

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

Correspondence concerning this article should be addressed to: Alexandra Sarafoglou,
Department of Psychology, PO Box 15906, 1001 NK Amsterdam, The Netherlands, E-mail:
alexandra.sarafoglou@gmail.com

Abstract

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

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

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. Here, again, the standard tests are inadequate. Generally, by specifying informed hypotheses researchers and practitioners are able to “add theoretical expectations to the traditional alternative hypothesis” (Hojtink, Klugkist, & Boelen, 2008, p. 2) and thus test hypotheses that relate more closely to their theories (Haaf, Klaassen, & Rouder, 2019; Rijkeboer & van 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:

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

where the subscript r denotes the informed hypothesis and e denotes the encompassing hypothesis. Several available R packages compute Bayes factors for informed hypotheses. For instance, the package **multinomineq** (Heck & Davis-Stober, 2019) evaluates informed hypotheses for multinomial models as well as models that feature independent binomials. The package **BFpack** (Joris Mulder et al., in press) evaluates informed hypotheses for statistical models such as univariate and multivariate normal linear models, generalized linear models, special cases of linear mixed models, survival models, and relational event models. The package **BAIN** (Gu, Hoijtink, Mulder, & Rosseel, 2019) evaluates informed hypotheses for structural equation models. Outside of R, the Fortran 90 program **BIEMS** (Joris Mulder, Hoijtink, & de Leeuw, 2012) evaluates informed hypotheses for multivariate linear models such as MANOVA, repeated measures, and multivariate regression. All these packages rely 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 (Klugkist et al., 2005 ;
 Hoijtink, 2011; Hoijtink et al., 2008) and the conditional encompassing method (Gu,
 Mulder, Deković, & Hoijtink, 2014; Laudy, 2006; Joris Mulder, 2014; J. Mulder, 2016; J.
 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 [as well](#)
[as mixtures between equality and inequality constraints](#).

This paper showcases how the proposed bridge sampling routine by
[@sarafoglou2020evaluatingPreprint](#) can be applied in a user-friendly way with
multibridge. In the remainder of this article, we will introduce the package and its
 functionalities and describe the methods used to compute the informed hypotheses in
 binomial and multinomial models. Then we will illustrate its core functions using three
 examples and end with a brief summary.

Multibridge

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

93 `binom_bf_informed`, return the Bayes factor estimate in favor of or against the informed
94 hypothesis (see Table 2 for a summary of the basic required arguments of the two core
95 functions). In addition, users can visualize the posterior parameter estimates under the
96 encompassing hypothesis using the `plot`-method, or get more detailed information on how
97 the Bayes factor is composed using the `summary`-method. For hypotheses that include
98 mixtures between equality and inequality constrained hypotheses the `bayes_factor`
99 method separately returns the Bayes factor for the equality constraints and the conditional
100 Bayes factor for the inequality constraints given the equality constraints. The informed
101 hypothesis can be conveniently specified using a string or character vector. Furthermore,
102 the transitivity property of Bayes factors can be used to test two informed hypotheses
103 against each other (see Example 1 for an illustration). The general workflow of
104 **multibridge** is illustrated in Figure 1. A list of all currently available functions and data
105 sets is given in Table 1.

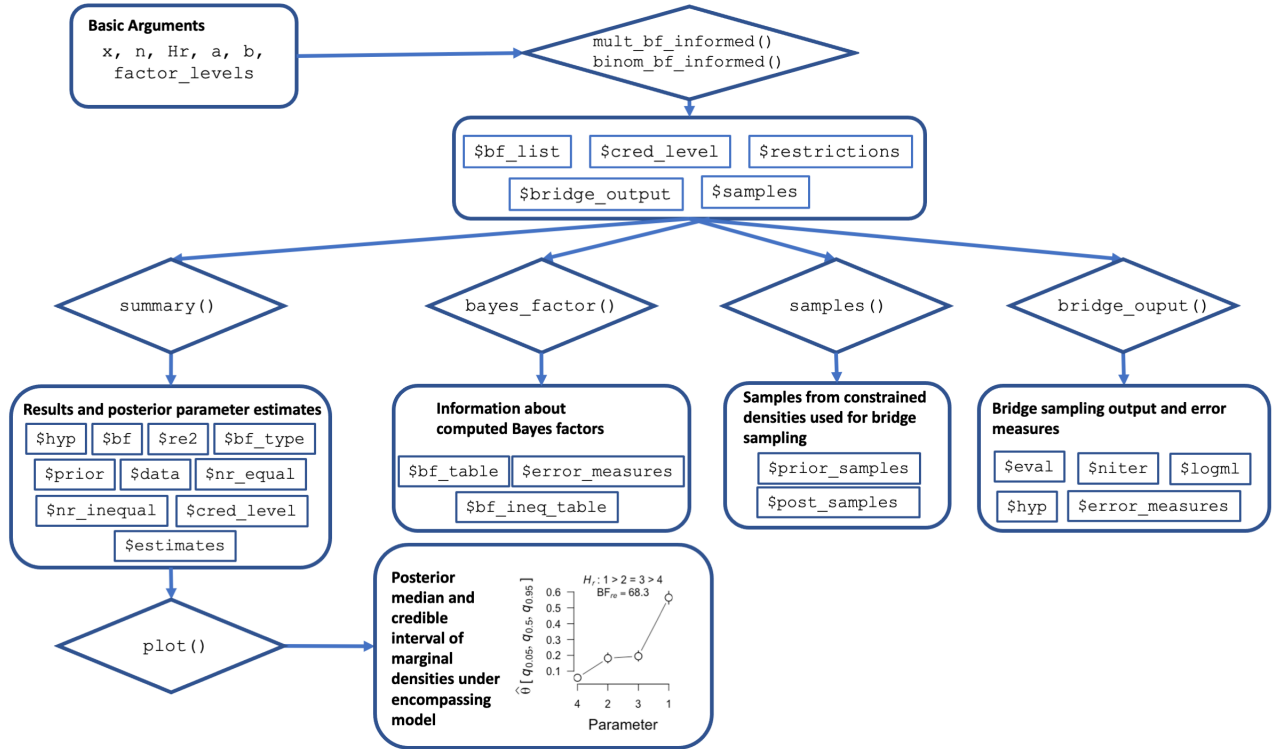


Figure 1. The **multibridge** workflow. When calling `mult_bf_informed` or `binom_bf_informed`, the user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and β parameters of the binomial prior distributions (`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

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

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

106

To compute a Bayes factor, the core functions require the observed counts, the

107

informed hypothesis, the parameters of the prior distribution under \mathcal{H}_e , and the category

108 labels. An overview of these arguments are provided in Table 2.

Table 2

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
<code>x</code>	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).
<code>n</code>	numeric. Vector with counts of trials. Must be the same length as <code>x</code> . Ignored if <code>x</code> is a matrix or a table. Included only in <code>binom_bf_informed</code> .
<code>Hr</code>	string or character. String or vector with the user specified informed hypothesis. Parameters may be referenced by the specified <code>factor_levels</code> or by numerical indices.
<code>a</code>	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 <code>x</code> . Default sets all parameters to 1.
<code>b</code>	numeric. Vector with β parameters. Must be the same length as <code>x</code> . Default sets all β parameters to 1. Included only in <code>binom_bf_informed</code> .
<code>factor_levels</code>	character. Vector with category labels. Must be the same length as <code>x</code> .

109 The package also includes S3 methods that, among other things, summarize the
110 results, plot the parameter estimates under \mathcal{H}_e , or extract the Bayes factors. Table 3
111 summarizes all S3 methods currently available in **multibridge**.

Table 3

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

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

Methodological Background

In this section we provide background information on the methods implemented in **multibridge**. Interested readers can find the mathematical details of the methods in @sarafoglou2020evaluatingPreprint and @gronau2017tutorial. Users who are mainly interested in the application of the package can safely skip this section. In this section we formalize multinomial models and models that feature independent binomial probabilities as they have been implemented in **multibridge**. In the multinomial model, we assume that the vector of observations \mathbf{x} in the K categories follows a multinomial distribution in which the parameters of interest, $\boldsymbol{\theta}$, represent the underlying category proportions. Since the K categories are dependent, the vector of probability parameters is constrained to sum to one, such that $\sum_{k=1}^K (\theta_1, \dots, \theta_K) = 1$. Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is the Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha}$:

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

$$\theta_1, \dots, \theta_K \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K), \quad (2)$$

where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* category counts. The formalization of the model for independent binomial probabilities is similar since the multinomial model above constitutes a generalization of the binomial model (for $K \geq 2$). In the binomial model, we assume that the elements in the vector of successes \mathbf{x} and the elements in the vector of total number of observations \mathbf{n} in the K categories follow independent binomial distributions. As in the multinomial model, the parameter vector of the binomial success probabilities $\boldsymbol{\theta}$ contains the underlying category proportions, however, in this model we assume that categories are independent which removes the sum-to-one constraint. Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is a vector of independent beta distributions with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

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

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

where α can be interpreted as vector of *a priori* successes that observations fall within the various categories and β can be interpreted as vector of *a priori* failures.

Bayes factor

multibridge features two different methods to compute Bayes factors: one method computes Bayes factors for equality constrained parameters and one method computes Bayes factors for inequality constrained parameters. Both methods will be outlined below. In cases where informed hypotheses feature mixtures between inequality and equality constraints, we compute the overall Bayes factor BF_{re} by multiplying the individual Bayes factors for both constraint types. This is motivated by the fact that the Bayes factor for mixtures will factor into a Bayes factor for the equality constraints and a conditional Bayes factor for the inequality constraints given the equality constraints (see Sarafoglou et al., in press, for the proof).

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:

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

146 Note that **multibridge** only supports the specification of one predicted value for all
 147 binomial probabilities. The package does not support the specification of different
 148 predicted values for different binomial probabilities. The reason for this is theoretical: we
 149 believe that such hypotheses are better tested using a hierarchical structure (thus modeling
 150 the binomial 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:

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

151 where θ_{0k} represent the predicted category proportions. When all category proportions are
 152 assumed to be exactly equal all θ_{0k} are set to $\frac{1}{K}$. Otherwise, $\boldsymbol{\theta}_0$ is replaced with the
 153 user-specified predicted values.

154 **The Bayes Factor For Inequality Constraints.** To approximate the Bayes
 155 factor for informed hypotheses, Klugkist et al. (2005) derived an identity that defines the
 156 Bayes factor BF_{re} as the ratio of proportions of posterior and prior parameter space
 157 consistent with the restriction. This identity forms the basis of the encompassing prior
 158 approach. Recently, Sarafoglou et al. (in press) highlighted that these proportions can be
 159 reinterpreted as the marginal likelihoods (i.e., the normalizing constants) of the constrained
 160 posterior and constrained prior distribution:

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

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., Quentin F. 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** (Q. F. 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:

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

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

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.

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 prior distribution with samples from the proposal distribution. The denominator evaluates the normalized proposal distribution with samples from the constrained prior distribution. Using this identity, we obtain the bridge sampling estimator for the marginal likelihood of the constrained prior distribution by applying the iterative scheme proposed by Meng and Wong (1996):

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

where N_1 denotes the number of samples drawn from the constrained distribution, that is,
 $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$, N_2 denotes the number of samples drawn from the proposal distribution,
that 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:

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

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

where $\boldsymbol{\xi}_n^* = \Phi^{-1} \left(\frac{\boldsymbol{\theta}_n^* - \mathbf{l}}{\mathbf{u} - \mathbf{l}} \right)$, and $\tilde{\boldsymbol{\theta}}_m = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_m) + \mathbf{l}) \mid J|$. 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 Quentin F. Gronau et al. (2017) is reached, that is:

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

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 associated relative mean square error, the number of iterations, and the quantities $q_{1,1}$, $q_{1,2}$, $q_{2,1}$, and $q_{2,2}$.

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. 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)
```

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 formalized as follows: for any given leading digit $d, d = (1, \dots, 9)$ the expected proportion is approximately equal to

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

This means that in an empirical data set, numbers with smaller leading digits are more common than numbers with larger leading digits. Specifically, a number has leading digit 1

in 30.1% of the cases, and leading digit 2 in 17.61% of the cases; leading digit 9 is the least frequent digit with an expected proportion of only 4.58% (see Table 4 for an overview of the expected proportions). Empirical data for which this relationship holds include population sizes, death rates, baseball statistics, atomic weights of elements, and physical constants (Benford, 1938). In contrast, artificially generated data, such as telephone numbers, do in general not obey Benford’s law (Hill, 1995). Given that Benford’s law applies to empirical data but not artificially generated data, a so-called Benford test can be used in fields like accounting and auditing to check for indications for poor data quality (for an overview, see e.g., Durtschi, Hillison, & Pacini, 2004; M. Nigrini, 2012; M. J. Nigrini & Mittermaier, 1997). Data that do not pass the Benford test, should raise audit risk concerns, meaning that it is recommended that they undergo additional follow-up checks (Nigrini, 2019).

Below we discuss four possible Bayesian adaptations of the Benford test. In a first scenario we simply conduct a Bayesian multinomial test in which we test the point-null hypothesis \mathcal{H}_0 which predicts a Benford distribution against the encompassing hypothesis \mathcal{H}_e . In a second scenario we test the null hypothesis against an alternative hypothesis, denoted as \mathcal{H}_{r1} , which predicts a decreasing trend in the proportions of leading digits. The hypothesis \mathcal{H}_{r1} exerts considerably more constraint than \mathcal{H}_e and provides a more sensitive test if our primary goal is to test whether data comply with Benford’s law or whether the data follow a similar but different trend. In the next two scenarios, our main goal is to identify fabricated data. The third scenario therefore tests the null hypothesis against the hypothesis that all proportions occur equally often. This hypothesis \mathcal{H}_{r2} could be considered if it is suspected that the data were generated randomly. In a fourth scenario we test the null hypothesis against a hypothesis which predicts a trend that is characteristic for manipulated data. This hypothesis, which we denote as \mathcal{H}_{r3} , could be derived from empirical research on fraud or be based on observed patterns from former fraud cases. For instance, Hill (1995) instructed students to produce a series of random numbers; in the resulting data the proportion of the leading digit 1 occurred most often and the digits 8

and 9 occurred least often which is consistent with the general pattern of Benford’s law. However, the proportion for the remaining leading digits were approximately equal. Note that the predicted distribution derived from Hill (1995) is not currently used as a test to detect fraud. However, for the sake of simplicity, if we assume that this pattern could be an indication of manipulated auditing data, the Bayes factor BF_{0r3} would quantify the evidence of whether the proportion of first digits resemble authentic or fabricated data.

Data and Hypothesis. The data we use to illustrate the computation of Bayes factors were originally published by the European statistics agency “Eurostat” and served as basis for reviewing the adherence to the Stability and Growth Pact of EU member states. Rauch, Göttzsche, Brähler, and Engel (2011) conducted a Benford test on data related to budget deficit criteria, that is, public deficit, public dept and gross national products. The data used for this example features the proportion of first digits from Greek fiscal data in the years between 1999 and 2010; a total of $N = 1,497$ numerical data were included in the analysis. We choose this data, since the Greek government deficit and debt statistics states has been repeatedly criticized by the European Commission in this time span (European Commision, 2004, 2010). In particular, the commission has accused the Greek statistical authorities to have misreported deficit and debt statistics. For further details on the data set see Rauch et al. (2011). The observed and expected proportions are displayed in Table 4; the expected proportions versus the posterior parameter estimates under the encompassing 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

281 In this example, the parameter vector of the multinomial model, $\theta_1, \dots, \theta_K$, reflects
 282 the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9.
 283 The hypotheses introduced above can then be formalized as follows. The null hypothesis
 284 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:

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

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

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

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

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

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

289 An overview of the relative plausibility of all $M = 5$ models simultaneously may be
 290 obtaining by presenting the posterior model probabilities $p(\mathcal{H}_i | x)$ (Berger & Molina,
 291 2005). Denoting the prior model probability for model \mathcal{H}_i by $p(\mathcal{H}_i)$, the posterior model
 292 probability for \mathcal{H}_0 is given by:

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

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

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

294 **Method.** Both BF_{0e} and BF_{r2e} may be readily computed by means of a Bayesian
 295 multinomial test which is implemented in the function `mult_bf_equality`. This function
 296 requires (1) a vector with observed counts, (2) a vector with concentration parameters of

the Dirichlet prior distribution under \mathcal{H}_e , and (3) the vector of expected proportions under \mathcal{H}_0 and under \mathcal{H}_{r2} . We do not incorporate specific expectations about the distribution of leading digits in the Greek fiscal data and therefore set all concentration parameters under \mathcal{H}_e to 1 (i.e., we assign θ a uniform Dirichlet prior distribution).

```
# 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)
```

```
## Warning in .checkProbability(p = p, x = x): Parameters have been rescaled.
```

```
logBFfe0 <- results_H0_He$bf$LogBFfe0
logBFer2 <- results_Hr2_He$bf$LogBFfe0
```

The hypotheses \mathcal{H}_{r1} and \mathcal{H}_{r3} contain inequality constraints, and this necessitates the use of the function `mult_bf_informed` to compute the Bayes factors BF_{r1e} and BF_{r3e} . This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under \mathcal{H}_e , (3) labels for the categories of interest (i.e., leading digits), and (4) the informed hypothesis \mathcal{H}_{r1} or \mathcal{H}_{r3} (e.g., as a string). In addition to the basic required arguments, we use two additional arguments here. The first argument sets the Bayes factor type, that is, whether the output should print the

309 Bayes factor in favor of the informed hypothesis (i.e., BF_{re}) or in favor of the encompassing
 310 hypothesis (i.e., BF_{er}). It is also possible to compute the log Bayes factor in favor of the
 311 hypothesis, which is the setting we choose for this example. The purpose of the second
 312 argument `seed` is to make the results reproducible:

```
# 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)

# 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)
```

313 ## Warning in .checkProbability(p = p, x = x): Parameters have been rescaled.

```
logBFer3 <- summary(results_He_Hr3)$bf
```

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

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

316 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	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mathbf{x})$	$\log(\text{BF}_{.0})$
\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	5.97×10^{-79}	-155.03
\mathcal{H}_{r2}	0.2	2.71×10^{-212}	-462.06

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

326 To summarize, the data offer overwhelming support for hypothesis \mathcal{H}_{r1} , which
 327 postulates a decreasing trend in the digit proportions. This model outperformed both
 328 simpler models (e.g., the Benford model) and a more complex model in which the
 329 proportions were free to vary. Detailed follow-up analyses are needed to discover why the
 330 data follow a monotonically decreasing pattern but not any of the two specific patterns
 331 that were put to the test (M. J. Nigrini, 2019).

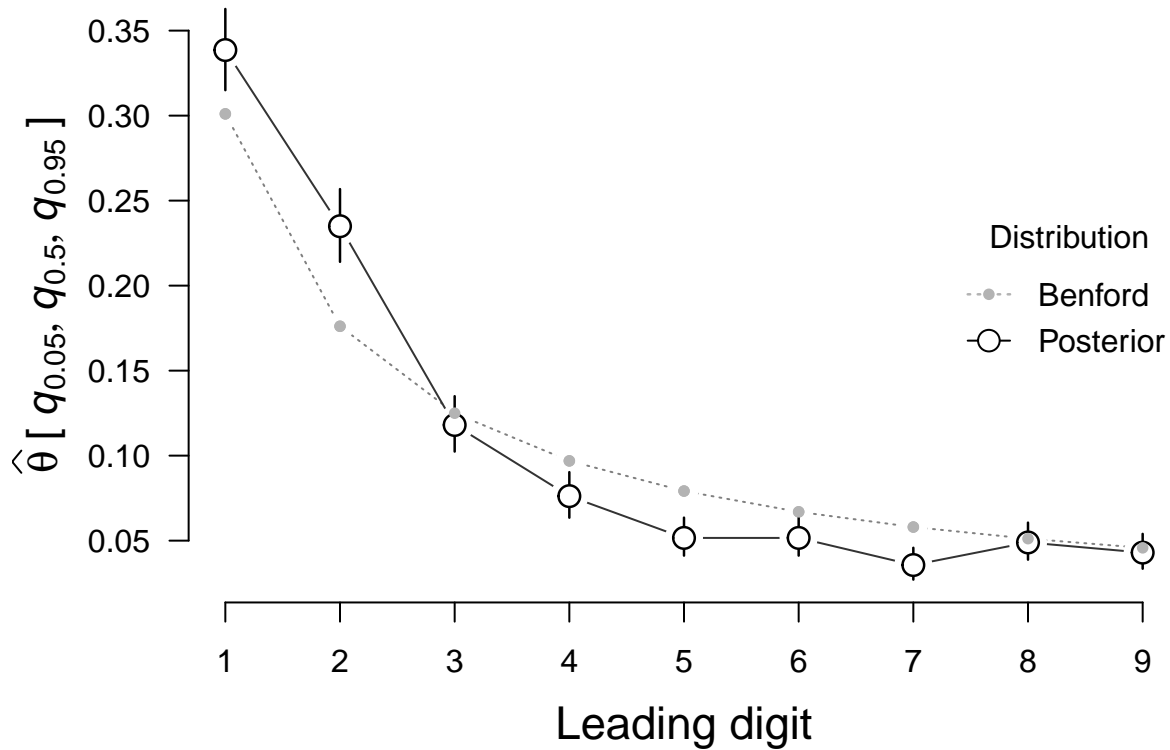


Figure 2. Predictions from Benford’s law (in grey) show together with the posterior medians (black circles) for the category proportions estimated under the encompassing model \mathcal{H}_e . The circle skewers show the 95% credible intervals. Only three of nine intervals encompass the expected proportions, 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**.

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 inconsistencies. This package allowed Nuijten et al. (2016) to estimate the prevalence of statistical reporting errors in the field of psychology. In total, the authors investigated a sample of 30,717 articles (which translates to over a quarter of a million p -values) published in eight major psychology journals between 1985 to 2013: *Developmental Psychology* (DP), the *Frontiers in Psychology* (FP), the *Journal of Applied Psychology* (JAP), the *Journal of Consulting and Clinical Psychology* (JCCP), *Journal of Experimental Psychology: General* (JEPG), the *Journal of Personality and Social Psychology* (JPSP), the *Public Library of Science* (PLoS), *Psychological Science* (PS).

Based on several background assumptions, Nuijten et al. (2016) predicted that the proportion of statistical reporting errors is higher for articles published in the *Journal of Personality and Social Psychology* (JPSP) than for articles published in the seven other journals.

Data and Hypothesis. Here we reuse the original data published by Nuijten et al. (2016), which we also distribute with the package `multibridge` under the name `journals`.

```
data(journals)
```

The Nuijten et al. (2016) hypothesis of interest, \mathcal{H}_r , states that the prevalence for statistical reporting errors is higher for JPSP than for the other journals.² We will consider two specific versions of the Nuijten et al. (2016) \mathcal{H}_r hypothesis. The first hypothesis, \mathcal{H}_{r1} , stipulates that JPSP has the highest prevalence of reporting inconsistencies, whereas the other seven journals share a prevalence that is lower. The second hypothesis, \mathcal{H}_{r2} , also stipulates that JPSP has the highest prevalence of reporting inconsistencies, but does not 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.

The **multibridge** package can be used to test \mathcal{H}_{r1} and \mathcal{H}_{r2} against the null hypothesis \mathcal{H}_0 that all eight journals have the same prevalence of statistical reporting errors. In addition, we will compare \mathcal{H}_{r1} , \mathcal{H}_{r2} , and \mathcal{H}_0 against the encompassing hypothesis \mathcal{H}_e that makes no commitment whatsoever about the prevalence of reporting inconsistencies across the eight journals. In this example, the parameter vector of the binomial success probabilities, θ , reflects the probabilities that articles contain at least one statistical reporting inconsistency across journals. Thus, the above hypotheses can be formalized as follows:

$$\begin{aligned}\mathcal{H}_0 : \theta_{\text{JAP}} &= \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}} = \theta_{\text{JPSP}} \\ \mathcal{H}_{r1} : (\theta_{\text{JAP}} &= \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}}) < \theta_{\text{JPSP}} \\ \mathcal{H}_{r2} : (\theta_{\text{JAP}}, \theta_{\text{PS}}, \theta_{\text{JCCP}}, \theta_{\text{PLOS}}, \theta_{\text{DP}}, \theta_{\text{FP}}, \theta_{\text{JEPG}}) &< \theta_{\text{JPSP}} \\ \mathcal{H}_e : \theta_{\text{JAP}} \cdots \theta_{\text{JPSP}} &\sim \prod_{k=1}^K \text{Beta}(\alpha_k, \beta_k).\end{aligned}$$

Method. To compute the Bayes factor BF_{0r} we need to specify (1) a vector with observed successes (i.e., the number of articles that contain a statistical inconsistency), (2) a vector containing the total number of observations (i.e., the number of articles), (3) a vector with prior parameter α_k for each binomial proportion of the beta prior distribution under \mathcal{H}_e , (4) a vector with prior parameter β_k for each binomial proportion of the beta prior distribution under \mathcal{H}_e , (5) the category labels (i.e., journal names), and (6) the informed hypothesis \mathcal{H}_{r1} or \mathcal{H}_{r2} (e.g., as a string). We also change the Bayes factor type to **LogBFr0** so that the function returns the log Bayes factor in favor for the informed hypothesis compared to the null hypothesis. Since we have no specific expectations about the distribution of statistical reporting errors in any given journal, we set all parameters α_k and β_k to one which corresponds to uniform beta distributions. With this 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$articles_with_NHST *
           (journals$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)

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	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mathbf{x})$	$\log(\text{BF}_{.0})$
\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 $\log(\text{BF}_{r20})$ 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; $\log(\text{BF}_{r20}) = 158.28$.

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.

```
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 $\log(\text{BF}_{e0}) = 156.27$, which confirms

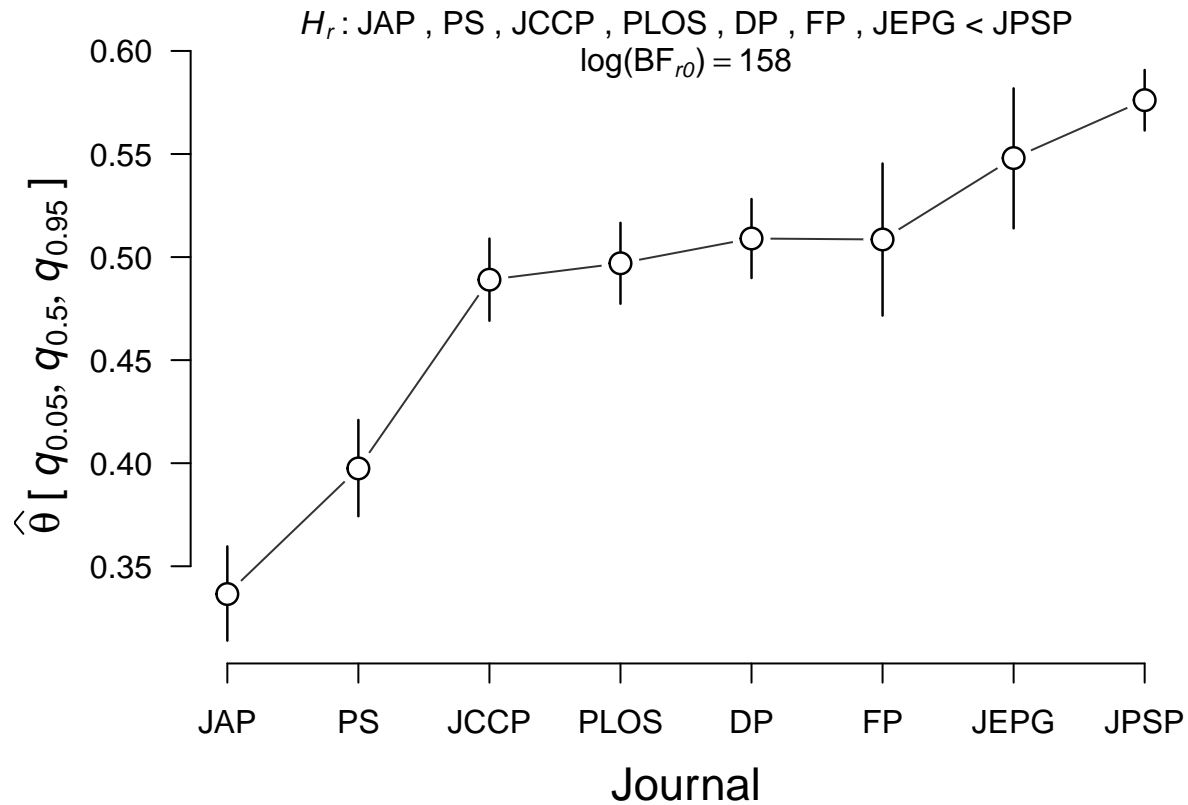


Figure 3. Posterior medians for the prevalence of statistical reporting inconsistencies across eight psychology journals, as obtained using the encompassing model. The circle skewers show the 95% credible intervals. Analysis based on data from Nuijten et al. (2016). This plot was created using the `plot-S3-method`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

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 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 higher than that in seven other psychology journals.

Summary

The R package **multibridge** facilitates the estimation of Bayes factors for informed hypotheses in both multinomial and independent binomial models. The efficiency gains of **multibridge** are particularly pronounced when the parameter restrictions are highly informative or when the number of categories is large.

multibridge supports the evaluation of informed hypotheses that feature equality constraints, inequality constraints, and free parameters, as well as mixtures between them. Moreover, users can choose to test the informative hypothesis against an encompassing hypothesis that lets all parameters vary freely or against the null hypothesis that states that category proportions are exactly equal. Beyond the core functions currently implemented in **multibridge**, there are several natural extensions we aim to include in future versions of this package. For instance, to compare several models with each other we plan to implement functions that compute the posterior model probabilities. Another extension is to facilitate the specification of hierarchical binomial and multinomial models which would allow users to analyze data where responses are nested within a higher-order structure such as participants, schools, or countries. Hierarchical multinomial models can be found, for instance, in source memory research where people need to select a previously studied item from a list (e.g., Arnold, Heck, Bröder, Meiser, & Boywitt, 2019). In addition, we aim to enable the specification of informed hypotheses that are more complex, including hypotheses on the size ratios of the parameters (e.g., $\theta_1 < 2 \times \theta_2$) of interest or the

difference between category proportions such that informed hypotheses can also be specified on odds ratios (e.g., $\frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$).

Declarations

Availability of data and code

The source code of the R package is available at:
<https://github.com/ASarafoglou/multibridge/>. In addition, readers can access the code for reproducing all analyses and plots via our project folder on the Open Science Framework:
<https://osf.io/2wf5y/>.

Funding

This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to AS (406-17-568), a Veni grant from the NWO to MM (451-17-017), a Vici grant from the NWO to EJW (016.Vici.170.083), as well as a a European Research Council (ERC) grant to EJW (283876). This paper was written in Rmarkdown, using the R package **papaja** (Aust & Barth, 2020).

Author contributions

The authors made the following contributions. Alexandra Sarafoglou: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Validation, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Frederik Aust: Conceptualization, Software, Supervision, Validation, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Maarten Marsman: Funding Acquisition, Conceptualization, Methodology, Supervision, Validation, Writing - Review & Editing; Eric-Jan Wagenmakers: Funding Acquisition, Methodology, Supervision, Validation, Writing - Review & Editing;

449 Julia M. Haaf: Conceptualization, Formal Analysis, Methodology, Software,
450 Supervision,15Validation, Writing - Original Draft Preparation, Writing - Review &
451 Editing.

452 **Conflicts of interest**

453 The authors declare that there were no conflicts of interest with respect to the
454 authorship or the publication of this article.

455 **Ethical Approval**

456 This is a methodological contribution which requires no ethical approval.

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/-/COM_2010_REPORT_GREEK.
- Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet distribution and related processes*. Department of Electrical Engineering, University of Washington.

- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., & Hothorn, F. S. T. (2020).
Mvtnorm: Multivariate normal and t distributions. Retrieved from
<http://CRAN.R-project.org/package=mvtnorm>
- Gronau, Quentin F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M.,
... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of*
Mathematical Psychology, 81, 80–97.
- Gronau, Q. F., Singmann, H., & Wagenmakers, E. –J. (2020). Bridgesampling: An
R package for estimating normalizing constants. *Journal of Statistical Software,*
Articles, 92, 1–29.
- Gu, X., Hoijsink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for
Bayesian testing of order constrained hypotheses in structural equation models.
Journal of Statistical Computation and Simulation, 89, 1526–1553.
- Gu, X., Mulder, J., Deković, M., & Hoijsink, H. (2014). Bayesian evaluation of
inequality constrained hypotheses. *Psychological Methods*, 19, 511–527.
- Haaf, J. M., Klaassen, F., & Rouder, J. (2019). Capturing ordinal theoretical
constraint in psychological science. *PsyArXiv*. Retrieved from
<https://doi.org/10.31234/osf.io/a4xu9>
- Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear
inequality constraints: Overview and improvements of computational methods
for Bayesian 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.
- Hoijsink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and*
social scientists. Boca Raton, FL: Chapman & Hall/CRC.
- Hoijsink, 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, Joris. (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, Joris, 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, Joris, 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,*

- 538 *and fraud detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.
- 539 Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud
540 schemes. *Managerial Auditing Journal*, 34, 602–622.
- 541 Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford’s law as an aid in
542 analytical procedures. *Auditing*, 16, 52–67.
- 543 Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J.
544 M. (2016). The prevalence of statistical reporting errors in psychology
545 (1985–2013). *Behavior Research Methods*, 48, 1205–1226.
- 546 Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination
547 methods for generalised linear mixed models. *Computational Statistics & Data*
548 *Analysis*, 54, 3269–3288.
- 549 Rauch, B., Göttische, M., Brähler, G., & Engel, S. (2011). Fact and fiction in
550 EU-governmental economic data. *German Economic Review*, 12, 243–255.
- 551 Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of
552 preferences. *Psychological Review*, 118, 42–56.
- 553 Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices
554 versus structural inconsistency of preferences. *Psychological Review*, 119,
555 408–416.
- 556 Rijkeboer, M., & van den Hout, M. (2008). A psychologists’s view on Bayesian
557 evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. A. Boelen
558 (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 299–309). Berlin:
559 Springer Verlag.
- 560 Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E. –J., &
561 Marsman, M. (in press). Evaluating multinomial order restrictions with bridge
562 sampling. *Psychological Methods*.
- 563 Sedransk, J., Monahan, J., & Chiu, H. (1985). Bayesian estimation of finite
564 population parameters in categorical data models incorporating order

565 restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*,
566 47, 519–527.

Transforming an Ordered Probability Vector to the Real Line

The bridge sampling routine in **multibridge** uses the multivariate normal distribution as proposal distribution, which requires moving the target distribution $\boldsymbol{\theta}$ to the real line. Crucially, the transformation needs to retain the ordering of the parameters, that is, it needs to take into account the lower bound l_k and the upper bound u_k of each θ_k . To meet these requirements, **multibridge** uses a probit transformation, as proposed in Sarafoglou et al. (in press), and subsequently transforms the elements in $\boldsymbol{\theta}$, moving from its lowest to its highest value. In the binomial model, we move all elements in $\boldsymbol{\theta}$ to the real line and thus construct a new vector $\mathbf{y} \in \mathbb{R}^K$. For multinomial models it follows from the sum-to-one constraint that the vector $\boldsymbol{\theta}$ is completely determined by its first $K - 1$ elements, where θ_K is defined as $1 - \sum_{k=1}^{K-1} \theta_k$. Hence, for multinomial models we will only consider the first $K - 1$ elements of $\boldsymbol{\theta}$ and we will transform them to $K - 1$ elements of a new vector $\mathbf{y} \in \mathbb{R}^{K-1}$.

Let ϕ denote the density of a normal variable with a mean of zero and a variance of one, Φ denote its cumulative density function, and Φ^{-1} denote the inverse cumulative density function. Then for each element θ_k , the transformation is

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

The inverse transformation is given by

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

To perform the transformations, we need to determine the lower bound l_k and the upper bound u_k of each θ_k . Assuming $\theta_{k-1} < \theta_k$ for $k \in \{2 \dots, K\}$ the lower bound for any element in $\boldsymbol{\theta}$ is defined as

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

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

$$u_k = \begin{cases} \frac{1}{K} & \text{if } k = 1 \\ \frac{1 - \sum_{i < k} \theta_i}{ERS} & \text{if } 1 < k < K, \end{cases} \quad (\text{C1})$$

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

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

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

account how many parameters were set equal to it (denoted as e_k) and how many parameters were set equal to its preceding value θ_{k-1} (denoted as e_{k-1}):

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K. \end{cases} \quad (\text{C2})$$

The upper bound for parameters in the binomial models still remains 1. To determine the upper bound for multinomial models we must, additionally for each element θ_k , take into account the number of free parameters that share common upper and lower bounds (denoted with f_k). The upper bound is then defined as:

$$u_k = \begin{cases} \frac{1 - (f_k \times l_k)}{K} = \frac{1}{K} & \text{if } k = 1 \\ \left(\frac{1 - \sum_{i < k} \theta_i - (f_k \times l_k)}{ERS} \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k \geq \max(\theta_{i < k}), \\ \left(2 \times \left(\frac{1 - \sum_{i < k} \theta_i - (f_k \times l_k)}{ERS} \right) - \max(\theta_{i < k}) \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k < \max(\theta_{i < k}). \end{cases} \quad (\text{C3})$$

The elements in the remaining stick are then computed as follows

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

The rationale behind these modifications will be described in more detail in the following sections. In **multibridge**, information that is relevant for the transformation of the parameter vectors is stored in the generated **restriction_list** which is returned by the main functions **binom_bf_informed** and **mult_bf_informed** but can also be generated separately with the function **generate_restriction_list**. This restriction list features the sublist **inequality_constraints** which encodes the number of equality constraints

616 collapsed in each parameter in `nr_mult_equal`. Similarly the number of free parameters
 617 that share common bounds are encoded under `nr_mult_free`.

618 Equality Constrained Parameters

619 In cases where informed hypotheses feature a mix of equality and inequality
 620 constrained parameters, we compute the Bayes factor BF_{re} , by multiplying the individual
 621 Bayes factors for both constraint types with each other:

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

622 where the subscript 1 denotes the hypothesis that only features equality constraints and
 623 the subscript 2 denotes the hypothesis that only features inequality constraints. To receive
 624 $\text{BF}_{2e} \mid \text{BF}_{1e}$, we collapse all equality constrained parameters in the constrained prior and
 625 posterior distributions into one category. This collapse has implications on the performed
 626 transformations.

627 When transforming the samples from the collapsed distributions, we need to account
 628 for the fact that the inequality constraints imposed under the original parameter values
 629 might not hold for the collapsed parameters. Consider, for instance, a multinomial model
 630 in 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:

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

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 $\frac{1 - 0.05}{5} \times 3 = 0.57$. The upper bound for θ_5 is then

$$\frac{(1 - 0.05 - 0.45)}{2} \times 1 = 0.25.$$

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

645 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:

$$u_k = \begin{cases} \frac{1 - (f_k \times l_k)}{K} & \text{if } k = 1 \\ \frac{1 - \sum_{i < k} \theta_i - (f_k \times l_k)}{ERS} & \text{if } 1 < k < K, \end{cases} \quad (\text{C5})$$

646 where f_k represents the number of free parameters that share common bounds with θ_k and
647 that have been not yet been transformed. Here, the number of elements in the remaining
648 stick is defined as the number of all parameters that are larger than θ_k :

649 $ERS = 1 + \sum_{j > k} f_j$. To illustrate this correction, assume that samples from the
650 encompassing distribution take the values (0.15, 0.29, 0.2, 0.36). The upper bound for θ_1 is
651 simply $\frac{1}{4}$. For θ_2 , we need to take into account that θ_2 and θ_3 share common bounds. To
652 compute the upper bound for θ_2 , we subtract from the length of the remaining stick the

653 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

$\frac{1 - 0.44 - 0}{1 + 1} = 0.28$, 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., $0.15 < (0.29, 0.275) \not\leq 0.285$). 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:

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

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

654 For our example the corrected upper bound for θ_3 would become $2 \times 0.28 - 0.29 = 0.27$
 655 which secures the proper ordering for the remainder of the parameters. If in this case θ_3
 656 would take on the value close to its upper bound, for instance $\theta_3 = 0.265$, θ_4 —due to the
 657 sum-to-one constraint—would take on the value 0.295 which would be in accordance with
 658 the constraint (i.e., $0.15 < (0.29, 0.265) < 0.295$).