multibridge: An R Package To Evaluate Multinomial Order Constraints

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

[1] University of Amsterdam

Author Note

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 evaluation of large models with many constraints and models with small parameter spaces. The package was developed in the R programming language and is freely available from the Comprehensive R Archive Network (CRAN). We illustrate the functions based on two empirical examples.

₂₄   multibridge: An R Package To Evaluate Multinomial Order Constraints

1   Introduction

₂₆   We present **multibridge**, an `R` package to evaluate informed hypotheses in

₂₇   multinomial models and models featuring independent binomials using Bayesian inference.

₂₈   The package allows users to specify constraints on the underlying category proportions

₂₉   including inequality constraints, equality constraints, free parameters and mixtures between

₃₀   them. The package is available from the Comprehensive `R` Archive Network (CRAN) at

₃₁   https://CRAN.R-project.org/package=multibridge. Here we introduce the methodology

₃₂   used to evaluate informed hypotheses in both cases and show how to use the

₃₃   implementations provided in **multibridge** through fully reproducible examples.

₃₄   The most common way to analyze categorical variables is to compare a null hypothesis

₃₅   where the underlying category proportions are exactly equal to an encompassing hypothesis

₃₆   where they are free to vary. Another common null hypothesis is one where category

₃₇   proportions are fixed and follow a predicted pattern. These two typical model comparisons

₃₈   are generally known as either chi-square goodness of fit tests, or binomial or multinomial

₃₉   tests. Although commonly used, this analytic strategy has been criticized, since the null

₄₀   hypotheses might reflect an unrealistic expectation about the real world and the

₄₁   encompassing hypothesis is too uninformative (Hoijtink, Klugkist, & Boelen, 2008). In

₄₂   addition, this strategy is often a vague test of the specific predictions that researchers and

₄₃   practitioners are interested in. A simple example for this are theories that predict ordinal

₄₄   relations among the underlying category proportions, such as increasing or decreasing trends.

₄₅   For instance, to check for irregularities in audit data, one could test whether the leading

₄₆   digits in the data are distributed according to an expected Benford distribution or whether

₄₇   they deviate from it, for example, by showing a general decreasing trend. Here, the Benford

₄₈   distribution can be tested with standard methods, however, the general decreasing trend

49 cannot be tested, since we cannot derive fixed underlying proportions for the leading digits.

50 Theories can also generate more complex predictions, including ones that feature

51 combinations of equality constraints, inequality constraints, and unconstrained category

52 proportions. In the following, we will denote such predictions as informed hypotheses, since

53 they "add theoretical expectations to the traditional alternative hypothesis, thus making it

54 more informative" (Hoijtink et al., 2008, p. 2). Such an informed hypothesis was expressed,

55 for instance, by Nuijten, Hartgerink, Assen, Epskamp, and Wicherts (2016) who studied the

56 prevalence of statistical reporting errors in articles published in different areas of

57 psychological science. Nuijten et al. (2016) hypothesized that articles published in social

58 psychology journals would have higher error rates than articles published in other

59 psychological journals while not expressing expectations about the error rate distribution

60 among the other journals. Here again it is not possible to apply standard tests, since we

61 cannot derive fixed proportions based on the hypothesis. Generally, if researchers and

62 practitioners can utilize statistical methods for testing informed hypotheses, they are able to

63 test hypotheses that relate more closely to their theories (Haaf, Klaassen, & Rouder, 2019;

64 Rijkeboer & van den Hout, 2008).

65   In the Bayesian framework, researchers can compare models that instantiate the

66 hypotheses of interest by means of Bayes factors (Jeffreys, 1935; Kass & Raftery, 1995). To

67 compute Bayes factors for informed hypotheses several R packages are already available. For

68 instance, with the package **multinomineq** (Heck & Davis-Stober, 2019) users can specify

69 inequality constrained hypotheses but also more general linear inequality constraints for

70 multinomial models as well as models that feature independent binomials. Outside of R, the

71 Fortran 90 program **BIEMS** (Mulder, Hoijtink, Leeuw, & others, 2012) allows for the

72 evaluation of order constraints for multivariate linear models such as MANOVA, repeated

73 measures, and multivariate regression. These packages rely on one of two implementations of

74 the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005; Sedransk, Monahan, &

75 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 these methods are currently widely used, they are known to become increasingly unreliable and inefficient as the number of constraints increases or the parameter space of the constrained model decreases (Sarafoglou et al., 2020). However, there are also approaches that are more efficient than the ones mentioned above. For instance, the R packages **BAIN** (Gu, Hoijtink, Mulder, & Rosseel, 2019) and **BFpack** (Mulder et al., 2020) implement a further developed more accurate version of the conditional encompassing approach. Unfortunately, however, these method are not available for categorical variables.

In contrast to these available packages, **multibridge** uses a bridge sampling routine that enables users to compute Bayes factors for informed hypotheses more reliably and efficiently (Bennett, 1976; Meng & Wong, 1996; Sarafoglou et al., 2020). The workhorse for this analysis, the bridge sampling algorithm, constitutes a special case of the algorithm implemented in the R package **bridgesampling** (Gronau, Singmann, & Wagenmakers, 2020). The **bridgesampling** package, allows users to estimate the marginal likelihood for a wide variety of models, including models implemented in Stan (Stan Development Team, 2020). However, the algorithm implemented in **bridgesampling** is not suitable for models that include constraints on probability vectors and hence is unsuitable for the analysis of categorical data. This is because the transformations used in the `bridgesampling` package cannot be applied to probability vectors due to their sum-to-one constraint. Therefore, in **multibridge**, we tailored the bridge sampling algorithm such that it accommodates the specification of informed hypotheses on probability vectors. The package then produces an estimate for the Bayes factor in favor of or against the informed hypothesis. The resulting Bayes factor compares the evidence for the informed hypotheses to the encompassing hypothesis that imposes no constraints on the underlying category proportions. Alternatively, the informed hypothesis can be tested against the null hypothesis that all

underlying category proportions are exactly equal. Users can additionally then either 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 1 summarizes the basic required arguments of the two core functions `mult_bf_informed` and `binom_bf_informed` and Table 2 summarizes all S3 methods currently available in **multibridge**.
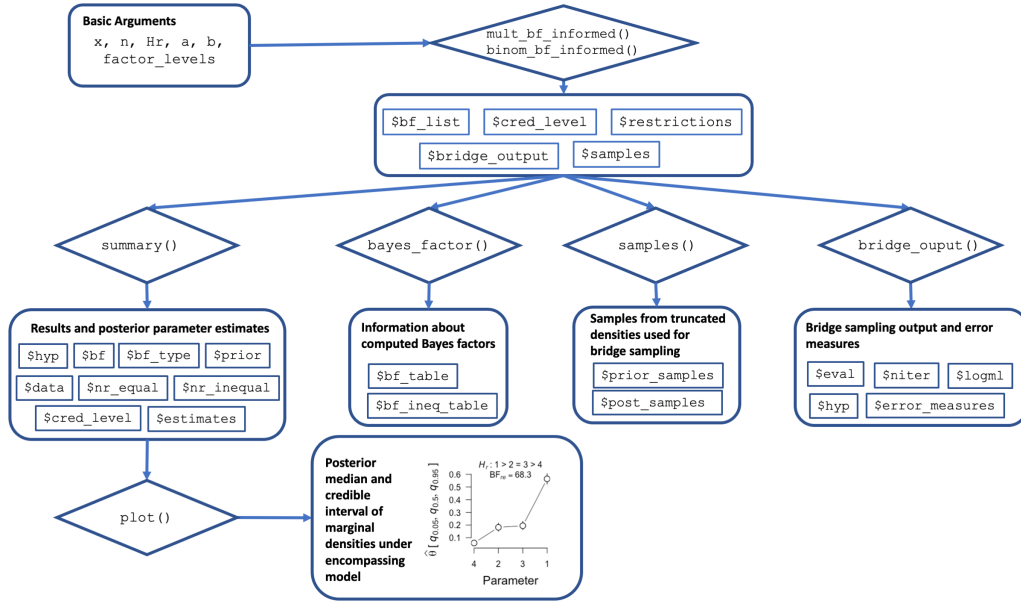
Figure 1. The **multibridge** workflow. The user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and β parameters of the Binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions `mult_bf_informed` and `binom_bf_informed` then produce 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

Argument	Description
x	numeric. a 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 of counts of trials. Must be the same length as x. Ignored if x is a matrix or a table
Hr	string or character. Encodes the user specified informed hypothesis. Users can either use the specified `factor_levels` or numerical indeces 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
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 95% credible interval of the parameter estimates of the encompassing model.
	`bayes_factor`	Contains all Bayes factors and log marginal likelihood estimates for inequality constraints.
	`samples`	Extracts prior and posterior samples from constrained distribution (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.

¹¹³   The methodology used in this package is based on the porposed bright sampling

¹¹⁴ routing by Sarafoglou et al. (2020). Here, we demonstrate how to apply the method in

¹¹⁵ practice using the `multibridge` package. Furthermore, we extend the bridge sampling

¹¹⁶ routine so that also ordinal constraints on binomial parameters can be evaluated. Thus, with

¹¹⁷ `multibridge`, for binomial and multinomial models the following hypotheses can be tested:

¹¹⁸ (a) hypotheses that postulate a monotonic increase and decrease for (a subset of) parameters

¹¹⁹ (e.g., $\theta_1 < \theta_2 < \theta_3$ or $\theta_1 > \theta_2 > \theta_3$); (b) hypotheses that postulate mixtures of inequality

¹²⁰ constraints and equality constraints (e.g., $\theta_1 < \theta_2 = \theta_3$); (c) hypotheses that postulate

¹²¹ mixtures of inequality constraints and free parameters (e.g., $\theta_1 < \theta_2, \theta_3$); (d) hypotheses that

¹²² postulate mixtures of (a)–(c) (e.g., $\theta_1 < (\theta_2 = \theta_3), \theta_4$). These hypotheses can be tested

¹²³ against the encompassing hypothesis, that all parameters vary freely or against the null

¹²⁴ hypothesis that all category proportions are equal.

¹²⁵   In the remainder of this article, we will first describe the Bayes factor identity for

¹²⁶ informed hypotheses in binomial and multinomial models, and present the bridge sampling

¹²⁷ routine implemented in the **multibridge** package including details of the necessary

¹²⁸ transformations required for this routine. Then, we illustrate how to use the **multibridge**

¹²⁹ package to estimate parameters, and compute Bayes factors using two examples and end

¹³⁰ with a brief summary.

2   Methods

¹³¹

¹³²   In this section we formalize multinomial models and models that feature independent

¹³³ binomial probabilities as we have implemented them in **multibridge**. In the multinomial

¹³⁴ model, we assume that the vector of observations $\mathbf{x}$ in the K categories follows multinomial

¹³⁵ distribution in which the parameters of interest, $\boldsymbol{\theta}$, represent the underlying category

¹³⁶ proportions. Since we assume a dependence between the K categories, the vector of

¹³⁷ probability parameters is constrained to sum to one, such that $\sum_{k=1}^{K}(\theta_1, \cdots, \theta_K) = 1$.

138 Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is the Dirichlet distribution with

139 concentration parameter vector $\boldsymbol{\alpha}$:

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

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

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

141 the model for independent binomial probabilities is very similar since the multinomial model

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

143 we assume that the elements in the vector of successes $\mathbf{x}$ and the elements in the vector of

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

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

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

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

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

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

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

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

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

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

₁₅₂ **2.1   Bayes factor**

₁₅₃   With **multibridge** package, it is possible to collect evidence for informed hypotheses

₁₅₄ on a parameter vector $\boldsymbol{\theta}$ by means of the Bayes factor. Bayes factors compare the relative

₁₅₅ evidence of two hypotheses in the light of the data. It is defined as the ratio of marginal

₁₅₆ likelihoods of the respective hypotheses. For instance, the Bayes factor for the informed

₁₅₇ hypothesis versus a hypothesis that lets all parameters free to vary is defined as:

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

₁₅₈   where the subscript r denotes the informed (restricted) hypothesis and e denotes the

₁₅₉ (encompassing) hypothesis which predicts that all parameters free to vary. In **multibridge**

₁₆₀ we use two different methods to compute Bayes factors, one method evaluates hypotheses

₁₆₁ that feature equality constraints on $\boldsymbol{\theta}$ and one method evaluates hypotheses that feature

₁₆₂ inequality constraints on $\boldsymbol{\theta}$. Both methods will be outlined below. In cases where informed

₁₆₃ hypotheses feature mixtures between inequality and equality constraints, we compute the

₁₆₄ corresponding Bayes factor BF_{re} by multiplying the individual Bayes factors for both

₁₆₅ constraint types with each other. A Bayes factor for mixtures thus factors into a Bayes

₁₆₆ factor for the equality constraints, and a conditional Bayes factor for the inequality

₁₆₇ constraints given the equality constraints (for the proof, see Sarafoglou et al., 2020).

₁₆₈ **2.2   The Bayes Factor For Equality Constraints**

   The Bayes factor for the equality constraints can be computed analytically both for

binomial and multinomial models. For binomial models, the function `binom_bf_equality` is

available to compute BF_{0e}. Assuming that the first i binomial probabilities in a model are equality constrained, the Bayes factor is defined as:

$$\mathrm{BF}_{0e} = \frac{\prod_{i<k} \mathrm{B}(\alpha_i, \beta_i)}{\prod_{i<k} \mathrm{B}(\alpha_i + x_i, \beta_i + n_i - x_i)} \times \frac{\mathrm{B}(\alpha_+ + x_+ - i + 1, \beta_+ + n_+ - x_+ - i + 1)}{\mathrm{B}(\alpha_+ - i + 1, \beta_+ - i + 1)}$$

169 where $\mathrm{B}()$ denotes the beta function and $\alpha_+ = \sum_{i<k} \alpha_i$, $\beta_+ = \sum_{i<k} \beta_i$, $x_+ = \sum_{i<k} x_i$ and

170 $n_+ = \sum_{i<k} n_i$. The latter factor introduces a correction for marginalizing which stems from

171 the change in degrees of freedom, when we collapse i equality constraint parameters: For i

172 collapsed categories, $i - 1$ degrees of freedom are lost which are subtracted from the prior

173 parameters in the corresponding Binomial distribution.

For multinomial models, the function `multBayes_bf_equality` is available. Assuming again that the first i category probabilities in a model are equality constraint, the Bayes factor BF_{0e} is defined as:

$$\mathrm{BF}_{0e} = \frac{\mathrm{B}(\boldsymbol{\alpha} + \mathbf{x})}{\mathrm{B}(\boldsymbol{\alpha})} \left(\frac{1}{i}\right)^{\sum_{i<k} x_i} \frac{\mathrm{B}\left(\sum_{i<k} \alpha_i - i + 1, \alpha_k, \ldots, \alpha_K\right)}{\mathrm{B}\left(\sum_{i<k} \alpha_i + x_i - i + 1, \alpha_k + x_k, \ldots, \alpha_K + x_K\right)}.$$

174 ## 2.3 The Bayes Factor For Inequality Constraints

175 To approximate the Bayes factor for informed hypotheses, Klugkist et al. (2005)

176 derived the following identity:

$$\mathrm{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}^{\substack{\text{Proportion of posterior parameter} \\ \text{space consistent with the restriction}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Proportion of prior parameter} \\ \text{space consistent with the restriction}}}}. \tag{5}$$

177 Recently, Sarafoglou et al. (2020) highlightes that the Bayes factor BF_{re} can also be

178 interpreted as ratio of two marginal likelihoods. In this identity, $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$ denotes

179 the marginal likelihood of the constrained posterior distribution and $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ denotes

180 the marginal likelihood of the constrained prior distribution. Even though both identities are

181 mathematically equivalent, the methods to estimate these identities differ substantially. In

182 the first case, the number of samples from the encompassing distribution in accordance with

183 the inequality constrained hypothesis serve as an estimate for the proportion of prior

184 parameter space consistent with the restriction. Although easy to implement, this definition

185 implies that the accuracy of this estimate is strongly dependent on the number of the

186 constrained parameters in the model and the size of the constrained parameter space. That

187 is, as the constraints become stronger, the constrained parameter space decreases. As a

188 result it becomes less likely that draws from the encompassing distribution will fall into the

189 constrained region, so that in some cases the estimation of the Bayes factor becomes

190 practically impossible (Sarafoglou et al., 2020).

191   However, when we interpret the Bayes factor BF_{re} as ratio of marginal likelihoods and

192 we are able to sample from the constrained prior and posterior distributions, we can utilize

193 numerical sampling methods such as bridge sampling to obtain the estimates. Crucially, in

194 this approach, it does not matter how small the constrained parameter space is in proportion

195 to the encompassing density. This gives the method a decisive advantage over the

196 encompassing prior approach in terms of accuracy and efficiency especially (1) when

197 binomial and multinomial models with relatively high number of categories (i.e., $K > 10$) are

198 evaluated and (2) when relatively little posterior mass falls in the constrained parameter

199 space.

2.4   The Bridge Sampling Method

201   Bridge sampling is a method to estimate the ratio of two marginal likelihoods (Bennett,

202 1976; Meng & Wong, 1996). In **multibridge**, we are using bridge sampling to estimate the

203 identity presented in Equation 5. But instead of estimating the ratio of marginal likelihoods

204 directly, we implemented a version of bridge sampling that estimates one marginal likelihood

₂₀₅ at the time. This approach has the benefit that it increases the accuracy of the method

₂₀₆ without considerably decreasing its computational efficiency (Overstall & Forster, 2010).

₂₀₇ Specifically, we subsequently estimate the marginal likelihood for the constrained prior

₂₀₈ distribution and the marginal likelihood of the constrained posterior distribution.

₂₀₉ When applying this modified version of the bridge sampling method, we estimate each

₂₁₀ marginal likelihood by means of a so-called proposal distribution. In **multibridge** this

₂₁₁ proposal distribution is the multivariate normal distribution. To estimate the marginal

₂₁₂ likelihood, bridge sampling only requires samples from the distribution of interest—the

₂₁₃ so-called target distribution—and samples from the proposal distribution.

₂₁₄ Samples from the target distribution—that is the constrained prior and posterior

₂₁₅ Dirichlet distribution for multinomial models and constrained prior and posterior beta

₂₁₆ distributions for binomial models—are drawn through the Gibbs sampling algorithms

₂₁₇ proposed by Damien and Walker (2001). For binomial models, we apply the suggested Gibbs

₂₁₈ sampling algorithm for constrained beta distributions. In the case of the multinomial models,

₂₁₉ we apply an algorithm that simulates values from constrained Gamma distributions which

₂₂₀ are then transformed into Dirichlet random variables (for details, see Appendix C in

₂₂₁ Sarafoglou et al., 2020). To sample efficiently from these distributions, **multibridge**

₂₂₂ provides a `C++` implementation of this algorithm.

₂₂₃ Samples from the proposal distribution can be generated using the standard

₂₂₄ `rmvnorm`-function from the `R` package **mvtnorm** (Genz et al., 2020). The vector of means

₂₂₅ and the covariance matrix of this distribution are derived from one part of the samples of the

₂₂₆ probit transformed target distribution. The reason for this approach is that the efficiency of

₂₂₇ the bridge sampling method is optimal only if the target and proposal distribution operate

₂₂₈ on the same parameter space and have sufficient overlap. We therefore probit transform the

₂₂₉ samples of the constrained distributions to move the samples from the probability space to

₂₃₀ the entire real line. Subsequently, we use half of these draws to construct the proposal

231 distribution using the method of moments. Details on the probit transformations are

232 provided in the appendix. Thus, for the marginal likelihood of the constrained prior

233 distribution, the modified bridge sampling identity is then defined as

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

234 where the term $h(\boldsymbol{\theta})$ refers to the bridge function proposed by Meng and Wong (1996) and

235 $g(\boldsymbol{\theta})$ refers to the proposal distribution. The numerator evaluates the unnormalized density

236 for the constrained prior distribution with samples from the proposal distribution. The

237 denominator evaluates the normalized proposal distribution with samples from the

238 constrained prior distribution. Using this identity, we receive the bridge sampling estimator

239 for the marginal likelihood of the constrained prior distribution by applying the iterative

240 scheme proposed by Meng and Wong (1996):

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

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

242 $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$, N_2 denotes the number of samples drawn from the proposal distribution, that

243 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:

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

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

where $\boldsymbol{\xi_n}^* = \Phi^{-1}\left(\dfrac{\boldsymbol{\theta}_n^* - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_m = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_m) + \mathbf{l})\,|J|)$. 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 bridge sampling estimate for the log marginal likelihood of the constrained distribution and its associate relative mean square error, the number of iterations, and the quantities $q_{1,2}$, $q_{1,2}$, $q_{1,2}$, and $q_{1,2}$ are included in the standard output in **multibridge**. The function to compute the relative mean square error was adopted from the R package **bridgesampling**.

3 Usage and Examples

The **multibridge** package can be installed from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=multibridge:

```
install.packages('multibridge')
library('multibridge')
```

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**—the

`mult_bf_informed`-function and the `binom_bf_informed`-function—can be illustrated

schematically as follows:

```
mult_bf_informed(x, Hr, a factor_levels)

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

The basic required arguments for these functions are listed in Table 1. 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.

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 truncated 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 truncated 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

265 is approximately equal to

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

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

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

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

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

270 expected proportions). Examples of empirical data for which this relationship holds include

271 data on population sizes, death rates, baseball statistics, atomic weights of elements, and

272 physical constants (Benford, 1938). In contrast, generated data, such as telephone numbers,

273 do in general not obey Benford's law (Hill, 1995). Given that Benford's law applies to

274 empirical data but not artificially generated data, a so-called Benford test can be used to

275 check whether a set of data obey Benford's law and therefore exhibit an important property

276 of empirical data sets. Benford's tests are used in fields like accounting and auditing to check

277 for indications for poor data quality, for instance, in fiscal statements (for an overview, see

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

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

280 recommended that the data undergo additional follow-up checks (Nigrini, 2019).

281   Below, we discuss three possible Bayesian adaptations of the Benford's test. In a first

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

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

284 which allows all proportions of first digits free to vary. Testing against the encompassing

285 hypothesis is considered standard practice, yet, it leads to an unfair comparison to the

286 detriment of the null hypothesis. In general, if we are dealing with a high-dimensional

287 parameter space and the competing hypotheses differ largely in their complexity, the Bayes

288 factor generally favors the less complex hypothesis (i.e., $\mathcal{H}_e$) even if the data follow the

289 predicted trend of the more complex hypothesis considerably well. In a second scenario we

290 therefore test the null hypothesis against an alternative hypothesis, denoted as $\mathcal{H}_{r1}$, which

291 predicts a monotonically decreasing trend in the proportions of leading digits. The

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

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

294 data follow a similar but different trend. In a third scenario, where the main goal is to

295 identify fabricated data, we could test the null hypothesis against a hypothesis which

296 predicts a trend that is characteristic for manipulated data. This hypothesis, which we

297 denote as $\mathcal{H}_{r2}$, could be derived from empirical research on fraud or be based on observed

298 patterns from former fraud cases. For instance, Hill (1988) instructed students to produce a

299 series of random numbers; in the resulting data the proportion of the leading digit 1

300 occurred most often and the digits 8 and 9 occurred least often which is consistent with the

301 general pattern of Benford's law. However, the proportion for the remaining leading digits

302 were approximately equal. We do want to note that the predicted distribution derived from

303 Hill (1988) is not currently used as a test to detect fraud. However, for the sake of simplicity,

304 if we assume that this pattern could be an indication of fabricated auditing data, the Bayes

305 factor would quantify the evidence of whether the proportion of first digits resemble

306 authentic or fabricated data.

307 **3.1.1 Data and Hypothesis.** The data we use to illustrate the computation of

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

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

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

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

312 data used for this example features the proportion of first digits from fiscal data from Greece

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

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

315 states has been repeatedly criticized by the European Commission in this time span

316 (European Commision, 2004, 2010). In particular, the commission has accused the Greek

317 statistical authorities to have misreported deficit and debt statistics. For further details on

318 the data set see Rauch et al. (2011). The observed proportions are displayed in Table 4, the

319 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

320 In this example, the parameter vector of the multinomial model, $\theta_1, \cdots, \theta_K$, reflects

321 the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9. Thus,

322 we can formalize the discussed hypotheses as follows. The null hypothesis specifies that the

323 proportions of first digits obeys Benford's law:

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

We are testing the null hypothesis against the following alternative hypotheses:

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

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

$$\mathcal{H}_{r2} : \theta_1 > (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7) > (\theta_8, \ \theta_9).$$

324   If we are interested in comparing two informed hypothesis with each other we can

325 make use of the transitivity property of the Bayes factor. For instance, if we would like to

326 compare the two informed hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ with each other, we would first compute

327 BF_{er1} and BF_{er2} and then yield BF_{r1r2} as follows:

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

328   **3.1.2 Method.** We can compare $\mathcal{H}_0$ and $\mathcal{H}_e$ by means of a Bayesian multinomial

329 test, that is, we stipulate equality constraints on the entire parameter vector $\boldsymbol{\theta}$. The

330 corresponding Bayes factor is thus computationally straightforward; we can calculate BF_{0e}

331 by applying the function `mult_bf_equality`. To evaluate $\mathcal{H}_0$, we only need to specify (1) a

332 vector with observed counts, (2) a vector with concentration parameters of the Dirichlet

333 prior distribution, and (3) the vector of proportions expected under the null. We do not

334 want to incorporate any specific expectations about the distribution of leading digits in the

335 Greek fiscal data. Hence, we set all concentration parameters to one which corresponds to a

336 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. 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
```

343 ##   BFType   LogBF

344 ## 1 LogBFe0   17.6715

345 ## 2 LogBFr10   25.0883

346 ## 3 LogBFr20 -154.5685

347   As the evidence is extreme in all three cases, we report all Bayes factors on the log
348 scale. The log Bayes factor $\log(\text{BF}_{e0})$ suggests extreme evidence against the hypothesis that
349 the first digits in the Greek fiscal data follow a Benford's distribution; $\log(\text{BF}_{e0}) = 17.67$.
350 The log Bayes factor $\log(\text{BF}_{r10})$ indicates extreme evidence in favor for a decreasing trend,
351 $\log(\text{BF}_{r10}) = 25.09$. Even though the Bayes factor suggests extreme evidence against the
352 hypothesis that the Greek fiscal data are an empirical data set, there is no support for the
353 hypothesis that the data are fabricated. The log Bayes factor $\log(\text{BF}_{r20})$ indicates extreme
354 evidence against $\mathcal{H}_{r2}$ with $\log(\text{BF}_{r20}) = -154.57$. When we compare the informed hypotheses
355 directly with each other, the data show evidence for a decreasing trend $(\log(\text{BF}_{r1r2}) = 180)$.

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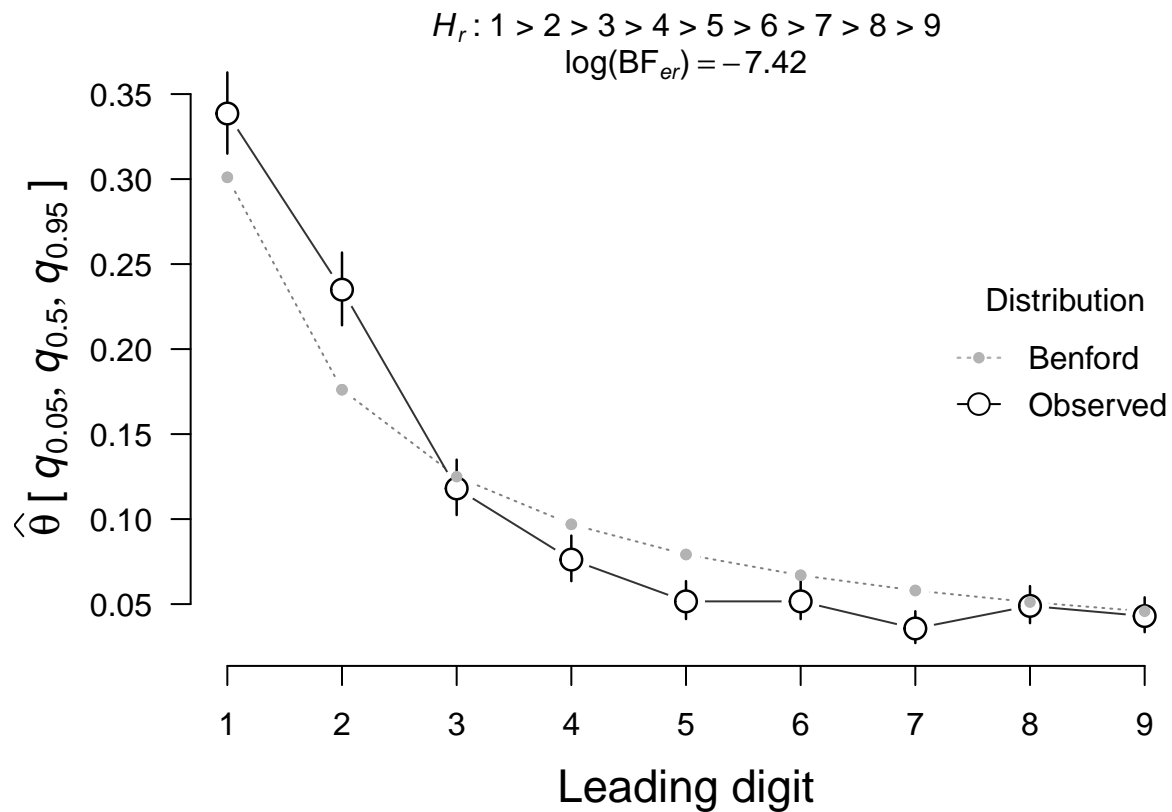364 thresholds and so on (Nigrini, 2019).

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

391 questionable research practices defensible and applicable to their research (John,

392 Loewenstein, & Prelec, 2012).

393 **3.2.1 Data and Hypothesis.** Here, we reuse the original data published by

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

395 name `journals`.

```
data(journals)
```

396 The hypothesis of interest, $\mathcal{H}_r$, formulated by Nuijten et al. (2016) states that the

397 prevalence for statistical reporting errors for articles published in social psychology journals

398 (i.e., JPSP) is higher than for articles published in other journals. Note that Nuijten et al.

399 (2016) did not make use of inferential statistics since their sample included the entire

400 population of articles from the eight flagship journals in psychology from 1985 to 2013. For

401 demonstration purposes, however, we will test the informed hypothesis stated by the authors.

402 We will test $\mathcal{H}_r$ against the the null hypothesis $\mathcal{H}_0$ that all journals have the same prevalence

403 for statistical reporting errors. In this example, the parameter vector of the binomial success

404 probabilities, $\boldsymbol{\theta}$, reflects the probabilities that articles using null hypothesis significance

405 testing (NHST) will have at least one statistical reporting error across journals. Thus, we

406 can formalize the discussed hypotheses as follows:

$$\mathcal{H}_r : (\theta_{\mathrm{DP}}, \theta_{\mathrm{FP}}, \theta_{\mathrm{JAP}}, \theta_{\mathrm{JCCP}}, \theta_{\mathrm{JEPG}}, \theta_{\mathrm{PLoS}}, \theta_{\mathrm{PS}}) < \theta_{\mathrm{JPSP}}$$

$$\mathcal{H}_0 : \theta_{\mathrm{DP}} = \theta_{\mathrm{FP}} = \cdots = \theta_{\mathrm{JPSP}}.$$

407 **3.2.2 Method.** To compute the Bayes factor BF_{0r} we need to specify (1) a vector

408 with observed successes (i.e., number of articles that contain a statistical reporting error),

409 and (2) a vector containing the total number of observations, (3) the informed hypothesis,

410 (4) a vector with prior parameter α_i for each binomial proportion, (5) a vector with prior

411 parameter β_i for each binomial proportion, and (6) the category labels (i.e., journal names).

412 Since we have no specific expectations about the distribution of statistical reporting errors in

413 any given journal, we set all parameters α_i and β_i to one which corresponds to uniform beta

414 distributions. With this information, we can now conduct the analysis with the function

415 `binom_bf_informed`.

```r
# 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 = 'BFr0', seed = 2020)

BFr0 <- summary(results_H0_Hr)$bf


BFe0 <- results_H0_Hr$bf_list$bf0_table[['BFe0']]
BFre <- results_H0_Hr$bf_list$bfr_table[['BFre']]


bayes_factor_table <- data.frame(
    BFType = c('BFe0', 'BFr0', 'BFre'),
    BF = c(BFe0, BFr0, BFre))
bayes_factor_table
```

416 `## BFType           BF`

417 `## 1  BFe0 7.381395e+67`

418 `## 2  BFr0 5.483500e+68`

419 `## 3  BFre 7.428873e+00`

420   The Bayes factor BF_{r0} suggests extreme evidence for the informed hypothesis that the
421 social psychology journal JPSP has the highest prevalence for statistical reporting errors
422 compared to the null hypothesis that the statistical reporting errors are equal across journals;
423 $\log(\text{BF}_{r0}) = 158.28$. When taking a closer look at the Bayes factors, we also see that the
424 data suggest that the null hypothesis that the statistical reporting errors are equal across
425 journals is highly unlikely compared to the encompassing hypothesis, $\log(\text{BF}_{e0}) = 156.27$. In
426 addition, the results suggest that the data are 7.43 more likely under the informed
427 hypothesis than under the hypothesis that the ordering of the journals can vary freely.

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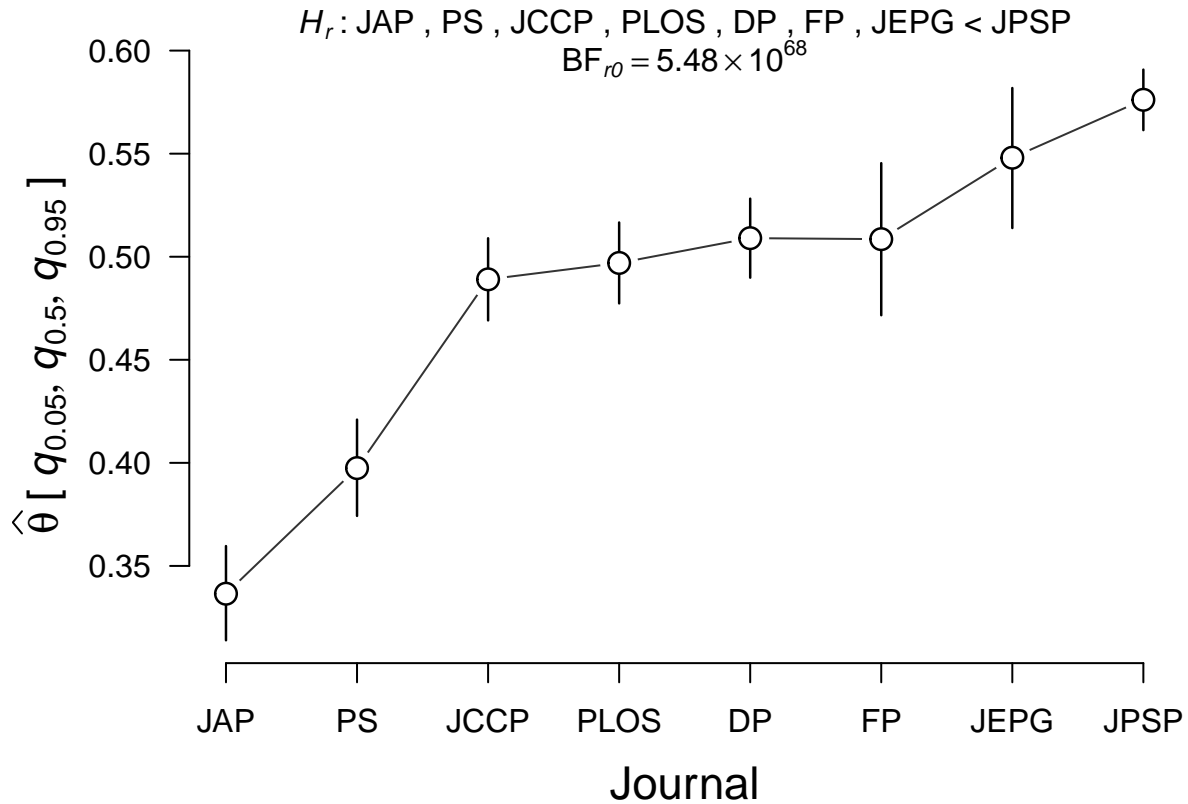
431   Figure 3.

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.

432   To summarize, we collected extreme evidence for the hypothesis stated by Nuijten et al.

433   (2016) that the prevalence of statistical reporting errors for articles published in a social

434   psychology journal (i.e., JPSP) is higher than for articles published in other journals.

435   However, this result should be interpreted with caution. It seems that the result is above all

436   an indication that the null hypothesis is highly misspecified and that the prevalence for a

437   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

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*

534   *Statistical Software, 46*, 1–39.

535   Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H.

536   (2009). Bayesian model selection of informative hypotheses for repeated

537   measurements. *Journal of Mathematical Psychology, 53*, 530–546.

538   Mulder, J., van Lissa, C., Williams, D. R., Gu, X., Olsson-Collentine, A., Boeing-Messing, F.,

539   & Fox, J.-P. (2020). *BFpack: Flexible bayes factor testing of scientific expectations.*

540   Retrieved from https://CRAN.R-project.org/package=BFpack

541   Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers.

542   *American Journal of Mathematics, 4*, 39–40.

543   Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud*

544   *detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.

545   Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud schemes.

546   *Managerial Auditing Journal, 34*, 602–622.

547   Nigrini, M. J., & Mittermaier, L. J. (1997). The use of benford's law as an aid in analytical

548   procedures. *Auditing, 16*, 52.

549   Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016).

550   The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior*

551   *Research Methods, 48*, 1205–1226.

552   Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for

553   generalised linear mixed models. *Computational Statistics & Data Analysis, 54*,

554   3269–3288.

555   Rauch, B., Göttsche, M., Brähler, G., & Engel, S. (2011). Fact and fiction in

556   EU-governmental economic data. *German Economic Review, 12*, 243–255.

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.

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/

Appendix

Transforming An Ordered Probability Vector To The Real Line

569   Since we choose the multivariate normal as proposal distribution, the mapping between the

570   proposal and target distribution requires us to move $\boldsymbol{\theta}$ to the real line. Crucially, the

571   transformation needs to retain the ordering of the parameters, that is, it needs to take into

572   account the lower bound l_k and the upper bound u_k of each θ_k. To achieve this goal,

573   **multibridge** uses a probit transformation as proposed in Sarafoglou et al. (2020) which

574   subsequently transforms the elements in $\boldsymbol{\theta}$ moving from its lowest to its highest value. In the

575   binomial model, we move all elements in $\boldsymbol{\theta}$ to the real line and thus construct a new vector

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

577   is completely determined by its first $K - 1$ elements, where θ_K is defined as $1 - \sum_{k=1}^{K} \theta_k$.

578   Hence, for multinomial models we will only consider the first $K - 1$ elements of $\boldsymbol{\theta}$ and we

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

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

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

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

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

583   The inverse transformation is given by

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

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

585   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

586   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 each parameter, we need to take into account

₆₀₃ for each element θ_k the number of equality constrained parameters that are collapsed within

₆₀₄ this element (denoted as e_k):

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

₆₀₅ The upper bound for parameters in the binomial models still remains 1. To determine the

₆₀₆ upper bound for multinomial models we must, additionally for each element θ_k, take into

₆₀₇ account the number of free parameters that share common upper and lower bounds (denoted

₆₀₈ with f_k). The upper bound is then defined as:

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

₆₀₉ The elements in the remaining stick are then computed as follows

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

₆₁₀ The rationale behind these modifications will be described in more detail in the following

₆₁₁ sections. In **multibridge**, information that is relevant for the transformation of the

₆₁₂ parameter vectors is stored in the generated `restriction_list` which is returned by the

₆₁₃ main functions `binom_bf_informed` and `mult_bf_informed` but can also be generated

₆₁₄ separately with the function `generate_restriction_list`. This restriction list features the

₆₁₅ 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   ***6.0.0.1   Equality Constrained Parameters.***   In cases where informed

619 hypotheses feature a mix of equality and inequality constrained parameters, we compute the

620 corresponding Bayes factor BF_{re}, by multiplying the individual Bayes factors for both

621 constrait 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 the

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

624 $\text{BF}_{2e} \mid \text{BF}_{1e}$, we collapse in the constrained prior and posterior distributions all equality

625 constrained parameters into one category which has implications on the performed

626 transformations.

627   When transforming the samples from these distributions, we need to account for the

628 fact that the inequality constraints imposed under the original parameter values might not

629 hold for the collapsed parameters. Consider, for instance, a multinomial model in which we

630 specify the following informed hypothesis

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

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 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 $0.05 \times 3 = 0.15$. In addition, to determine the lower bound of θ_5 we divide the preceding value θ_* by three, that is, $0.6/3 = 0.2$. 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. Similarly, to determine the upper bound for a given parameter value, we need to multiple the upper bound the number of equality constrained parameters within the current constraint:

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

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 $\dfrac{1 - 0.05}{5} \times 3 = 0.57$. The upper bound for θ_5 is then $\dfrac{(1 - 0.05 - 0.45)}{2} \times 1 = 0.25$.

6.0.0.2   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

$$\mathcal{H}_r : \theta_1 \leq \theta_2, \theta_3 \leq \theta_4.$$

This hypothesis specifies that θ_2 and θ_3 have the shared lower bound θ_1 and the shared upper bound 1, 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 bounds, we need to subtract for each θ_k from the length of the remaining stick the lower bounds of all parameters that share common bounds with θ_k and that have not yet been accounted for in the transformation:

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

where f_k represents the number of free parameters that share common upper and lower bounds with θ_k and that have been not yet been accounted for. 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 $(0.15, 0.3, 0.2, 0.35)$. 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 upper and lower bounds. Thus, to compute the upper bound for θ_2, we subtract from the length of the remaining stick the lower bound of θ_3: $\dfrac{1 - 0.15 - (0.15 \times 1)}{2} = 0.35$.

A further correction is required, if a preceding free parameter (i.e., a free parameter that was already accounted for in the stick) is larger than the upper bound of the current parameter. For instance, in our example the upper bound for θ_3 would be $\dfrac{1 - 0.15 - 0.3}{2} = 0.275$, but the preceding free parameter is 0.3. However, if θ_3 would actually take on the value 0.275, then θ_4 would have to be 0.275 as well, which would violate the constraint (i.e., $0.15 \leq 0.3, 0.275 \nleq 0.275$). In these cases, the upper bound needs to be corrected downwards. To do this, we subtract the difference between the largest preceding

free parameter in the sequence with the current upper bound. 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 $2 \times 0.275 - 0.3 = 0.25$ which secures the proper ordering for the remainder of the parameters: if θ_3 would take on the value 0.25, θ_4 would be 0.3 which would be in accordance with the constraint, that is, $0.15 \le 0.3, 0.25 \le 0.3$.

6.1   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/