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

Alexandra Sarafoglou[1], Frederik Aust[1], Maarten Marsman[1], Eric-Jan Wagenmakers[1], & Julia M. Haaf[1]

[1] University of Amsterdam

Author Note

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

Abstract

The **multibridge** package efficiently computes Bayes factors for binomial and multinomial models that feature inequality constraints, equality constraints, free parameters and mixtures between them. By using the bridge sampling algorithm to compute the Bayes factor, **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. The package was developed in the R programming language and is freely available from the Comprehensive `R` Archive Network (CRAN). This paper introduces the underlying methodology and illustrates how to use the implementations provided in **multibridge** through fully reproducible examples.

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

1 Introduction

We present **multibridge**, an `R` package to evaluate informed hypotheses in multinomial models and models featuring independent binomials using Bayesian inference. The package allows users to specify for both models informed hypotheses about the underlying category proportions $\boldsymbol{\theta}$. For binomial and multinomial models the following informed hypotheses $\mathcal{H}_r$ on $\boldsymbol{\theta}$ can be tested: (a) hypotheses that postulate equality constraints for (a subset of) parameters (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., $\theta_1 < \theta_2 = \theta_3$); (d) hypotheses that postulate mixtures of inequality constraints and free parameters (e.g., $\theta_1 < \theta_2, \theta_3$); (e) hypotheses that postulate mixtures of (a)–(d) (e.g., $\theta_1 < (\theta_2 = \theta_3), \theta_4$). The informed hypothesis is passed to **multibridge** conveniently using a string or character vector. The user can choose whether the respective Bayes factor should compare the informed hypothesis 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. The package is available from the Comprehensive `R` Archive Network (CRAN) at https://CRAN.R-project.org/package=multibridge.

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 the underlying category proportions are exactly equal, or equal to fixed values. In some cases, it is possible to derive these fixed values from a theory. For instance, Benford's law makes exact predictions about the distribution of leading digits in empirical datasets (Benford, 1938; Newcomb, 1881). Often,

however, this comparison cannot adequately test the predictions researchers are interested in. For instance, the weak-order mixture model of decision-making (Regenwetter & Davis-Stober, 2012) 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 would also prefer A over C (Regenwetter, Dana, & Davis-Stober, 2011). This model provides a precise prediction of behavior. However, one cannot derive fixed values for the choice preferences from the theory of weakly-ordered preference, which makes the comparison between $\mathcal{H}_e$ and $\mathcal{H}_0$ an inadequate test. Instead, the model predictions need to be translated into an informed hypothesis $\mathcal{H}_r$ that stipulates ordinal relations among the parameters. Only then it is possible to adequately test whether the theory of weakly-ordered preference describes participants choice behavior. Theories can also generate more complex predictions, 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 psychological journals while not expressing expectations about the error rate distribution among the latter. Here again it is not possible to apply standard tests, since we cannot derive fixed proportions from the stated hypothesis. Generally, by specifying informed hypotheses researchers and practitioners are able to "add theoretical expectations to the traditional alternative hypothesis" (Hoijtink, 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 test the hypotheses of interest by means of Bayes factors (Jeffreys, 1935; Kass & Raftery, 1995). Bayes factors compare the relative evidence of two hypotheses in the light of the data. It is defined as the ratio of marginal likelihoods of the respective hypotheses. For instance, the Bayes factor for the informed

hypothesis versus encompassing hypothesis is defined as:

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

where the subscript r denotes the informed hypothesis and e denotes the encompassing hypothesis. To compute Bayes factors for informed hypotheses several `R` packages are already available. 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** (Mulder et al., 2020) 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** (Mulder, Hoijtink, Leeuw, & others, 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 (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 has been critiqued for becoming increasingly unreliable and inefficient as the number of restrictions increases or the parameter space of the restricted model decreases (Sarafoglou et al., 2020).

As alternative to the encompassing prior approach, Sarafoglou et al. (2020) 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** 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). This split results in the Bayes factor estimate being more accurate and requiring less time to compute since parts of it are analytically available. When calling the core functions of **multibridge**, that is `mult_bf_informed` and `binom_bf_informed`, they 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 receive a visualization of 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 shows the conditional Bayes factor for the inequality constraints given the equality constraints and a Bayes factor for the equality constraints. The general workflow of **multibridge** is illustrated in Figure 1. Table 2 summarizes all S3 methods currently available in **multibridge**.

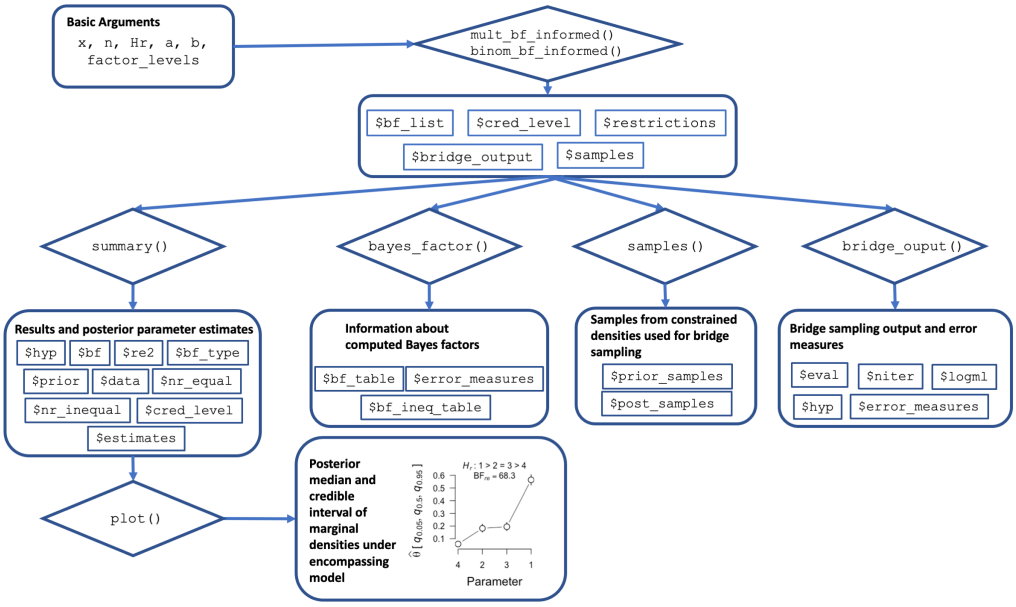

Figure 1. The **multibridge** workflow. The user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and β parameters of the binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions `mult_bf_informed` and `binom_bf_informed` then return an estimate for the Bayes factor of the informed hypothesis versus 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 following 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. Users can either use the specified `factor_levels` or numerical indices to refer to parameters.
`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.

This paper showcases how the proposed bridge sampling routine by Sarafoglou et al. (2020) can be applied in a user-friendly way with **multibridge**. In the remainder of this article, we will describe the Bayes factor identity for informed hypotheses in binomial and multinomial models, and briefly describe the bridge sampling method. Then, we illustrate the core functions of **multibridge** package using two examples and end with a brief summary.

2 Methods

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 $\sum_{k=1}^{K}(\theta_1, \cdots, \theta_K) = 1$. Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is the Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha}$:

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

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

where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* category counts. The formalization of the model for independent binomial probabilities is very 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 $\boldsymbol{\beta}$:

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

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

where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* successes that observations fall within the various categories and $\boldsymbol{\beta}$ can be interpreted as vector of *a priori* failures.

2.1 Bayes factor

In **multibridge** we use 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 with each other. That is, the Bayes factor for mixtures factors into a Bayes factor for the equality constraints, and a conditional Bayes factor for the inequality constraints given the equality constraints (for the proof, see Sarafoglou et al., 2020).

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 $\mathrm{B}(\cdot)$ 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:

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

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

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

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

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

2.1.2 The Bayes Factor For Inequality Constraints.

To approximate the Bayes factor for informed hypotheses, Klugkist et al. (2005) derived an identity that defines the Bayes factor BF_{re} as ratio of proportions of posterior and prior parameter space consistent with the restriction. This identity forms the basis of the encompassing prior approach. Recently, Sarafoglou et al. (2020) highlighted that these proportions can be reinterpreted as the marginal likelihoods of the constrained 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 subsequently 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

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 receive the bridge sampling estimator for the marginal likelihood of the constrained prior distribution by applying the iterative scheme proposed by Meng and Wong (1996):

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

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:

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

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

where ${\boldsymbol{\xi}_{\boldsymbol{n}}}^* = \Phi^{-1}\left(\frac{\boldsymbol{\theta}^*_{\boldsymbol{n}} - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, 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:

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

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 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. In addition, the first example shows how two informed hypotheses can be compared to each other.

A list of all currently available functions and data sets is given in Table 3. Additional examples are available as vignettes (see https://cran.r-project.org/package=multibridge, or `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

formalized as follows: for any given leading digit $d, d = (1, \cdots, 9)$ 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, 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; Nigrini, 2012; 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 three possible Bayesian adaptations of the Benford's 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 constraints 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 a third scenario, where the main goal is to identify fabricated data, we test the null hypothesis against a hypothesis which predicts a trend that is characteristic for manipulated data. This hypothesis, which we denote as $\mathcal{H}_{r2}$, could be

derived from empirical research on fraud or be based on observed patterns from former fraud cases. For instance, Hill (1988) 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 (1988) 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 fabricated auditing data, the Bayes factor would quantify the evidence of whether the proportion of first digits resemble authentic or fabricated data.

3.1.1 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, Göttsche, 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 fiscal data from Greece 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 proportions are displayed in Table 4, the figure displaying the observed versus the expected proportions are displayed in Figure 2.

Table 4

The Table shows the Observed Counts, Observed Proportions, and Expected Proportions of first digits in Greece governmental data. 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. Thus, we can formalize the discussed hypotheses as follows. The null hypothesis specifies that the proportions of first digits obeys Benford's law:

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

Here, we are testing the null hypothesis against the following three alternative

hypotheses:

$$\mathcal{H}_e : \boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}),$$
$$\mathcal{H}_{r1} : \theta_1 > \theta_2 > \theta_3 > \theta_4 > \theta_5 > \theta_6 > \theta_7 > \theta_8 > \theta_9,$$
$$\mathcal{H}_{r2} : \theta_1 > (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7) > (\theta_8,\ \theta_9).$$

We could also compare the three alternative hypothesis directly with each other. To do so, we can make use of the transitivity property of the Bayes factor. For instance, if we would like to compare $\mathcal{H}_{r1}$ with $\mathcal{H}_{r2}$, we would first compute BF_{er1} and BF_{er2} and then yield BF_{r1r2} by dividing the two quantities:

$$\text{BF}_{r1r2} = \frac{\text{BF}_{er2}}{\text{BF}_{er1}}.$$

3.1.2 Method. We can compare $\mathcal{H}_0$ and $\mathcal{H}_e$ by means of a Bayesian multinomial test which is implemented in the function `mult_bf_equality`. To evaluate $\mathcal{H}_0$, we only need to specify (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution, and (3) the vector of proportions expected under the null hypothesis. We do not want to incorporate any specific expectations about the distribution of leading digits in the Greek fiscal data. Hence, we set all concentration parameters to one which corresponds to a uniform Dirichlet distribution.

```
# Observed counts
x <- c(509, 353, 177, 114,  77,  77,  53,  73,  64)
# Concentration parameters
a <-  rep(1, 9)
# Expected proportions
p <- log10((1:9 + 1)/1:9)
# Execute the analysis
```

```
results_H0_He  <- mult_bf_equality(x = x, a = a, p = p)
logBFe0 <- results_H0_He$bf$LogBFe0
```

Since the hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ contain inequality constraints, we use the function `mult_bf_informed` to compute the Bayes factor of the informed hypotheses to the encompassing hypothesis. We then make use of the transitivity property of the Bayes factor to compare the alternative hypotheses to the null hypothesis. In this function, we need to specify (1) a vector with observed counts, (2) the informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r2}$ (e.g., as character vector), (3) a vector with concentration parameters of the Dirichlet prior distribution, and (4) labels for the categories of interest (i.e., leading digits):

```
# Labels for categories of interest
factor_levels <- 1:9
# Specifying the informed Hypothesis
Hr1 <- c('1 > 2 > 3 > 4 > 5 > 6 > 7 > 8 > 9')
Hr2 <- 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)$bf
results_He_Hr2 <- mult_bf_informed(x = x, Hr = Hr2, a = a,
                                   factor_levels = factor_levels,
                                   bf_type = 'LogBFer', seed = 2020)
logBFer2 <- summary(results_He_Hr2)$bf
```

```
bayes_factor_table <- data.frame(
   BFType = c('LogBFe0', 'LogBFr10', 'LogBFr20'),
```

```
   LogBF  = c(logBFe0, -logBFer1 + logBFe0, -logBFer2 + logBFe0))
bayes_factor_table
```

```
##      BFType      LogBF
## 1   LogBFe0   17.6715
## 2 LogBFr10   25.0883
## 3 LogBFr20 -154.5685
```

As the evidence is extreme in all three cases, we report all Bayes factors on the log scale. We can make the following statements concerning the comparison of the null hypothesis to the three alternative hypotheses. The first Bayes factor $\log(\text{BF}_{e0})$ suggests extreme evidence *in favor of* the hypothesis that the first digits vary freely; $\log(\text{BF}_{e0}) = 17.67$. The second Bayes factor $\log(\text{BF}_{r10})$ suggests extreme evidence *in favor of* the hypothesis that the first digits follow a decreasing trend, $\log(\text{BF}_{r10}) = 25.09$. The third Bayes factor $\log(\text{BF}_{r20})$ suggests extreme evidence *against* the hypothesis that the first digits follow a fraudulent pattern with $\log(\text{BF}_{r20}) = -154.57$. When we compare the informed hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ directly with each other, the data show most evidence for a decreasing trend ($\log(\text{BF}_{r1r2}) = 180$).

To summarize, the preferred hypothesis is $\mathcal{H}_{r1}$ that postulates an decreasing trend. The second best performing hypothesis is the encompassing hypothesis $\mathcal{H}_e$, followed by $\mathcal{H}_0$ that postulates a Benford distribution. The worst performing hypothesis is $\mathcal{H}_{r2}$, the hypothesis that the data are fabricated. Hence, the result suggests that the leading digits in the fiscal statistics do not follow a Benford distribution but they also do not seem to be fabricated. Therefore, it might be reasonable to assume that the data have poor overall quality. Further follow-up checks of these numbers could provide information on whether financial statements were actually materially misstated, for instance, by rounding up or down numbers, avoiding certain thresholds and so on (Nigrini, 2019).

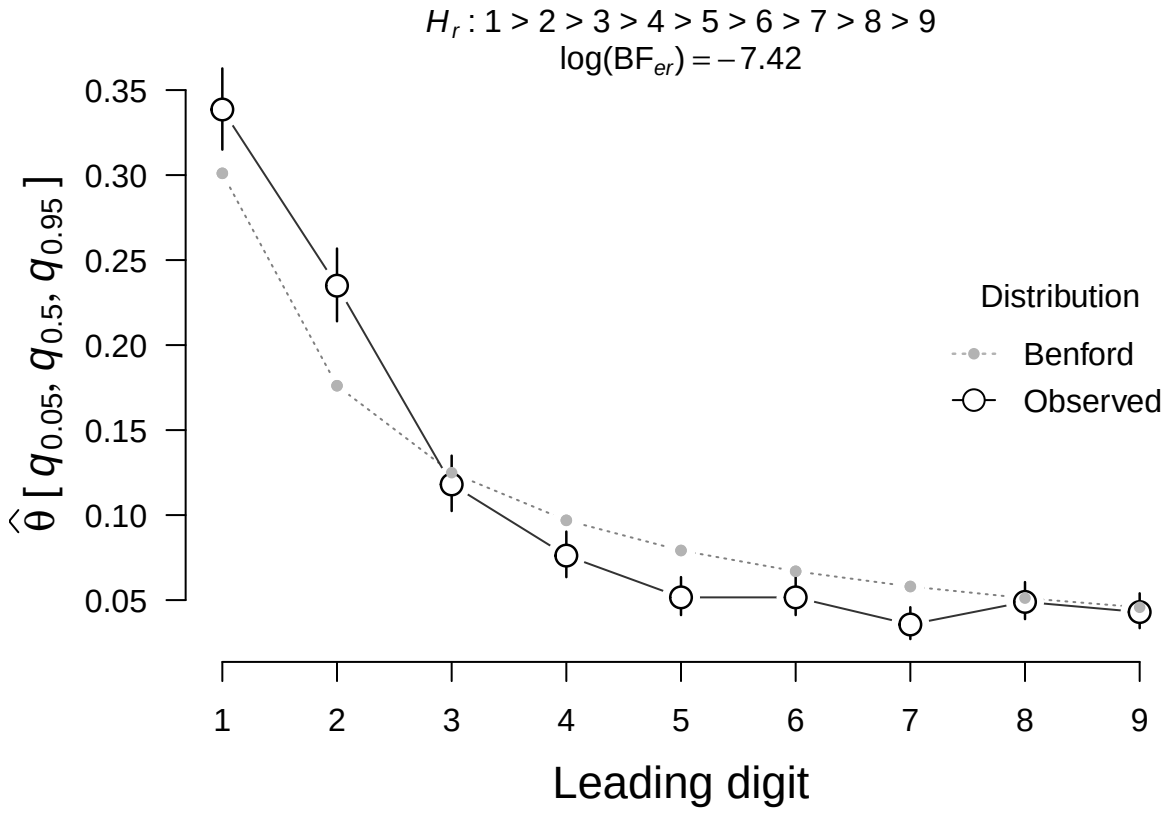

Figure 2. Proportions of leading digits observed in the fiscal statistics from Greece in comparison to the proportions expected according to Benford's law. The black-rimmed dots indicate the the posterior median estimates and corresponding 95% credible intervals based on the encompassing model. The grey filled dots indicate the proportions predicted by Benford's law. Only three out of nine estimates cover the expected proportions. This plot was created using the `plot`-S3-method for `summary.bmult` objects.

3.2 Example 2: Prevalence of Statistical Reporting Errors

In any scientific 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. In most cases this is because researchers copy the relevant test statistics by hand into their articles and there are no automatic checks to detect mistakes. Therefore, Epskamp and Nuijten (2014) developed the R package `statcheck`, which only requires the PDF of a given scientific article to detect these reporting errors automatically and efficiently. 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 psychological 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).

Besides the overall prevalence of statistical reporting errors across these journals, the authors were interested whether there is a higher prevalence for reporting inconsistencies in certain subfields in psychology compared to others. In this context, the possibility was raised that there exists a relationship between the prevalence for reporting inconsistencies and questionable research practices. Specifically, the authors argued that besides honest mistakes when transferring the test statistics into the manuscript, statistical reporting errors occur when authors misreport p-values, for instance, by incorrectly rounding them down to or below 0.05. Based on this assumption, Nuijten et al. (2016) predicted that the proportion of statistical reporting errors should be highest in articles published in the *Journal of Personality and Social Psychology* (JPSP), compared to other journals, because compared to other areas of psychology researchers in social psychology most frequently deemed

questionable research practices defensible and applicable to their research (John, Loewenstein, & Prelec, 2012).

3.2.1 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 hypothesis of interest, $\mathcal{H}_r$, formulated by Nuijten et al. (2016) states that the prevalence for statistical reporting errors for articles published in social psychology journals (i.e., JPSP) is higher than for articles published in other journals. Note that Nuijten et al. (2016) did not make use of inferential statistics since their sample included the entire population of articles from the eight flagship journals in psychology from 1985 to 2013. For demonstration purposes, however, we will test the informed hypothesis stated by the authors. We will test $\mathcal{H}_r$ against the the null hypothesis $\mathcal{H}_0$ that all journals have the same prevalence for statistical reporting errors. In this example, the parameter vector of the binomial success probabilities, $\boldsymbol{\theta}$, reflects the probabilities that articles using null hypothesis significance testing (NHST) will have at least one statistical reporting error across journals. Thus, we can formalize the discussed hypotheses as follows:

$$\mathcal{H}_r : (\theta_{\text{DP}}, \theta_{\text{FP}}, \theta_{\text{JAP}}, \theta_{\text{JCCP}}, \theta_{\text{JEPG}}, \theta_{\text{PLoS}}, \theta_{\text{PS}}) < \theta_{\text{JPSP}}$$
$$\mathcal{H}_0 : \theta_{\text{DP}} = \theta_{\text{FP}} = \cdots = \theta_{\text{JPSP}}.$$

3.2.2 Method.

To compute the Bayes factor BF_{0r} we need to specify (1) a vector with observed successes (i.e., number of articles that contain a statistical reporting error), and (2) a vector containing the total number of observations, (3) the informed hypothesis, (4) a vector with prior parameter α_i for each binomial proportion, (5) a vector with prior

parameter β_i for each binomial proportion, and (6) the category labels (i.e., journal names). Since we have no specific expectations about the distribution of statistical reporting errors in any given journal, we set all parameters α_i and β_i 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
# We assign a uniform beta distribution to each binomial proportion
a <- rep(1, 8)
b <- rep(1, 8)

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

# Category labels
journal_names <- journals$journal

# Execute the analysis
results_H0_Hr <- binom_bf_informed(x = x, n = n, Hr = Hr, a = a, b = b,
                                   factor_levels = journal_names,
```

```
                                    bf_type = 'LogBFr0', seed = 2020)
```

```
LogBFr0 <- summary(results_H0_Hr)$bf
LogBFe0 <- results_H0_Hr$bf_list$bf0_table[['LogBFe0']]
LogBFre <- -results_H0_Hr$bf_list$bfr_table[['LogBFer']]


bayes_factor_table <- data.frame(
   BFType = c('LogBFe0', 'LogBFr0', 'LogBFre'),
   BF = c(LogBFe0, LogBFr0, LogBFre))
bayes_factor_table
```

```
##    BFType          BF
## 1 LogBFe0 156.272164
## 2 LogBFr0 158.280000
## 3 LogBFre   2.005374
```

Again, as the evidence is extreme in all three cases, we report all Bayes factors on the log scale. The Bayes factor $\log(\text{BF}_{r0})$ suggests extreme evidence for the informed hypothesis that the social psychology journal JPSP has the highest prevalence for statistical reporting errors compared to the null hypothesis that the statistical reporting errors are equal across journals; $\log(\text{BF}_{r0}) = 158.28$.

In order to get a clearer picture about the ordering of the journals, we can investigate the posterior estimates under the encompassing model as the next step. The posterior median and 95% credible interval are returned by the `summary`-method and can be plotted, Figure 3.

When comparing $\mathcal{H}_r$ and $\mathcal{H}_0$ with the encompassing hypothesis, we also see that the data suggest that the null hypothesis that the statistical reporting errors are equal across

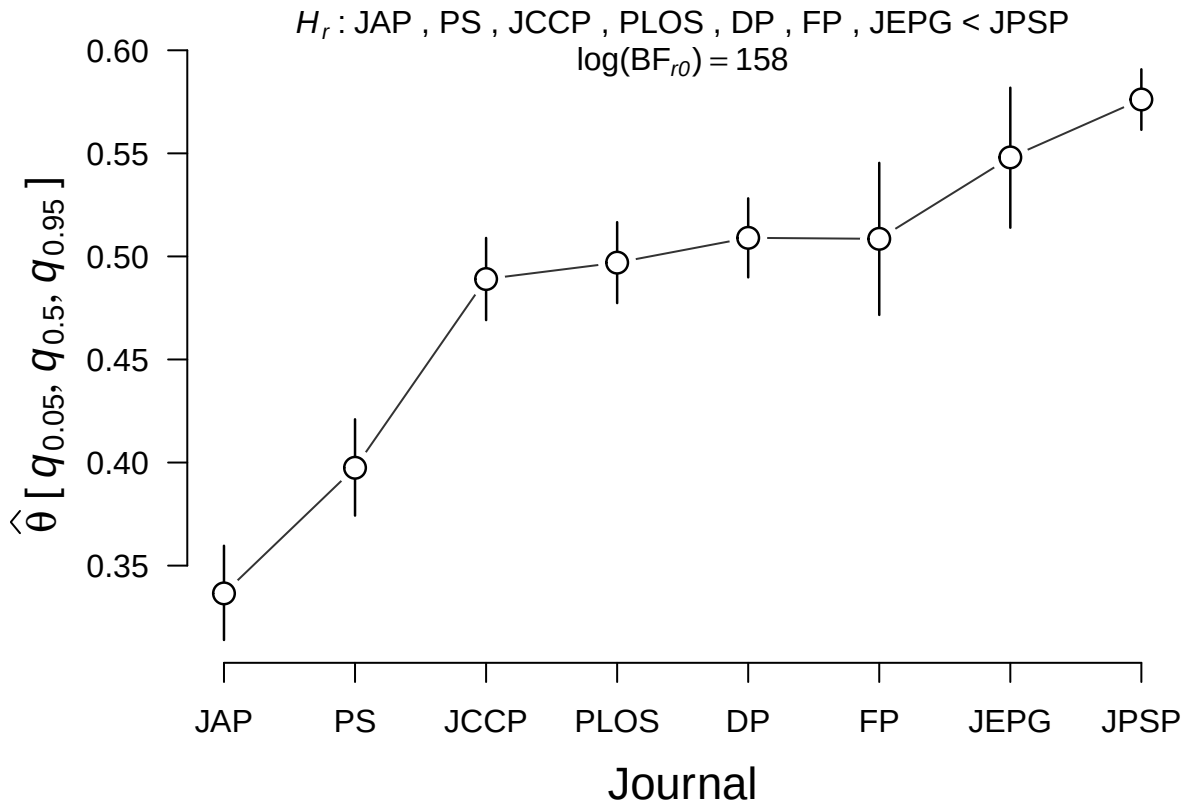

Figure 3. The figure displays for each journal the posterior estimates for the prevalence that an article includes a statistical reporting error and the corresponding 95% credible intervals based on the encompassing model. It appears that all journals show a relatively similar prevalence for statistical reporting errors, with the exception of the *Journal of Applied Psychology* (JAP) and *Psychological Science* (PS), whose prevalence is much lower. This plot was created using the `plot`-S3-method for `summary.bmult` objects.

journals is highly unlikely compared to the encompassing hypothesis, $\log(\text{BF}_{e0}) = 156.27$. In addition, the results the data suggest moderate evidence for the informed hypothesis compared to the hypothesis that the ordering of the journals can vary freely, $\log(\text{BF}_{re}) = 2.01$.

To summarize, we collected extreme evidence for the hypothesis stated by Nuijten et al. (2016) that the prevalence of statistical reporting errors for articles published in a social psychology journal (i.e., JPSP) is higher than for articles published in other journals.

However, this result should be interpreted with caution. It seems that the result is above all an indication that the null hypothesis is highly misspecified and that the prevalence for a statistical reporting error varies greatly from journal to journal. Evidence that JPSP stands out and has a higher prevalence than the other journals is relatively small; the data provided only moderate evidence against the encompassing hypotheses.

4 Summary

The `R` package **multibridge** facilitates the estimation of Bayes factors for informed hypotheses in binomial and multinomial models. Compared to existing packages, this new package efficiently estimates Bayes factors for models with large number of categories which occur frequently in empirical studies. This efficient and reliable estimation is made possible by a recently developed bridge sampling routine (Sarafoglou et al., 2020). The package offers researchers and practitioners the opportunity to specify informed hypotheses that relate closely to their theories. Specifically, informed hypotheses that feature equality constraints, inequality constraints, and free parameters as well as mixtures between them are supported. Moreover, users can also choose whether the informative hypothesis should be tested against an encompassing hypothesis that lets all parameters vary freely or 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, one extension is to facilitate the specification of hierarchical binomial and multinomial models which would allow users to analyze data where responses are nested within participants. Hierarchical multinomial models can be found, for instance, in source memory research where participants need to select a previously studied item from a list of multiple stimuli (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 of interest or the difference between category proportions such that informed hypotheses can also be specified on odds ratios.

5 Acknowledgements

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).

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.

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

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, Q. 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. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*(10), 1–29.

Gu, X., Hoijtink, 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., & Hoijtink, 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. (1988). Random-number guessing and the first digit phenomenon. *Psychological Reports*, *62*, 967–971.

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

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

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., Hoijtink, H., Leeuw, C. de, & others. (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.

Mulder, J., van Lissa, C., Williams, D. R., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., & Fox, J.-P. (2020). *BFpack: Flexible bayes factor testing of scientific expectations*. Retrieved from https://CRAN.R-project.org/package=BFpack

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.

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. (2020). Evaluating multinomial order restrictions with bridge sampling. *PsyArXiv*. Retrieved from https://psyarxiv.com/bux7p/

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.

Appendix A

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 us to move 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 achieve this goal, **multibridge** uses a probit transformation as proposed in Sarafoglou et al. (2020) which 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 $\boldsymbol{y} \in \mathbb{R}^K$. 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 $1 - \sum_{k=1}^{K} \theta_k$. 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 $\boldsymbol{y} \in \mathbb{R}^{K-1}$.

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 $k \in \{1 \cdots, K\}$ the lower bound for any element in $\boldsymbol{\theta}$ is defined as

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

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; Stan Development Team, 2020). By using this so-called stick-breaking method we can define the upper bound for any θ_k as follows:

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{9}$$

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

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}):

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K. \end{cases} \tag{10}$$

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:

$$u_k = \begin{cases} \dfrac{1 - (f_k \times l_k)}{K} & \text{if } k = 1 \\ \left(\dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k \geq \max(\theta_{i<k}), \\ \left(2 \times \left(\dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} \right) - \max(\theta_{i<k}) \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k < \max(\theta_{i<k}). \end{cases} \tag{11}$$

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

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

6.1 Equality Constrained Parameters

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

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

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

When transforming the samples from the collapsed distributions, we need to account for the fact that the inequality constraints imposed under the original parameter values might not hold for the collapsed parameters. Consider, for instance, a multinomial model 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. In general, lower bounds for the parameters need to be adjusted as follows:

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K, \end{cases} \tag{12}$$

where 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:

$$u_k = \begin{cases} \frac{1}{ERS} \times e_k & \text{if } k = 1 \\ \frac{1 - \sum_{i<k} \theta_i}{ERS} \times e_k & \text{if } 1 < k < K, \end{cases} \tag{13}$$

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

6.2 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:

$$u_k = \begin{cases} \dfrac{1 - (f_k \times l_k)}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{14}$$

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 $1/4$. 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: $\dfrac{1 - 0.15 - (1 \times 0.15)}{1 + 1} = 0.35$.

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., $0.15 < 0.29$, $0.275 \not< 0.285$). 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:

$$u_k = u_k - (\max(\theta_{i<k}) - u_k) \tag{15}$$

$$= 2 \times u_k - \max(\theta_{i<k}). \tag{16}$$

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., $0.15 < 0.29, 0.265 < 0.295$).

Appendix B

References

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

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E., & Marsman, M. (2020). Evaluating multinomial order restrictions with bridge sampling. *PsyArXiv*. Retrieved from https://psyarxiv.com/bux7p/

Stan Development Team. (2020). *Stan modeling language user's guide and reference manual, version 2.23.0*. R Foundation for Statistical Computing. Retrieved from http://mc-stan.org/