_informed` and `binom_bf_informed`, return the

Bayes factor estimate in favor of or against the informed hypothesis (see Table 1 for a

summary of the basic required arguments of the two core functions). In addition, users can

visualize the posterior parameter estimates under the encompassing hypothesis using the

`plot`-method, or get more detailed information on how the Bayes factor is composed using

the `summary`-method. For hypotheses that include mixtures between equality and inequality

constrained hypotheses the `bayes_factor` method separately returns the Bayes factor for

the equality constraints and the conditional Bayes factor for the inequality constraints given

the equality constraints. The informed hypothesis can be conveniently specified using a

string or character vector. Furthermore, the transitivity property of Bayes factors can be

102 used to test two informed hypotheses against each other (see Example 1 for an illustration).

103 The general workflow of **multibridge** is illustrated in Figure 1. Table 2 summarizes all S3

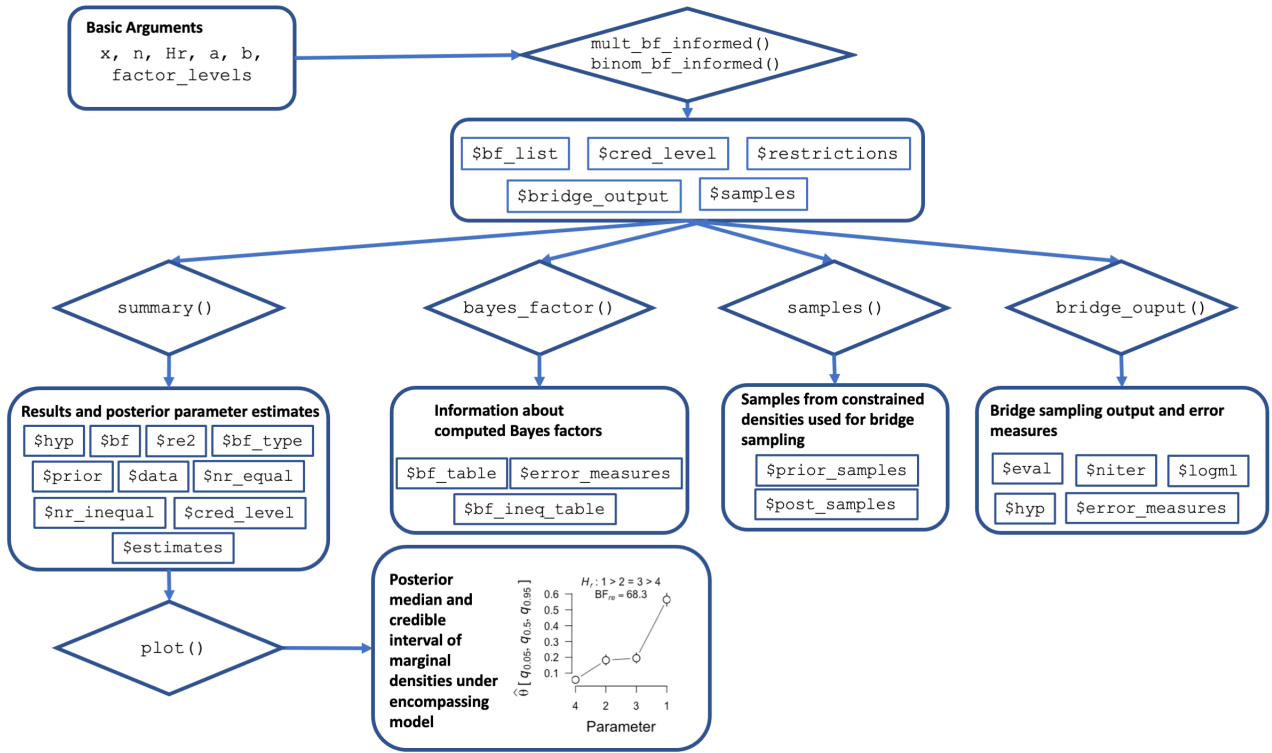104 methods currently available in **multibridge**.

Figure 1.   The **multibridge** workflow.   When calling `mult_bf_informed` or `binom_bf_informed`, the user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and β parameters of the binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions then return the estimated Bayes factor for the informed hypothesis relative to the encompassing or the null hypothesis. Based on these results different S3 methods can be used to get more detailed information on the individual components of the analysis (e.g., `summary`, `bayes_factor`), and parameter estimates of the encompassing distribution (`plot`).

Table 1

To estimate the Bayes factor in favor for or against the specified informed hypothesis, the user provides the core functions `mult_bf_informed` *and* `binom_bf_informed` *with the basic required arguments listed below.*

Argument	Description
`x`	`numeric`. Vector with data (for multinomial models) or a vector of counts of successes, or a two-dimensional table (or matrix) with 2 columns, giving the counts of successes and failures, respectively (for binomial models).
`n`	`numeric`. Vector with counts of trials. Must be the same length as `x`. Ignored if `x` is a matrix or a table. Included only in `binom_bf_informed`.
`Hr`	`string` or `character`. String or vector with the user specified informed hypothesis. Parameters may be referenced by the specified `factor_levels` or by numerical indices.
`a`	`numeric`. Vector with concentration parameters of Dirichlet distribution (for multinomial models) or α parameters for independent beta distributions (for binomial models). Must be the same length as `x`. Default sets all parameters to 1.
`b`	`numeric`. Vector with β parameters. Must be the same length as `x`. Default sets all β parameters to 1. Included only in `binom_bf_informed`.
`factor_levels`	`character`. Vector with category labels. Must be the same length as `x`.

Table 2

*S3 methods available in **multibridge**.*

Function Name(s)	S3 Method	Description
`mult_bf_informed`, `binom_bf_informed`	`print`	Prints model specifications and descriptives.
	`summary`	Prints and returns the Bayes factor and associated hypotheses for the full model, and all equality and inequality constraints.
	`plot`	Plots the posterior median and credible interval of the parameter estimates of the encompassing model. Default sets credible interval to 95%.
	`bayes_factor`	Contains all Bayes factors and log marginal likelihood estimates for inequality constraints.
	`samples`	Extracts prior and posterior samples from constrained densities (if bridge sampling was applied).
	`bridge_output`	Extracts bridge sampling output and associated error measures.
	`restriction_list`	Extracts restriction list and associated informed hypothesis.
`mult_bf_inequality`, `binom_bf_inequality`	`print`	Prints the bridge sampling estimate for the log marginal likelihood and the corresponding percentage error.
	`summary`	Prints and returns the bridge sampling estimate for the log marginal likelihood and associated error terms.

105   This paper showcases how the proposed bridge sampling routine by Sarafoglou et al. (in

106   press) can be applied in a user-friendly way with **multibridge**. In the remainder of this

107   article, we will describe the Bayes factor identity for informed hypotheses in binomial and

108   multinomial models, and briefly describe the bridge sampling method. Then, we illustrate the

109   core functions of **multibridge** package using two examples and end with a brief summary.

110   **2   Methods**

111   In this section we formalize multinomial models and models that feature independent

112   binomial probabilities as they have been implemented in **multibridge**. In the multinomial

113   model, we assume that the vector of observations $\mathbf{x}$ in the K categories follows a multinomial

114   distribution in which the parameters of interest, $\boldsymbol{\theta}$, represent the underlying category

115   proportions. Since the K categories are dependent, the vector of probability parameters is

116   constrained to sum to one, such that $\sum_{k=1}^{K}(\theta_1, \cdots, \theta_K) = 1$. Therefore, a suitable choice for

117   a prior distribution for $\boldsymbol{\theta}$ is the Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha}$:

$$x_1, \cdots, x_K \sim \text{Multinomial}(\sum_{k=1}^{K} x_k, \theta_1, \cdots, \theta_K) \tag{1}$$

$$\theta_1, \cdots, \theta_K \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K), \tag{2}$$

118   where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* category counts. The formalization of the

119   model for independent binomial probabilities is similar since the multinomial model above

120   constitutes a generalization of the binomial model (for $K \geq 2$). In the binomial model, we

121   assume that the elements in the vector of successes $\mathbf{x}$ and the elements in the vector of total

122   number of observations $\mathbf{n}$ in the K categories follow independent binomial distributions. As

123   in the multinomial model, the parameter vector of the binomial success probabilities $\boldsymbol{\theta}$

124   contains the underlying category proportions, however, in this model we assume that

125   categories are independent which removes the sum-to-one constraint. Therefore, a suitable

126 choice for a prior distribution for $\boldsymbol{\theta}$ is a vector of independent beta distributions with

127 parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$x_1 \cdots x_K \sim \prod_{k=1}^{K} \text{Binomial}(\theta_k, n_k) \tag{3}$$

$$\theta_1 \cdots \theta_K \sim \prod_{k=1}^{K} \text{Beta}(\alpha_k, \beta_k), \tag{4}$$

128 where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* successes that observations fall within the

129 various categories and $\boldsymbol{\beta}$ can be interpreted as vector of *a priori* failures.

130 **2.1   Bayes factor**

131 **multibridge** features two different methods to compute Bayes factors: one method

132 computes Bayes factors for equality constrained parameters and one method computes Bayes

133 factors for inequality constrained parameters. Both methods will be outlined below. In cases

134 where informed hypotheses feature mixtures between inequality and equality constraints, we

135 compute the overall Bayes factor BF_{re} by multiplying the individual Bayes factors for both

136 constraint types. This is motivated by the fact that the Bayes factor for mixtures will factor

137 into a Bayes factor for the equality constraints and a conditional Bayes factor for the

138 inequality constraints given the equality constraints (see Sarafoglou et al., in press, for the

139 proof).

2.1.1   The Bayes Factor For Equality Constraints.   In **multibridge** the

Bayes factor for the equality constraints can be computed analytically both for binomial and

multinomial models using the functions `binom_bf_equality` and `mult_bf_equality`. For

binomial models, assuming that the all binomial probabilities in a model are exactly equal,

the Bayes factor is defined as:

$$\text{BF}_{0e} = \frac{\prod_{k=1}^{K} \text{B}(\alpha_k, \beta_k)}{\prod_{k=1}^{K} \text{B}(\alpha_k + x_k, \beta_k + n_k - x_k)} \times \frac{\text{B}(\alpha_+ + x_+ + 1, \beta_+ + n_+ - x_+ + 1)}{\text{B}(\alpha_+ + 1, \beta_+ + 1)},$$

where $\text{B}(\cdot)$ denotes the beta function and $\alpha_+ = \sum_{k=1}^{K} \alpha_k$, $\beta_+ = \sum_{k=1}^{K} \beta_k$, $x_+ = \sum_{k=1}^{K} x_k$ and $n_+ = \sum_{k=1}^{K} n_k$. If all binomial probabilities in a model are assumed to be exactly equal *and* equal to a predicted value θ_0, the Bayes factor is defined as:

$$\text{BF}_{0e} = \frac{\prod_{k=1}^{K} \text{B}(\alpha_k, \beta_k)}{\prod_{k=1}^{K} \text{B}(\alpha_k + x_k, \beta_k + n_k - x_k)} \times \theta_0^{x_+} (1 - \theta_0)^{n_+ - x_+}.$$

140 Note that **multibridge** only supports the specification of one predicted value for all

141 binomial probabilities. The package does not support the specification of different predicted

142 values for different binomial probabilities. The reason for this is theoretical: we believe that

143 such hypotheses are better tested using a hierarchical structure (thus modeling the binomial

144 probabilities as dependent).

For multinomial models, assuming that all category proportions in a model are equality constrained, the Bayes factor BF_{0e} is defined as:

$$\text{BF}_{0e} = \frac{\text{B}(\alpha_1, \ldots, \alpha_K)}{\text{B}(\alpha_1 + x_1, \ldots, \alpha_K + x_K)} \times \frac{\text{B}(\boldsymbol{\alpha} + \mathbf{x})}{\text{B}(\boldsymbol{\alpha})} \times \prod_{k=1}^{K} \theta_{0k}^{x_k},$$

145 where θ_{0k} represent the predicted category proportions. When all category proportions are

146 assumed to be exactly equal all θ_{0k} are set to $\frac{1}{K}$. Otherwise, $\boldsymbol{\theta}_0$ is replaced with the

147 user-specified predicted values.

148   **2.1.2   The Bayes Factor For Inequality Constraints.**   To approximate the

149 Bayes factor for informed hypotheses, Klugkist et al. (2005) derived an identity that defines

150 the Bayes factor BF_{re} as the ratio of proportions of posterior and prior parameter space

151 consistent with the restriction. This identity forms the basis of the encompassing prior

152 approach. Recently, Sarafoglou et al. (in press) highlighted that these proportions can be

153 reinterpreted as the marginal likelihoods (i.e., the normalizing constants) of the constrained

154 posterior and constrained prior distribution:

$$\text{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}^{\substack{\text{Marginal likelihood of} \\ \text{constrained posterior distribution}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Marginal likelihood of} \\ \text{constrained prior distribution}}}}. \tag{5}$$

The benefit of reinterpreting the identity by Klugkist et al. (2005) is that we can estimate the Bayes factor by utilizing numerical sampling methods such as bridge sampling. For that we only need to be able to sample from the constrained densities. Crucially, when using bridge sampling, it does not matter how small the constrained parameter space is in proportion to the encompassing density. This gives the method a decisive advantage over the encompassing prior approach in terms of accuracy and efficiency especially (1) when binomial and multinomial models with moderate to high number of categories (i.e., $K > 10$) are evaluated and (2) when relatively little posterior mass falls in the constrained parameter space.

The bridge sampling algorithm implemented in **multibridge** estimates one marginal likelihood at the time (cf., Gronau et al., 2017; Overstall & Forster, 2010). Specifically, we separately estimate the marginal likelihood for the constrained prior distribution and the marginal likelihood of the constrained posterior distribution. Here we describe how to estimate the marginal likelihood for the constrained prior distribution; the steps presented can then be applied accordingly to the posterior distribution. It should be noted that the bridge sampling algorithm implemented in **multibridge** is an adapted version of the algorithm implemented in the `R` package **bridgesampling** (Gronau, Singmann, & Wagenmakers, 2020) and allows for the specification of informed hypotheses on probability vectors.[1] The bridge sampling identity for the marginal likelihood of the constrained prior distribution is defined as:

---

[1] In addition, the function to compute the relative mean square error for bridge sampling estimates in **multibridge** is based on the code of the `error_measures`-function from the **bridgesampling** package.

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\text{prior}}\left(g(\boldsymbol{\theta})h(\boldsymbol{\theta})\right)}, \tag{6}$$

where the term $h(\boldsymbol{\theta})$ refers to the bridge function proposed by Meng and Wong (1996), $g(\boldsymbol{\theta})$ refers to a so-called proposal distribution, and $p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$ is the part of the prior parameter space under the encompassing hypothesis that is in accordance with the constraint. To estimate the marginal likelihood, bridge sampling requires samples from the target distribution, that is, the constrained Dirichlet distribution for multinomial models and constrained beta distributions for binomial models, and samples from the proposal distribution which in principle can be any distribution with a known marginal likelihood; in **multibridge** the proposal distribution is the multivariate normal distribution. Samples from the target distribution are generated using the Gibbs sampling algorithms proposed by Damien and Walker (2001). For binomial models, we apply the suggested Gibbs sampling algorithm for constrained beta distributions. In the case of the multinomial models, we apply an algorithm that simulates values from constrained Gamma distributions which are then transformed into Dirichlet random variables. To sample efficiently from these distributions, **multibridge** provides a `C++` implementation of this algorithm. Samples from the proposal distribution are generated using the standard `rmvnorm`-function from the `R` package **mvtnorm** (Genz et al., 2020).

The efficiency of the bridge sampling method is optimal only if the target and proposal distribution operate on the same parameter space and have sufficient overlap. We therefore probit transform the samples of the constrained distributions to move the samples from the probability space to the entire real line. Subsequently, we use half of these draws to construct the proposal distribution using the method of moments. Details on the probit transformations are provided in the appendix.

The numerator in Equation 6 evaluates the unnormalized density for the constrained

197   prior distribution with samples from the proposal distribution. The denominator evaluates

198   the normalized proposal distribution with samples from the constrained prior distribution.

199   Using this identity, we obtain the bridge sampling estimator for the marginal likelihood of

200   the constrained prior distribution by applying the iterative scheme proposed by Meng and

201   Wong (1996):

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} \approx \frac{\frac{1}{N_2} \sum_{m=1}^{N_2} \frac{\ell_{2,m}}{s_1 \ell_{2,m} + s_2 p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}}{\frac{1}{N_1} \sum_{n=1}^{N_1} \frac{1}{s_1 \ell_{1,n} + s_2 p(\boldsymbol{\theta}_{\boldsymbol{n}}^* \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}},$$

202   where N_1 denotes the number of samples drawn from the constrained distribution, that is,

203   $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$, N_2 denotes the number of samples drawn from the proposal distribution, that

204   is $\tilde{\boldsymbol{\theta}} \sim g(\boldsymbol{\theta})$, $s_1 = \frac{N_1}{N_2 + N_1}$, and $s_2 = \frac{N_2}{N_2 + N_1}$. The quantities $\ell_{1,n}$ and $\ell_{2,m}$ are defined as follows:

$$\ell_{1,n} = \frac{q_{1,1}}{q_{1,2}} = \frac{p(\boldsymbol{\theta}_{\boldsymbol{n}}^* \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta}_{\boldsymbol{n}}^* \in \mathcal{R}_r)}{g(\boldsymbol{\xi}_{\boldsymbol{n}}^{\,*})}, \tag{7}$$

$$\ell_{2,m} = \frac{q_{2,1}}{q_{2,2}} = \frac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}})}, \tag{8}$$

where $\boldsymbol{\xi}_{\boldsymbol{n}}^{\,*} = \Phi^{-1}\left(\dfrac{\boldsymbol{\theta}_{\boldsymbol{n}}^* - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}}) + \mathbf{l})|J|)$. The quantity $q_{1,1}$ refers to the

evaluations of the constrained distribution for constrained samples and $q_{1,2}$ refers to the

proposal distribution evaluated at the probit-transformed samples from the constrained

distribution, respectively. The quantity $q_{2,1}$ refers to evaluations of the constrained

distribution at the inverse probit-transformed samples from the proposal distribution and

$q_{2,2}$ refers to the proposal evaluations for samples from the proposal, respectively. Note that

the quantities $\ell_{1,n}$ and $\ell_{2,m}$ have been adjusted to account for the necessary parameter

transformations to create overlap between the constrained distributions and the proposal

distribution. **multibridge** runs the iterative scheme until the tolerance criterion suggested by Gronau et al. (2017) is reached, that is:

$$\frac{\mid \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} - \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)} \mid}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)}} \leq 10^{-10}.$$

The sampling from the target and proposal distribution, the transformations and computational steps are performed automatically within the core functions of **multibridge**. The user only needs to provide the functions with the data, a prior and a specification of the informed hypothesis. As part of the standard output of `binom_bf_informed` and `mult_bf_informed`, the functions return the bridge sampling estimate for the log marginal likelihood of the target distribution, its associate relative mean square error, the number of iterations, and the quantities $q_{1,1}$, $q_{1,2}$, $q_{2,1}$, and $q_{2,2}$.

3   Usage and Examples

In the following, we will outline two examples on how to use **multibridge** to compare an informed hypothesis to a null or encompassing hypothesis. The first example concerns multinomial data and the second example concerns independent binomial data.

A list of all currently available functions and data sets is given in Table 3. Additional examples are available as vignettes (see `vignette(package = "multibridge")`). The two core functions of **multibridge**—`mult_bf_informed` and the `binom_bf_informed`—can be illustrated schematically as follows:

```
mult_bf_informed(x, Hr, a, factor_levels)
binom_bf_informed(x, n, Hr, a, b, factor_levels)
```

Table 3

*Core functions available in **multibridge**.*

Function Name(s)	Description
`mult_bf_informed`	Evaluates informed hypotheses on multinomial parameters.
`mult_bf_inequality`	Estimates the marginal likelihood of a constrained prior or posterior Dirichlet distribution.
`mult_bf_equality`	Computes Bayes factor for equality constrained multinomial parameters using the standard Bayesian multinomial test.
`mult_tsampling`	Samples from constrained prior or posterior Dirichlet density.
`lifestresses`, `peas`	Data sets associated with informed hypotheses in multinomial models.
`binom_bf_informed`	Evaluates informed hypotheses on binomial parameters.
`binom_bf_inequality`	Estimates the marginal likelihood of constrained prior or posterior beta distributions.
`binom_bf_equality`	Computes Bayes factor for equality constrained binomial parameters.
`binom_tsampling`	Samples from constrained prior or posterior beta densities.
`journals`	Data set associated with informed hypotheses in binomial models.
`generate_restriction_list`	Encodes the informed hypothesis.

3.1   Example 1: Applying A Benford Test to Greek Fiscal Data

The first-digit phenomenon, otherwise known as Benford's law (Benford, 1938; Newcomb, 1881) states that the expected proportion of leading digits in empirical data can be

₂₂₃ formalized as follows: for any given leading digit $d, d = (1, \cdots, 9)$ the expected proportion is

₂₂₄ approximately equal to

$$\mathbb{E}_{\theta_d} = \log_{10}((d+1)/d).$$

₂₂₅ This means that in an empirical data set, numbers with smaller leading digits are more

₂₂₆ common than numbers with larger leading digits. Specifically, a number has leading digit 1

₂₂₇ in 30.1% of the cases, and leading digit 2 in 17.61% of the cases; leading digit 9 is the least

₂₂₈ frequent digit with an expected proportion of only 4.58% (see Table 4 for an overview of the

₂₂₉ expected proportions). Empirical data for which this relationship holds include population

₂₃₀ sizes, death rates, baseball statistics, atomic weights of elements, and physical constants

₂₃₁ (Benford, 1938). In contrast, artificially generated data, such as telephone numbers, do in

₂₃₂ general not obey Benford's law (Hill, 1995). Given that Benford's law applies to empirical

₂₃₃ data but not artificially generated data, a so-called Benford test can be used in fields like

₂₃₄ accounting and auditing to check for indications for poor data quality (for an overview, see

₂₃₅ e.g., Durtschi, Hillison, & Pacini, 2004; Nigrini, 2012; Nigrini & Mittermaier, 1997). Data

₂₃₆ that do not pass the Benford test, should raise audit risk concerns, meaning that it is

₂₃₇ recommended that they undergo additional follow-up checks (Nigrini, 2019).

₂₃₈   Below we discuss four possible Bayesian adaptations of the Benford test. In a first

₂₃₉ scenario we simply conduct a Bayesian multinomial test in which we test the point-null

₂₄₀ hypothesis $\mathcal{H}_0$ which predicts a Benford distribution against the encompassing hypothesis

₂₄₁ $\mathcal{H}_e$. In a second scenario we test the null hypothesis against an alternative hypothesis,

₂₄₂ denoted as $\mathcal{H}_{r1}$, which predicts a decreasing trend in the proportions of leading digits. The

₂₄₃ hypothesis $\mathcal{H}_{r1}$ exerts considerably more constraint than $\mathcal{H}_e$ and provides a more sensitive

₂₄₄ test if our primary goal is to test whether data comply with Benford's law or whether the

₂₄₅ data follow a similar but different trend. In the next two scenarios, our main goal is to

₂₄₆ identify fabricated data. The third scenario therefore tests the null hypothesis against the

₂₄₇ hypothesis that all proportions occur equally often. This hypothesis $\mathcal{H}_{r2}$ could be considered

248 if it is suspected that the data were generated randomly. In a fourth scenario we test the

249 null hypothesis against a hypothesis which predicts a trend that is characteristic for

250 manipulated data. This hypothesis, which we denote as $\mathcal{H}_{r3}$, could be derived from empirical

251 research on fraud or be based on observed patterns from former fraud cases. For instance,

252 Hill (1995) instructed students to produce a series of random numbers; in the resulting data

253 the proportion of the leading digit 1 occurred most often and the digits 8 and 9 occurred

254 least often which is consistent with the general pattern of Benford's law. However, the

255 proportion for the remaining leading digits were approximately equal. Note that the

256 predicted distribution derived from Hill (1995) is not currently used as a test to detect fraud.

257 However, for the sake of simplicity, if we assume that this pattern could be an indication of

258 manipulated auditing data, the Bayes factor BF_{0r3} would quantify the evidence of whether

259 the proportion of first digits resemble authentic or fabricated data.

260 **3.1.1   Data and Hypothesis.**   The data we use to illustrate the computation of

261 Bayes factors were originally published by the European statistics agency "Eurostat" and

262 served as basis for reviewing the adherence to the Stability and Growth Pact of EU member

263 states. Rauch, Göttsche, Brähler, and Engel (2011) conducted a Benford test on data related

264 to budget deficit criteria, that is, public deficit, public dept and gross national products. The

265 data used for this example features the proportion of first digits from Greek fiscal data in the

266 years between 1999 and 2010; a total of $N = 1{,}497$ numerical data were included in the

267 analysis. We choose this data, since the Greek government deficit and debt statistics states

268 has been repeatedly criticized by the European Commission in this time span (European

269 Commision, 2004, 2010). In particular, the commission has accused the Greek statistical

270 authorities to have misreported deficit and debt statistics. For further details on the data set

271 see Rauch et al. (2011). The observed and expected proportions are displayed in Table 4;

272 the expected proportions versus the posterior parameter estimates under the encompassing

273 hypothesis are displayed in Figure 2.

Table 4

Observed counts, observed proportions, and expected proportions of first digits in the Greek fiscal data set. The total sample size was $N = 1{,}497$ observations. Note that the observed proportions and counts deviate slightly from those reported in Rauch et al. (2011) (probably due to rounding errors).

Leading digit	Observed Counts	Observed Proportions	Expected Proportions: Benford's Law
1	509	0.340	0.301
2	353	0.236	0.176
3	177	0.118	0.125
4	114	0.076	0.097
5	77	0.051	0.079
6	77	0.051	0.067
7	53	0.035	0.058
8	73	0.049	0.051
9	64	0.043	0.046

In this example, the parameter vector of the multinomial model, $\theta_1, \cdots, \theta_K$, reflects the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9. The hypotheses introduced above can then be formalized as follows. The null hypothesis specifies that the proportions of first digits obeys Benford's law:

$$\mathcal{H}_0 : \boldsymbol{\theta}_0 = (0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046).$$

This null hypothesis can then be tested against each of the following four alternative

hypotheses:

$$\mathcal{H}_e : \boldsymbol{\theta} \sim \text{Dirichlet}(\mathbf{1}),$$

$$\mathcal{H}_{r1} : \theta_1 > \theta_2 > \theta_3 > \theta_4 > \theta_5 > \theta_6 > \theta_7 > \theta_8 > \theta_9,$$

$$\mathcal{H}_{r2} : \boldsymbol{\theta}_0 = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9} \right),$$

$$\mathcal{H}_{r3} : \theta_1 > (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7) > (\theta_8, \; \theta_9).$$

The comparison of any two informed hypotheses with one another follows from the fact that Bayes factors are transitive. For instance, the Bayes factor comparison between $\mathcal{H}_0$ and $\mathcal{H}_{r1}$ can be obtained by first computing BF_{e0} and BF_{er1}, and then dividing out the common hypothesis $\mathcal{H}_e$:

$$\text{BF}_{0r1} = \frac{\text{BF}_{e0}}{\text{BF}_{er1}}.$$

An overview of the relative plausibility of all $M = 5$ models simultaneously may be obtaining by presenting the posterior model probabilities $p(\mathcal{H}_i \,|\, x)$ (Berger & Molina, 2005). Denoting the prior model probability for model $\mathcal{H}_i$ by $p(\mathcal{H}_i)$, the posterior model probability for $\mathcal{H}_0$ is given by:

$$p(\mathcal{H}_0 \mid \mathbf{x}) = \frac{\dfrac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)} \times p(\mathcal{H}_0)}{\displaystyle\sum_{i=1}^{M} \dfrac{p(\mathbf{x} \mid \mathcal{H}_i)}{p(\mathbf{x} \mid \mathcal{H}_e)} \times p(\mathcal{H}_i)}.$$

When all hypotheses are equally likely *a priori*, this simplifies to:

$$p(\mathcal{H}_0 \mid \mathbf{x}) = \frac{\text{BF}_{0e}}{\text{BF}_{0e} + \text{BF}_{r1e} + \text{BF}_{r2e} + \text{BF}_{r3e} + \text{BF}_{ee}}.$$

3.1.2   Method.   Both BF_{0e} and BF_{r2e} may be readily computed by means of a Bayesian multinomial test which is implemented in the function `mult_bf_equality`. This

289 function requires (1) a vector with observed counts, (2) a vector with concentration

290 parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, and (3) the vector of expected

291 proportions under $\mathcal{H}_0$ and under $\mathcal{H}_{r2}$. We do not incorporate specific expectations about the

292 distribution of leading digits in the Greek fiscal data and therefore set all concentration

293 parameters under $\mathcal{H}_e$ to 1 (i.e., we assign $\boldsymbol{\theta}$ a uniform Dirichlet prior distribution).

```r
# Observed counts
x <- c(509, 353, 177, 114,  77,  77,  53,  73,  64)
# Prior specification for Dirichlet prior distribution under H_e
a <-  rep(1, 9)
# Expected proportions for H_0 and H_r2
p0  <- log10((1:9 + 1)/1:9)
pr2 <- rep(1/9, 9)
# Execute the analysis
results_H0_He  <- mult_bf_equality(x = x, a = a, p = p0)
results_Hr2_He <- mult_bf_equality(x = x, a = a, p = pr2)


logBFe0  <- results_H0_He$bf$LogBFe0
logBFer2 <- results_Hr2_He$bf$LogBFe0
```

294   The hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r3}$ contain inequality constraints, and this necessitates the

295 use of the function `mult_bf_informed` to compute the Bayes factors BF_{r1e} and BF_{r3e}. This

296 function requires (1) a vector with observed counts, (2) a vector with concentration

297 parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, (3) labels for the categories of

298 interest (i.e., leading digits), and (4) the informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r3}$ (e.g., as a string):

```r
# Observed counts
x <- c(509, 353, 177, 114,  77,  77,  53,  73,  64)
# Prior specification for Dirichlet prior distribution under H_e
```

```r
a <-  rep(1, 9)
# Labels for categories of interest
factor_levels <- 1:9
# Specifying the informed hypotheses as a string
Hr1 <- c('1 > 2 > 3 > 4 > 5 > 6 > 7 > 8 > 9')
Hr3 <- c('1 > 2 = 3 = 4 = 5 = 6 = 7 > 8 > 9')
# Execute the analysis
results_He_Hr1 <- mult_bf_informed(x = x, Hr = Hr1, a = a,
                                   factor_levels = factor_levels,
                                   bf_type = 'LogBFer', seed = 2020)
logBFer1 <- summary(results_He_Hr1)$bf
results_He_Hr3 <- mult_bf_informed(x = x, Hr = Hr3, a = a,
                                   factor_levels = factor_levels,
                                   bf_type = 'LogBFer', seed = 2020)
logBFer3 <- summary(results_He_Hr3)$bf
```

299   We may now exploit transitivity to compare all alternative hypotheses to the Benford

300   null hypothesis $\mathcal{H}_0$. We also compute the posterior model probabilities for all hypotheses.

301   The results are shown in Table 5.

Table 5

Prior model probabilities, posterior model probabilities, and Bayes factors for five rival accounts of first digit frequencies in the Greek fiscal data set.

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mid \mathbf{x})$	$\log(\mathrm{BF}_{.0})$
$\mathcal{H}_0$	0.2	1.27×10^{-11}	0
$\mathcal{H}_{r1}$	0.2	0.9994	25.09
$\mathcal{H}_e$	0.2	0.0006	17.67
$\mathcal{H}_{r3}$	0.2	9.46×10^{-79}	-154.57
$\mathcal{H}_{r2}$	0.2	2.71×10^{-212}	-462.06

The results indicate strong support for $\mathcal{H}_{r1}$ –the model in which the proportions are assumed to decrease monotonically– over all other models. The log Bayes factor of $\mathcal{H}_{r1}$ against Benford's law $\mathcal{H}_0$ is an overwhelming 25.09; the evidence for $\mathcal{H}_{r1}$ is even stronger when it is compared against models that feature equality constraints (i.e., $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$). Finally, $\mathcal{H}_{r1}$ also outperforms model $\mathcal{H}_e$, the unconstrained model in which all parameters are free to vary. The latter result demonstrates how a parsimonious model that makes precise predictions can be favored over a model that is more complex (e.g., Jefferys & Berger, 1992). The strong Bayes factor support for $\mathcal{H}_{r1}$ translates to a relatively extreme posterior model probability of 0.9994.

To summarize, the data offer overwhelming support for hypothesis $\mathcal{H}_{r1}$, which postulates a decreasing trend in the digit proportions. This model outperformed both simpler models (e.g., the Benford model) and a more complex model in which the proportions were free to vary. Detailed follow-up analyses are needed to discover why the data follow a monotonically decreasing pattern but not any of the two specific patterns that were put to the test (Nigrini, 2019).
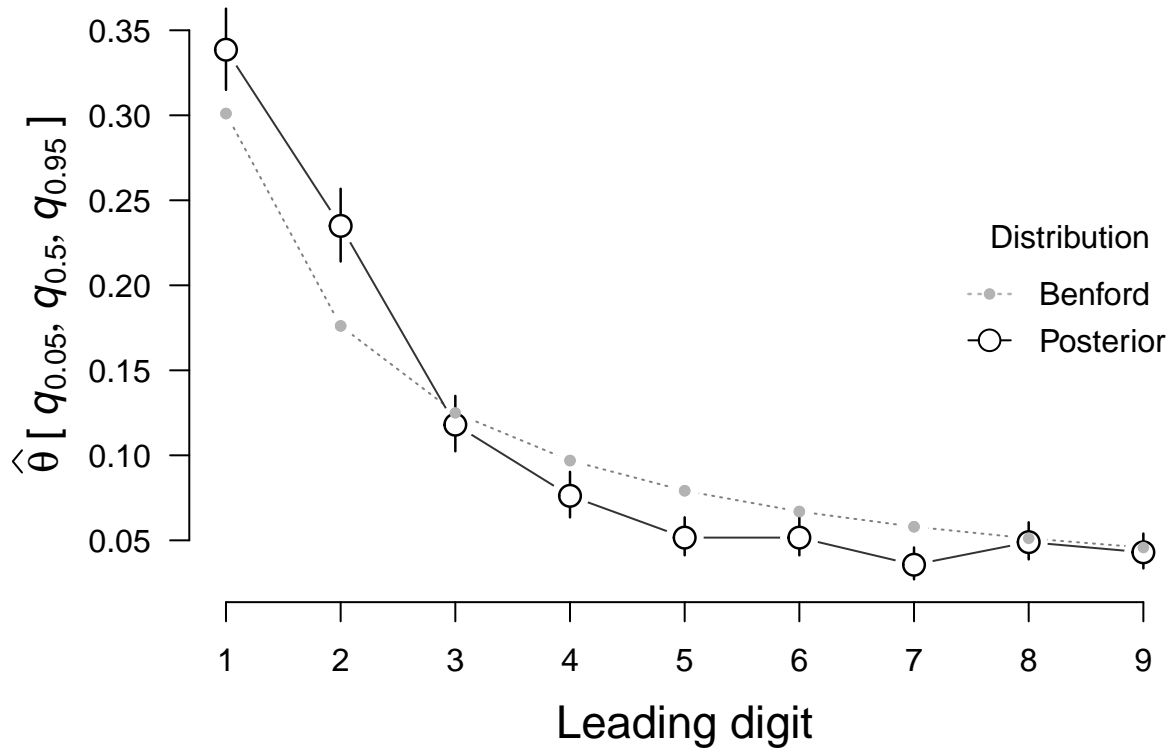
Figure 2. Predictions from Benford's law (in grey) show together with the posterior medians (black circles) for the category proportions estimated under the encompassing model $\mathcal{H}_e$. The circle skewers show the 95% credible intervals. Only three of nine intervals encompass the expected proportions, suggesting that the data do not follow Benford's law. This plot was created using the `plot`-S3-method for `summary.bmult` objects in **multibridge**.

3.2   Example 2: Prevalence of Statistical Reporting Errors

This section illustrates how **multibridge** may be used to evaluate models for independent binomial data rather than multinomial data. Our example concerns the prevalence of statistical reporting errors across eight different psychology journals. In any article that uses null hypothesis significance testing, there is a chance that the reported test statistic and degrees of freedom do not match the reported p-value, possibly because of copy-paste errors. To flag these errors, Epskamp and Nuijten (2014) developed the R package `statcheck`, which scans the PDF of a given scientific article and automatically detects statistical

₃₂₅ inconsistencies. This package allowed Nuijten et al. (2016) to estimate the prevalence of

₃₂₆ statistical reporting errors in the field of psychology. In total, the authors investigated a

₃₂₇ sample of 30,717 articles (which translates to over a quarter of a million p-values) published

₃₂₈ in eight major psychology journals between 1985 to 2013: *Developmental Psychology* (DP),

₃₂₉ the *Frontiers in Psychology* (FP), the *Journal of Applied Psychology* (JAP), the *Journal of*

₃₃₀ *Consulting and Clinical Psychology* (JCCP), *Journal of Experimental Psychology: General*

₃₃₁ (JEPG), the *Journal of Personality and Social Psychology* (JPSP), the *Public Library of*

₃₃₂ *Science* (PLoS), *Psychological Science* (PS).

₃₃₃   Based on several background assumptions, Nuijten et al. (2016) predicted that the

₃₃₄ proportion of statistical reporting errors is higher for articles published in the *Journal of*

₃₃₅ *Personality and Social Psychology* (JPSP) than for articles published in the seven other

₃₃₆ journals.

₃₃₇   **3.2.1   Data and Hypothesis.**   Here we reuse the original data published by

₃₃₈ Nuijten et al. (2016), which we also distribute with the package **multibridge** under the

₃₃₉ name `journals`.

```
data(journals)
```

₃₄₀   The Nuijten et al. (2016) hypothesis of interest, $\mathcal{H}_r$, states that the prevalence for

₃₄₁ statistical reporting errors is higher for JPSP than for the other journals.[2] We will consider

₃₄₂ two specific versions of the Nuijten et al. (2016) $\mathcal{H}_r$ hypothesis. The first hypothesis, $\mathcal{H}_{r1}$,

₃₄₃ stipulates that JPSP has the highest prevalence of reporting inconsistencies, whereas the

₃₄₄ other seven journals share a prevalence that is lower. The second hypothesis, $\mathcal{H}_{r2}$, also

₃₄₅ stipulates that JPSP has the highest prevalence of reporting inconsistencies, but does not

₃₄₆ commit to any particular structure on the prevalence for the other seven journals.

---

[2]Nuijten et al. (2016) did not report inferential tests because they had sampled the entire population. We do report inferential tests here because we wish to learn about the latent data-generating process.

347   The **multibridge** package can be used to test $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ against the null hypothesis

348   $\mathcal{H}_0$ that all eight journals have the same prevalence of statistical reporting errors. In addition,

349   we will compare $\mathcal{H}_{r1}$, $\mathcal{H}_{r2}$, and $\mathcal{H}_0$ against the encompassing hypothesis $\mathcal{H}_e$ that makes no

350   commitment whatsoever about the prevalence of reporting inconsistencies across the eight

351   journals. In this example, the parameter vector of the binomial success probabilities, $\boldsymbol{\theta}$,

352   reflects the probabilities that articles contain at least one statistical reporting inconsistency

353   across journals. Thus, the above hypotheses can be formalized as follows:

$$\mathcal{H}_0 : \theta_{\text{JAP}} = \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}} = \theta_{\text{JPSP}}$$

$$\mathcal{H}_{r1} : (\theta_{\text{JAP}} = \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}}) < \theta_{\text{JPSP}}$$

$$\mathcal{H}_{r2} : (\theta_{\text{JAP}}, \theta_{\text{PS}}, \theta_{\text{JCCP}}, \theta_{\text{PLOS}}, \theta_{\text{DP}}, \theta_{\text{FP}}, \theta_{\text{JEPG}}) < \theta_{\text{JPSP}}$$

$$\mathcal{H}_e : \theta_{\text{JAP}} \cdots \theta_{\text{JPSP}} \sim \prod_{k=1}^{K} \text{Beta}(\alpha_k, \beta_k).$$

354   **3.2.2   Method.**   To compute the Bayes factor BF_{0r} we need to specify (1) a vector

355   with observed successes (i.e., the number of articles that contain a statistical inconsistency),

356   (2) a vector containing the total number of observations (i.e., the number of articles), (3) a

357   vector with prior parameter α_k for each binomial proportion of the beta prior distribution

358   under $\mathcal{H}_e$, (4) a vector with prior parameter β_k for each binomial proportion of the beta

359   prior distribution under $\mathcal{H}_e$, (5) the category labels (i.e., journal names), and (6) the

360   informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r2}$ (e.g., as a string). Since we have no specific expectations

361   about the distribution of statistical reporting errors in any given journal, we set all

362   parameters α_k and β_k to one which corresponds to uniform beta distributions. With this

363   information, we can now conduct the analysis with the function `binom_bf_informed`.

```
# Since percentages are rounded to two decimal values, we round the
# articles with an error to obtain integer values
x <- round(journals$articles_with_NHST *
```

```r
                (journals$perc_articles_with_errors/100))
# Total number of articles
n <- journals$articles_with_NHST
# Prior specification for beta prior distributions under H_e
a <- rep(1, 8)
b <- rep(1, 8)
# Labels for categories of interest
journal_names <- journals$journal


# Specifying the informed Hypothesis
Hr1 <- c('JAP = PS = JCCP = PLOS = DP = FP = JEPG < JPSP')
Hr2 <- c('JAP , PS , JCCP , PLOS , DP , FP , JEPG < JPSP')


# Execute the analysis for Hr1
results_H0_Hr1 <- binom_bf_informed(x = x, n = n, Hr = Hr1, a = a, b = b,
                             factor_levels = journal_names,
                             bf_type = 'LogBFr0', seed = 2020)
# Execute the analysis for Hr2
results_H0_Hr2 <- binom_bf_informed(x = x, n = n, Hr = Hr2, a = a, b = b,
                             factor_levels = journal_names,
                             bf_type = 'LogBFr0', seed = 2020)
```

```r
LogBFe0  <- results_H0_Hr1$bf_list$bf0_table[['LogBFe0']]
LogBFr10 <- summary(results_H0_Hr1)$bf
LogBFr20 <- summary(results_H0_Hr2)$bf
```

Table 6

Prior model probabilities, posterior model probabilities, and Bayes factors for four hypotheses concerning the prevalence of statistical reporting errors across psychology journals.

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mid \mathbf{x})$	$\log(\mathrm{BF}_{.0})$
$\mathcal{H}_0$	0.25	1.6073×10^{-69}	0
$\mathcal{H}_{r2}$	0.25	0.8814	158.28
$\mathcal{H}_e$	0.25	0.1186	156.27
$\mathcal{H}_{r1}$	0.25	1.9517×10^{-37}	73.88

As the evidence is extreme in all four cases, we again report all Bayes factors on the log scale. The Bayes factor $\log(\mathrm{BF}_{r20})$ indicates overwhelming evidence for the informed hypothesis that JPSP has the highest prevalence for statistical reporting inconsistencies compared to the null hypothesis that the statistical reporting errors are equal across all eight journals; $\log(\mathrm{BF}_{r20}) = 158.28$.

For a clearer picture about the ordering of the journals we can investigate the posterior distributions for the prevalence rates obtained under the encompassing model.

```r
plot(summary(results_H0_Hr2), xlab = "Journal")
```

The posterior medians and 95% credible intervals are returned by the `summary`-method and are shown in Figure 3. The figure strongly suggests that the prevalence of reporting inconsistencies is not equal across all eight journals. This impression may be quantified by comparing the null hypothesis $\mathcal{H}_0$ to the encompassing hypothesis $\mathcal{H}_e$. The corresponding Bayes factor equals $\log(\mathrm{BF}_{e0}) = 156.27$, which confirms that the data dramatically undercut the null hypothesis that the prevalence of statistical reporting inconsistencies is equal across journals.

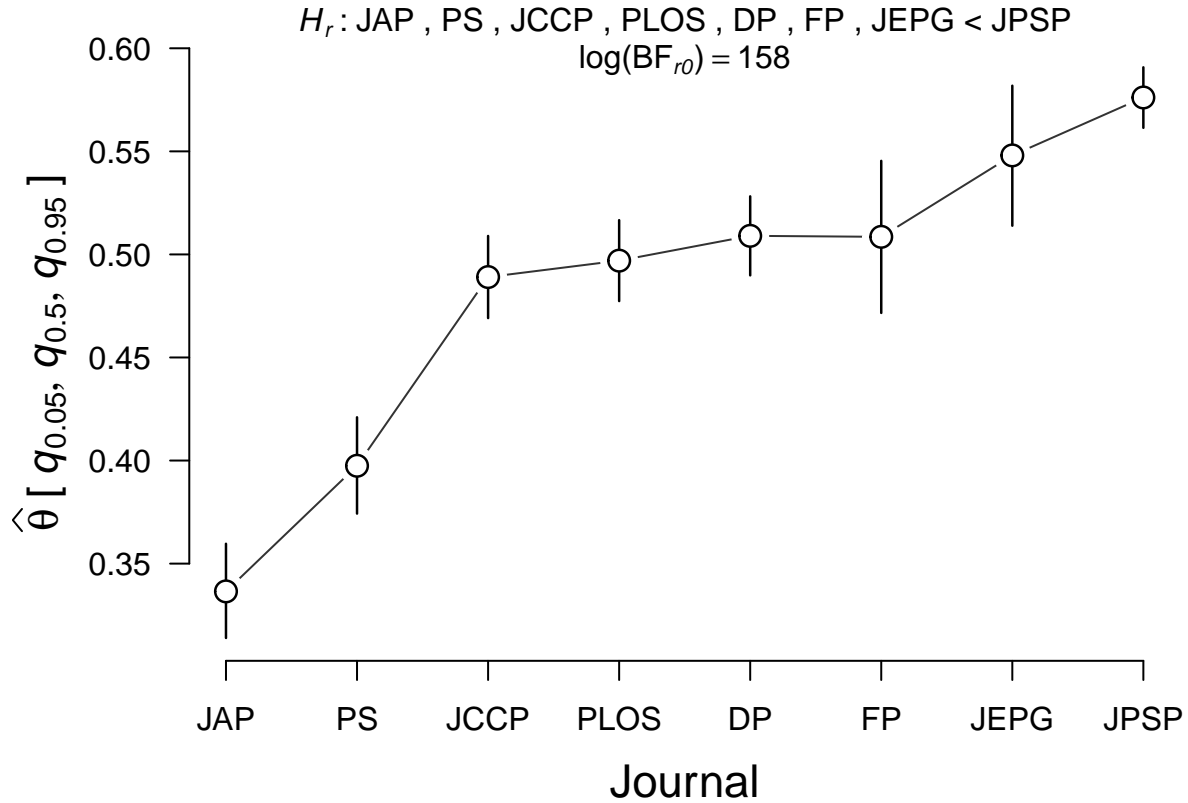The data offer most support for the Nuijten hypothesis $\mathcal{H}_{r2}$, which posits that JPSP

Figure 3. Posterior medians for the prevalence of statistical reporting inconsistencies across eight psychology journals, as obtained using the encompassing model. The circle skewers show the 95% credible intervals. Analysis based on data from Nuijten et al. (2016). This plot was created using the `plot`-S3-method for `summary.bmult` objects.

has the highest prevalence but does not commit to any restriction on the prevalences for the remaining seven journals. This hypothesis may be compared to the encompassing hypothesis $\mathcal{H}_e$, which yields $\log(\text{BF}_{r2e}) = 2.01$. This means that the observed data are $\exp(2.01) \approx 7.45$ times more likely under $\mathcal{H}_{r2}$ than under $\mathcal{H}_e$; this is moderate evidence for the restriction suggested by Nuijten et al. (2016). Under equal prior probability for the models, this Bayes factor translates to a posterior probability on $\mathcal{H}_e$ of 0.119, an amount that researchers may deem too large to discard in an all-or-none fashion.

To summarize, the data provide moderate evidence for the hypothesis stated by Nuijten et al. (2016) that the prevalence of statistical reporting inconsistencies in JPSP is

388 higher than that in seven other psychology journals.

4 Summary

390 The R package **multibridge** facilitates the estimation of Bayes factors for informed

391 hypotheses in both multinomial and independent binomial models. The efficiency gains of

392 **multibridge** are particularly pronounced when the parameter restrictions are highly

393 informative or when the number of categories is large.

394 **multibridge** supports the evaluation of informed hypotheses that feature equality

395 constraints, inequality constraints, and free parameters, as well as mixtures between them.

396 Moreover, users can choose to test the informative hypothesis against an encompassing

397 hypothesis that lets all parameters vary freely or against the null hypothesis that states that

398 category proportions are exactly equal. Beyond the core functions currently implemented in

399 **multibridge**, there are several natural extensions we aim to include in future versions of

400 this package. For instance, to compare several models with each other we plan to implement

401 functions that compute the posterior model probabilities. Another extension is to facilitate

402 the specification of hierarchical binomial and multinomial models which would allow users to

403 analyze data where responses are nested within a higher-order structure such as participants,

404 schools, or countries. Hierarchical multinomial models can be found, for instance, in source

405 memory research where people need to select a previously studied item from a list (e.g.,

406 Arnold, Heck, Bröder, Meiser, & Boywitt, 2019). In addition, we aim to enable the

407 specification of informed hypotheses that are more complex, including hypotheses on the size

408 ratios of the parameters (e.g., $\theta_1 < 2 \times \theta_2$) of interest or the difference between category

409 proportions such that informed hypotheses can also be specified on odds ratios (e.g.,

410 $\frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$).

5   Acknowledgements

6   References

Arnold, N. R., Heck, D. W., Bröder, A., Meiser, T., & Boywitt, C. D. (2019). Testing
        hypotheses about binding in context memory with a hierarchical multinomial
        modeling approach. *Experimental Psychology*, *66*, 239–251.

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R
        Markdown.* Retrieved from https://github.com/crsh/papaja

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American
        Philosophical Society*, *78*, 551–572.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data.
        Journal of Computational Physics, *22*, 245–268.

Damien, P., & Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities.
        Journal of Computational and Graphical Statistics, *10*, 206–215.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in
        detecting fraud in accounting data. *Journal of Forensic Accounting*, *5*, 17–34.

Epskamp, S., & Nuijten, M. (2014). *Statcheck: Extract statistics from articles and recompute
        p values (R package version 1.0.0.).* Comprehensive R Archive Network. Retrieved
        from https://cran.r-project.org/web/packages/statcheck

European Commision. (2004). *Report by Eurostat on the revision of the Greek government
        deficit and debt figures* [Eurostat Report].
        https://ec.europa.eu/eurostat/web/products-eurostat-news/-/GREECE.

European Commision. (2010). *Report on Greek government deficit and debt statistics*
        [Eurostat Report]. https://ec.europa.eu/eurostat/web/products-eurostat-news/-

439   /COM_2010_REPORT_GREEK.

440   Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet distribution*

441   *and related processes.* Department of Electrical Engineering, University of

442   Washington.

443   Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., & Hothorn, F. S. T. (2020). *Mvtnorm:*

444   *Multivariate normal and t distributions.* Retrieved from

445   http://CRAN.R-project.org/package=mvtnorm

446   Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., . . .

447   Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical*

448   *Psychology*, *81*, 80–97.

449   Gronau, Q. F., Singmann, H., & Wagenmakers, E. (2020). Bridgesampling: An R package for

450   estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*, 1–29.

451   Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing

452   of order constrained hypotheses in structural equation models. *Journal of Statistical*

453   *Computation and Simulation*, *89*, 1526–1553.

454   Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality

455   constrained hypotheses. *Psychological Methods*, *19*, 511–527.

456   Haaf, J. M., Klaassen, F., & Rouder, J. (2019). Capturing ordinal theoretical constraint in

457   psychological science. *PsyArXiv.* Retrieved from

458   https://doi.org/10.31234/osf.io/a4xu9

459   Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality

460   constraints: Overview and improvements of computational methods for Bayesian

461   inference. *Journal of Mathematical Psychology*, *91*, 70–87.

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354–363.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., & Boelen, P. (Eds.). (2008). *Bayesian evaluation of informative hypotheses.* New York: Springer Verlag.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*, 64–72.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society, 31*, 203–222.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica, 59*, 57–69.

Laudy, O. (2006). *Bayesian inequality constrained models for categorical data* (PhD thesis). Utrecht University.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica, 6*, 831–860.

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis, 71*, 448–463.

Mulder, J. (2016). Bayes factors for testing order–constrained hypotheses on correlations. *Journal of Mathematical Psychology, 72*, 104–115.

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., . . .
          van Lissa, C. (in press). BFpack: Flexible Bayes factor testing of scientific theories in
          R. *Journal of Statistical Software.*

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for
          calculating Bayes factors for inequality and equality constrained models. *Journal of
          Statistical Software*, *46*, 1–39.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H.
          (2009). Bayesian model selection of informative hypotheses for repeated
          measurements. *Journal of Mathematical Psychology*, *53*, 530–546.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers.
          American Journal of Mathematics, *4*, 39–40.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud
          detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.

Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud schemes.
          Managerial Auditing Journal, *34*, 602–622.

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford's law as an aid in analytical
          procedures. *Auditing*, *16*, 52–67.

Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016).
          The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior
          Research Methods*, *48*, 1205–1226.

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for
          generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*,
          3269–3288.

Rauch, B., Göttsche, M., Brähler, G., & Engel, S. (2011). Fact and fiction in
      EU-governmental economic data. *German Economic Review*, *12*, 243–255.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences.
      Psychological Review, *118*, 42–56.

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus
      structural inconsistency of preferences. *Psychological Review*, *119*, 408–416.

Rijkeboer, M., & van den Hout, M. (2008). A psychologists's view on Bayesian evaluation of
      informative hypotheses. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian*
      evaluation of informative hypotheses (pp. 299–309). Berlin: Springer Verlag.

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E., & Marsman, M. (in
      press). Evaluating multinomial order restrictions with bridge sampling. *Psychological*
      Methods.

Sedransk, J., Monahan, J., & Chiu, H. (1985). Bayesian estimation of finite population
      parameters in categorical data models incorporating order restrictions. *Journal of the*
      Royal Statistical Society. Series B (Methodological), *47*, 519–527.

A Transforming an Ordered Probability Vector to the Real Line

523 The bridge sampling routine in **multibridge** uses the multivariate normal distribution

524 as proposal distribution, which requires moving the target distribution $\boldsymbol{\theta}$ to the real line.

525 Crucially, the transformation needs to retain the ordering of the parameters, that is, it needs

526 to take into account the lower bound l_k and the upper bound u_k of each θ_k. To meet these

527 requirements, **multibridge** uses a probit transformation, as proposed in Sarafoglou et al. (in

528 press), and subsequently transforms the elements in $\boldsymbol{\theta}$, moving from its lowest to its highest

529 value. In the binomial model, we move all elements in $\boldsymbol{\theta}$ to the real line and thus construct a

530 new vector $\boldsymbol{y} \in \mathbb{R}^K$. For multinomial models it follows from the sum-to-one constraint that

531 the vector $\boldsymbol{\theta}$ is completely determined by its first $K-1$ elements, where θ_K is defined as

532 $1 - \sum_{k=1}^{K-1} \theta_k$. Hence, for multinomial models we will only consider the first $K-1$ elements of

533 $\boldsymbol{\theta}$ and we will transform them to $K-1$ elements of a new vector $\boldsymbol{y} \in \mathbb{R}^{K-1}$.

534 Let ϕ denote the density of a normal variable with a mean of zero and a variance of

535 one, Φ denote its cumulative density function, and Φ^{-1} denote the inverse cumulative density

536 function. Then for each element θ_k, the transformation is

$$\xi_k = \Phi^{-1}\left(\frac{\theta_k - l_k}{u_k - l_k}\right),$$

537 The inverse transformation is given by

$$\theta_k = (u_k - l_k)\Phi(\xi_k) + l_k.$$

538 To perform the transformations, we need to determine the lower bound l_k and the

539 upper bound u_k of each θ_k. Assuming $\theta_{k-1} < \theta_k$ for $k \in \{2 \cdots, K\}$ the lower bound for any

540 element in $\boldsymbol{\theta}$ is defined as

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \theta_{k-1} & \text{if } 1 < k < K. \end{cases}$$

This definition holds for both binomial models and multinomial models. Differences in these two models appear only when determining the upper bound for each parameter. For binomial models, the upper bound for each θ_k is simply 1. For multinomial models, however, due to the sum-to-one constraint the upper bounds depend on the values of smaller elements as well as on the number of remaining larger elements in $\boldsymbol{\theta}$. To be able to determine the upper bounds, we represent $\boldsymbol{\theta}$ as unit-length stick which we subsequently divide into K elements (Frigyik, Kapila, & Gupta, 2010, p. stan2020). By using this so-called stick-breaking method we can define the upper bound for any θ_k as follows:

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{C1}$$

where $1 - \sum_{i<k} \theta_i$ represents the length of the remaining stick, that is, the proportion of the unit-length stick that has not yet been accounted for in the transformation. The elements in the remaining stick are denoted as ERS, and are computed as follows:

$$ERS = K - 1 + k.$$

The transformations outlined above are suitable only for ordered probability vectors, that is, for informed hypotheses in binomial and multinomial models that only feature inequality constraints. However, when informed hypotheses also feature equality constrained parameters, as well as parameters that are free to vary we need to modify the formula. Specifically, to determine the lower bounds for any θ_k, we need to take into account how

557 many parameters were set equal to it (denoted as e_k) and how many parameters were set

558 equal to its preceding value θ_{k-1} (denoted as e_{k-1}):

$$
l_k = \begin{cases} 0 & \text{if } k = 1 \\[2ex] \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K. \end{cases} \tag{C2}
$$

559 The upper bound for parameters in the binomial models still remains 1. To determine the

560 upper bound for multinomial models we must, additionally for each element θ_k, take into

561 account the number of free parameters that share common upper and lower bounds (denoted

562 with f_k). The upper bound is then defined as:

$$
u_k = \begin{cases} \dfrac{1 - (f_k \times l_k)}{K} = \dfrac{1}{K} & \text{if } k = 1 \\[3ex] \left(\dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k \geq \max(\theta_{i<k}), \\[3ex] \left(2 \times \left(\dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} \right) - \max(\theta_{i<k}) \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k < \max(\theta_{i<k}). \end{cases}
$$

$$\tag{C3}$$

563   The elements in the remaining stick are then computed as follows

$$
ERS = e_k + \sum_{j>k} e_j \times f_j.
$$

564 The rationale behind these modifications will be described in more detail in the following

565 sections. In **multibridge**, information that is relevant for the transformation of the

566 parameter vectors is stored in the generated `restriction_list` which is returned by the

567 main functions `binom_bf_informed` and `mult_bf_informed` but can also be generated

568 separately with the function `generate_restriction_list`. This restriction list features the

569 sublist `inequality_constraints` which encodes the number of equality constraints

570 collapsed in each parameter in `nr_mult_equal`. Similarly the number of free parameters

571 that share common bounds are encoded under `nr_mult_free`.

A.1 Equality Constrained Parameters

573   In cases where informed hypotheses feature a mix of equality and inequality

574 constrained parameters, we compute the Bayes factor BF_{re}, by multiplying the individual

575 Bayes factors for both constraint types with each other:

$$\text{BF}_{re} = \text{BF}_{1e} \times \text{BF}_{2e} \mid \text{BF}_{1e},$$

576 where the subscript 1 denotes the hypothesis that only features equality constraints and the

577 subscript 2 denotes the hypothesis that only features inequality constraints. To receive

578 $\text{BF}_{2e} \mid \text{BF}_{1e}$, we collapse all equality constrained parameters in the constrained prior and

579 posterior distributions into one category. This collapse has implications on the performed

580 transformations.

581   When transforming the samples from the collapsed distributions, we need to account

582 for the fact that the inequality constraints imposed under the original parameter values

583 might not hold for the collapsed parameters. Consider, for instance, a multinomial model in

584 which we specify the following informed hypothesis

$$\mathcal{H}_r : \theta_1 < \theta_2 = \theta_3 = \theta_4 < \theta_5 < \theta_6,$$

where samples from the encompassing distribution take the values
$(0.05, 0.15, 0.15, 0.15, 0.23, 0.27)$. For these parameter values the inequality constraints hold
since 0.05 is smaller than 0.15, 0.23, and 0.27. However, the same constraint does not hold
when we collapse the categories θ_2, θ_3, and θ_4 into θ_*. That is, the collapsed parameter

$\theta_* = 0.15 + 0.15 + 0.15 = 0.45$ is now larger than 0.23 and 0.27. In general, to determine the lower bound for a given parameter θ_k we thus need to take into account both the number of collapsed categories in the preceding parameter e_{k-1} as well as the number of collapsed categories in the current parameter e_k. Thus, lower bounds for the parameters need to be adjusted as follows:

$$
l_k =
\begin{cases}
0 & \text{if } k = 1 \\
\frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K,
\end{cases}
$$

which leads to Equation C2. In this equation, e_{k-1} and e_k refer to the number of equality constrained parameters that are collapsed in θ_{k-1} and θ_k, respectively. In the example above, this means that to determine the lower bound for θ_* we multiply the preceding value θ_1 by three, such that the lower bound is $\left(\frac{0.05}{1}\right) \times 3 = 0.15$. In addition, to determine the lower bound of θ_5 we divide the preceding value θ_* by three, that is, $\left(\frac{0.45}{3}\right) \times 1 = 0.15$. Similarly, to determine the upper bound for a given parameter value θ_k, we need to multiple the upper bound by the number of parameters that are collapsed within it:

$$
u_k =
\begin{cases}
\dfrac{1}{ERS} \times e_k & \text{if } k = 1 \\
\dfrac{1 - \sum_{i<k} \theta_i}{ERS} \times e_k & \text{if } 1 < k < K,
\end{cases}
\tag{C4}
$$

where $1 - \sum_{i<k} \theta_i$ represents the length of the remaining stick and the number of elements in the remaining stick are computed as follows: $ERS = \sum_k^K e_k$. For the example above, the upper bound for θ_* is $\dfrac{1 - 0.05}{5} \times 3 = 0.57$. The upper bound for θ_5 is then $\dfrac{(1 - 0.05 - 0.45)}{2} \times 1 = 0.25$.

A.2 Corrections for Free Parameters

Different adjustments are required for a sequence of inequality constrained parameters that share upper and lower bounds. Consider, for instance, a multinomial model in which we specify the informed hypothesis

$$\mathcal{H}_r : \theta_1 < (\theta_2\,,\,\theta_3) < \theta_4.$$

This hypothesis specifies that θ_2 and θ_3 have the shared lower bound θ_1 and the shared upper bound θ_4, however, θ_2 can be larger than θ_3 or vice versa. To integrate these cases within the stick-breaking approach one must account for these potential changes of order. For these cases, the lower bounds for the parameters remain unchanged. To determine the upper bound for θ_k, we need to subtract from the length of the remaining stick the lower bound from the parameters that are free to vary. However, only those parameters are included in this calculation that have not yet been transformed:

$$u_k = \begin{cases} \dfrac{1 - (f_k \times l_k)}{K} & \text{if } k = 1 \\[2mm] \dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{C5}$$

where f_k represents the number of free parameters that share common bounds with θ_k and that have been not yet been transformed. Here, the number of elements in the remaining stick is defined as the number of all parameters that are larger than θ_k: $ERS = 1 + \sum_{j>k} f_j$. To illustrate this correction, assume that samples from the encompassing distribution take the values $(0.15, 0.29, 0.2, 0.36)$. The upper bound for θ_1 is simply $\frac{1}{4}$. For θ_2, we need to take into account that θ_2 and θ_3 share common bounds. To compute the upper bound for θ_2, we subtract from the length of the remaining stick the lower bound of θ_3:

$$\frac{1 - 0.15 - (1 \times 0.15)}{1 + 1} = 0.35.$$

A further correction is required if a preceding free parameter (i.e., a parameter with common bounds that was transformed already) is larger than the upper bound of the current

parameter. For instance, in our example the upper bound for θ_3 would be

$\dfrac{1 - 0.44 - 0}{1 + 1} = 0.28$, which is smaller than the value of the preceding free parameter, which

was 0.29. If in this case θ_3 would actually take on the value close to its upper bound, for

instance $\theta_3 = 0.275$, then—due to the sum-to-one constraint—θ_4 would violate the constraint

(i.e., $0.15 < (0.29, 0.275) \nless 0.285$). In these cases, the upper bound for the current θ_k needs

to be corrected downwards. To do this, we subtract from the current upper bound the

difference to the largest preceding free parameter. Thus, if $u_k < \max(\theta_{i<k})$, the upper bound

becomes:

$$u_k = u_k - (\max(\theta_{i<k}) - u_k) \tag{C6}$$

$$= 2 \times u_k - \max(\theta_{i<k}). \tag{C7}$$

For our example the corrected upper bound for θ_3 would become $2 \times 0.28 - 0.29 = 0.27$

which secures the proper ordering for the remainder of the parameters. If in this case θ_3

would take on the value close to its upper bound, for instance $\theta_3 = 0.265$, θ_4—due to the

sum-to-one constraint—would take on the value 0.295 which would be in accordance with

the constraint (i.e., $0.15 < (0.29, 0.265) < 0.295$).