

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

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

University of Amsterdam

Correspondence concerning this article should be addressed to: Alexandra Sarafoglou,

<sup>7</sup> Department of Psychology, PO Box 15906, 1001 NK Amsterdam, The Netherlands, E-mail:  
<sup>8</sup> alexandra.sarafooglou@gmail.com

9

## Abstract

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

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

26 **Introduction**

27 The most common way to analyze categorical variables is to conduct either binomial  
28 tests, multinomial tests, or chi-square goodness of fit tests. These tests compare the  
29 encompassing hypothesis to a null hypothesis that all underlying category proportions are  
30 either exactly equal, or follow a specific distribution. Accordingly, these tests are suitable  
31 when theories predict either the invariance of all category proportions or specific values. For  
32 instance, chi-square goodness of fit tests are commonly used to test Benford's law, which  
33 predicts the distribution of leading digits in empirical datasets (Benford, 1938; Newcomb,  
34 1881). Often, however, the predictions that researchers are interested in are of a different  
35 kind. Consider for instance the weak-order mixture model of decision-making (Regenwetter  
36 & Davis-Stober, 2012). The theory predicts that individuals' choice preferences are weakly  
37 ordered at all times, that is, if they prefer choice  $A$  over  $B$  and  $B$  over  $C$  then they will  
38 also prefer  $A$  over  $C$  (Regenwetter, Dana, & Davis-Stober, 2011)—a well-constrained  
39 prediction of behavior. The theory is, however, silent about the exact values of each choice  
40 preference. Hence, the standard tests that compare  $\mathcal{H}_e$  to  $\mathcal{H}_0$  are unsuited to test the  
41 derived predictions. Instead, the predictions need to be translated into an informed  
42 hypothesis  $\mathcal{H}_r$  that reflects the predicted ordinal relations among the parameters. Only  
43 then is it possible to adequately test whether the theory of weakly-ordered preference  
44 describes participants' choice behavior. Of course, researchers may be interested in more  
45 complex hypotheses, including ones that feature combinations of equality constraints,  
46 inequality constraints, and unconstrained category proportions. For instance, Nuijten,  
47 Hartgerink, Assen, Epskamp, and Wicherts (2016) hypothesized that articles published in  
48 social psychology journals would have higher error rates than articles published in other  
49 psychology journals. As in the previous example, the authors had no expectations about

50 the exact error rate distribution across journals. Here, again, the standard tests are  
 51 inadequate. Generally, by specifying informed hypotheses researchers and practitioners are  
 52 able to “add theoretical expectations to the traditional alternative hypothesis” (Hoijtink,  
 53 Klugkist, & Boelen, 2008, p. 2) and thus test hypotheses that relate more closely to their  
 54 theories (Haaf, Klaassen, & Rouder, 2019; Rijkeboer & van den Hout, 2008).

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

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

55 where the subscript  $r$  denotes the informed hypothesis and  $e$  denotes the encompassing  
 56 hypothesis. Several available R packages compute Bayes factors for informed hypotheses.  
 57 For instance, the package **multinomineq** (Heck & Davis-Stober, 2019) evaluates informed  
 58 hypotheses for multinomial models as well as models that feature independent binomials.  
 59 The package **BFpack** (Joris Mulder et al., 2021) evaluates informed hypotheses for  
 60 statistical models such as univariate and multivariate normal linear models, generalized  
 61 linear models, special cases of linear mixed models, survival models, and relational event  
 62 models. The package **bain** (Gu, Hoijtink, Mulder, & Rosseel, 2019) evaluates informed  
 63 hypotheses for structural equation models. Outside of R, the Fortran 90 program **BIEMS**  
 64 (Joris Mulder, Hoijtink, & de Leeuw, 2012) evaluates informed hypotheses for multivariate  
 65 linear models such as MANOVA, repeated measures, and multivariate regression. All these  
 66 packages rely on one of two implementations of the encompassing prior approach (Klugkist,  
 67 Kato, & Hoijtink, 2005; Sedransk, Monahan, & Chiu, 1985) to approximate order

68 constrained Bayes factors: the unconditional encompassing method (Klugkist et al., 2005 ;  
69 Hoijtink, 2011; Hoijtink et al., 2008) and the conditional encompassing method (Gu,  
70 Mulder, Deković, & Hoijtink, 2014; Laudy, 2006; Joris Mulder, 2014; J. Mulder, 2016; J.  
71 Mulder et al., 2009). Even though the encompassing prior approach is currently the most  
72 common method to evaluate informed hypotheses, it becomes increasingly unreliable and  
73 inefficient as the number of restrictions increases or the parameter space of the restricted  
74 model decreases (Sarafoglou et al., 2021). For instance, simulation studies conducted by  
75 Sarafoglou et al. (2021) have illustrated that the unconditional encompassing approach is  
76 not able to produce Bayes factors when hypotheses with a large number of constrained  
77 parameters are considered (i.e., they considered 18 categories). For hypotheses with fewer  
78 categories (i.e., 5 or 6), the method worked well when the data provided either weak or  
79 moderate evidence in favor of or against the informed hypothesis. However, when the data  
80 provided extreme evidence against the predicted constraints, the method again failed to  
81 compute Bayes factors.

82 As alternative to the encompassing prior approach, Sarafoglou et al. (2021) recently  
83 proposed a bridge sampling routine (Bennett, 1976; Meng & Wong, 1996) that computes  
84 Bayes factors for informed hypotheses more reliably and efficiently. This routine is  
85 implemented in **multibridge** (<https://CRAN.R-project.org/package=multibridge>) and is  
86 suitable to evaluate inequality constraints for multinomial and binomial models as well as  
87 combinations between equality and inequality constraints.

88 Here we showcase how the proposed bridge sampling routine by Sarafoglou et al.  
89 (2021) can be performed with **multibridge**. In the remainder of this article, we will  
90 introduce the package and its functionalities and describe the methods used to compute  
91 the informed hypotheses in binomial and multinomial models. We will illustrate its core  
92 functions using three examples and end with a brief discussion and future directions.

## Multibridge

The general workflow of **multibridge** is illustrated in Figure 1. The core functions of **multibridge**, that is `mult_bf_informed` and `binom_bf_informed`, return the Bayes factor estimate in favor of or against the informed hypothesis. To compute a Bayes factor, the core functions require the observed counts, the informed hypothesis, the parameters of the prior distribution under  $\mathcal{H}_e$ , and the category labels. An overview of the basic required arguments of the two core functions are provided in Table 1.

When calling `mult_bf_informed` or `binom_bf_informed`, the user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the  $\alpha$  and  $\beta$  parameters of the binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions then return the estimated Bayes factor for the informed hypothesis relative to the encompassing hypothesis that imposes no constraints on the category proportions or the null hypothesis which states that all category proportions are equal. Based on these results different S3 methods can be used to get more detailed information on the individual components. For instance, users can extract the Bayes factor with the `bayes_factor`-method, visualize the posterior parameter estimates under the encompassing hypothesis using the `plot`-method, or get more detailed information on how the Bayes factor is composed using the `summary`-method. Table 2 summarizes all S3 methods currently available in **multibridge**.

## Supported Hypotheses

The following hypotheses are supported in **multibridge**. Users can test hypotheses on equality and inequality constraints among parameters (left column in Figure 2). We consider inequality constraints, for instance, in Example 3 of this manuscript, when we test whether the probability to violate stochastic dominance decreases for persons with higher

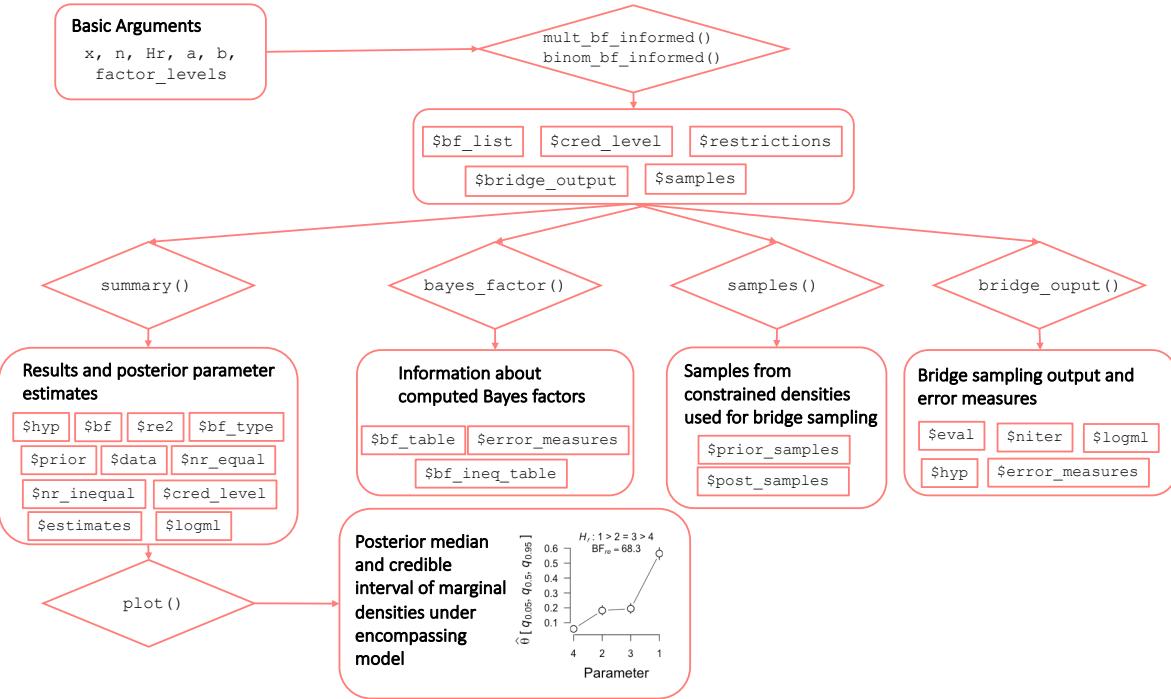
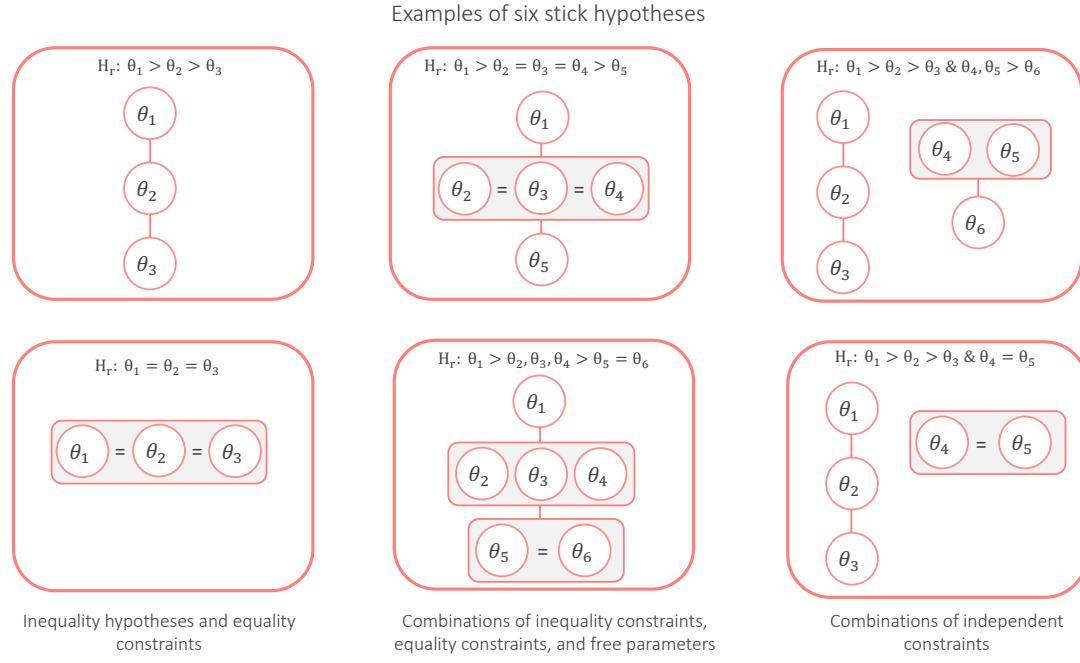


Figure 1. The **multibridge** workflow. The functions `mult_bf_informed` or `binom_bf_informed` return the estimated Bayes factor for the informed hypothesis relative to the encompassing or the null hypothesis. Based on these results different S3 methods can be used to get more detailed information on the individual components of the analysis (e.g., `summary`, `bayes_factor`), and parameter estimates of the encompassing distribution (`plot`).

118 education levels (Myung, Karabatsos, & Iverson, 2005).

119 Additionally, **multibridge** supports the evaluation of combinations of equality  
 120 constraints, inequality constraints, and free parameters (middle column). As an example,  
 121 the hypothesis in the top middle panel identifies a largest parameter ( $\theta_1$ ) and a smallest  
 122 parameter ( $\theta_5$ ), and equates the remaining parameters ( $\theta_2 = \theta_3 = \theta_4$ ). Combinations of  
 123 constraints are considered, for instance, in Example 2 of this manuscript. Based on Nuijten  
 124 et al. (2016) we test whether the proportion of statistical reporting errors is higher for  
 125 articles published in the *Journal of Personality and Social Psychology* (JPSP) than for  
 126 articles published in seven other high-profile psychology journals.

The package also supports the computation of Bayes factors for multiple independent constraints, representing, for instance, two main effects (right column). For instance, the hypothesis in the bottom right panel describes an inequality constraint on the first three category proportions ( $\theta_1 > \theta_2 > \theta_3$ ) and an equality constraint on the fourth and fifth category proportion ( $\theta_4 = \theta_5$ ).



*Figure 2.* **multibridge** supports informed hypotheses including inequality and equality constraints (left column), combinations of inequality and equality constraints and free parameters (middle column), and multiple independent constraints (right column). Parameters with larger values appear higher in the drawing. A prerequisite of **multibridge** is that all elements within a constraint can be arranged as a linearly ordered set.

An important requirement for the hypotheses supported in **multibridge** is that

within each independent constraint, all elements are arranged as a linearly ordered set.

Elements can refer to individual parameters as shown in the top left panel of Figure 2. In

this example, for each pair of elements one precedes the other in the sequence (i.e.,  $\theta_1$

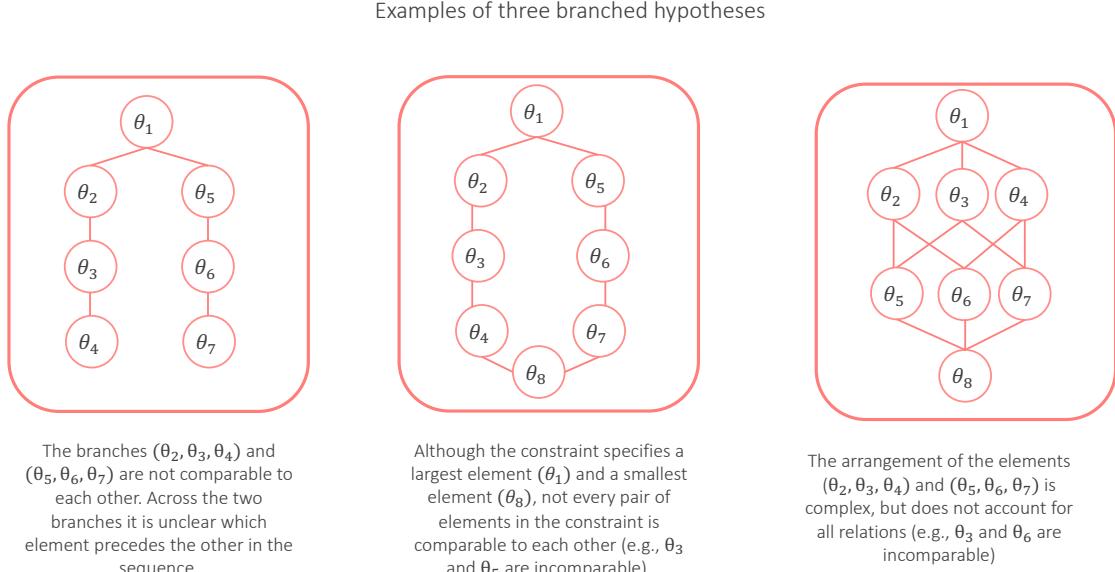
precedes  $\theta_2$  and  $\theta_2$  precedes  $\theta_3$ ). Elements can also refer to a group of equality constrained

parameters or a group of free parameters as shown in the middle panel of Figure 2. In the top middle panel, too, for each pair of elements one precedes the other in the sequence (e.g.,  $\theta_1$  precedes  $(\theta_2 = \theta_3 = \theta_4)$  and  $(\theta_2 = \theta_3 = \theta_4)$  precedes  $\theta_5$ ). That is, if the constraint was to be drawn as a Hasse diagram or specified as a character vector, the constrained elements should string together like a chain, ranging from the smallest element to the largest. We refer to these hypotheses as “stick hypotheses”.

Conversely, “branched hypotheses”, are hypotheses in which elements are not arranged as a linearly ordered set but as a partial order, meaning that some but not all pairs of elements precede one another. These hypotheses are *not* supported in **multibridge**. Examples for branched hypotheses are shown in Figure 3. For instance, the hypothesis illustrated in the left panel states that  $\theta_1$  precedes all other parameters. In addition, the hypothesis orders the branches  $(\theta_2, \theta_3, \theta_4)$  and  $(\theta_5, \theta_6, \theta_7)$ . However, it remains unclear whether, for instance,  $\theta_3$  or  $\theta_5$  precede the other in the sequence. Thus, not all pairs of elements are comparable. Similarly, in all three examples of branched hypotheses it is unclear whether  $\theta_3$  precedes or follows  $\theta_6$ . Researchers whose theories give rise to branched hypotheses and wish to test them can do so using one of the alternative R packages, for instance, **multinomineq** by Heck and Davis-Stober (2019).

When an informed hypothesis includes combinations of equality and inequality constraints, the core functions in **multibridge** split the hypothesis to compute Bayes factors separately for imposed equality constraints (for which the Bayes factor has an analytic solution) and inequality constraints (for which the Bayes factor is estimated using bridge sampling). Hence, for hypotheses that include combinations of equality and inequality constraints the `bayes_factor` method separately returns the Bayes factor for the equality constraints and the conditional Bayes factor for the inequality constraints given the equality constraints.

The informed hypothesis `Hr` can be conveniently specified as a string or a character



*Figure 3.* Examples of three hypotheses in which elements in a constraint are arranged as a partial order. In each panel there exist elements that are not comparable with each other, that is, for which neither element precedes the other in the sequence. The partial order shows itself in the branching of the Hasse diagram. These branched hypotheses are currently not supported in **multibridge**.

<sup>163</sup> vector describing the relations among the category proportions. A simple ordering of three category proportions,  $\theta_1 > \theta_2 > \theta_3$ , can be specified either as `c("t1", ">", "t2", ">, "t3")`, or as `"t1 > t2 > t3"`. To assign labels to the parameters, they must be passed to the argument `factor_levels`. **multibridge** then assumes that the order within the category labels correspond to the order of the data vector. Alternatively, the informed hypotheses can be specified using indices (e.g., `"1 > 2 > 3"`). To avoid circularity, an index or category label can be used only once within an informed hypothesis.

Table 1

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

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

170 Signs permitted to specify informed hypotheses are the `$<$ ''-sign and>''-sign for`

171 inequality constraints, the `=$$''-sign for equality constraints, the,''-sign for`

172 parameters that vary freely within a constraint, and the “&”-sign to connect multiple

173 independent constraints. For instance, the informed hypothesis in the top right panel in

174 Figure 2, that is, "t1 > t2 > t3 & t4 , t5 > t6", states that t1 is bigger than t2,

175 and that t2, is bigger than t3. In addition, the hypothesis states that t4 and t5 are bigger

176 than t6, with no further constraints imposed among t4 and t5.

```
x <- c(2, 1, 5, 4, 1, 8)
n <- c(10, 7, 13, 7, 9, 14)
a <- b <- c(1, 1, 1, 1, 1, 1)

# Testing ordinal relations and equality constraints
mult_bf_informed(x=x, a=a, Hr='1 > 2 > 3')
mult_bf_informed(x=x, a=a, Hr='1 = 2 = 3')

binom_bf_informed(x=x, n=n, a=a, b=b, Hr='1 > 2 > 3')
binom_bf_informed(x=x, n=n, a=a, b=b, Hr='1 = 2 = 3')

# Testing combinations of ordinal constraints,
# equality constraints, and free parameters
mult_bf_informed(x=x, a=a, Hr='1 = 2 = 3 > 4 > 5 = 6')
binom_bf_informed(x=x, n=n, a=a, b=b, Hr='1 < 2 , 3 , 4 < 5 = 6')

# Testing combinations of independent constraints
mult_bf_informed(x=x, a=a, Hr='1 > 2 > 3 & 4 , 5 = 6')
binom_bf_informed(x=x, n=n, a=a, b=b, Hr='1 > 2 > 3 & 5 = 6')
```

177 When testing equality constrained hypotheses, users should be aware that there is a

178 difference between assuming equality of category proportions and adding categories

179 together, that is, the hypothesis  $\mathcal{H}_r : \theta_1 = \theta_2 > \theta_3 = \theta_4$  differs from the hypothesis

180  $\mathcal{H}_r : \theta_1 + \theta_2 > \theta_3 + \theta_4$ . The first hypothesis concerns four category proportions of which two

181 pairs are expected to be equal; as a result, we assign a  $K = 4$  Dirichlet prior to this

182 distribution. The second hypothesis concerns only two categories since we assume that  $\theta_1$

183 and  $\theta_2$  belong to one group and  $\theta_3$  and  $\theta_4$  belong to the other. Consequently, one assigns a

184  $K = 2$  Dirichlet prior to this distribution. Therefore, to test the second hypothesis, the

185 respective counts of the categories should first be combined and the analysis should be

186 performed on the basis of these new data.

```
# Merging categories or setting them equal do not yield the same results

# Hr: t1 = t2 < t3 = t4

x <- c(20, 7, 5, 9)
a <- c(1, 1, 1, 1)

summary(mult_bf_informed(x=x, a=a, Hr='1 = 2 > 3 = 4'))$bf

# Hr: t1 + t2 < t3 + t4

x <- c(20 + 7, 5 + 9)
a <- c(1, 1)

summary(mult_bf_informed(x=x, a=a, Hr='1 > 2'))$bf
```

Table 2

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

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

187 In **multibridge**, the functions `mult_bf_informed` and `binom_bf_informed` perform

188 all necessary analysis steps. Other available functions compute Bayes factors for

189 hypotheses that postulate only equality or only inequality constraints, and draw from

190 constrained multinomial distributions and distributions of multiple independent binomials.

191 A list of all currently available functions and data sets is given in Table 3.

Table 3

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

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

192

## Methodological Background

193 In this section we provide background information on the methods implemented in  
 194 **multibridge**. Specifically, we formalize multinomial models and models that feature  
 195 independent binomial probabilities, and define Bayes factors for the Bayesian multinomial  
 196 test and testing equality of multiple independent binomial probabilities. Furthermore, the  
 197 section discusses the influence of priors on the Bayes factors, illustrates how to compute  
 198 posterior model probabilities and how to compare two informed hypotheses with each other,  
 199 and provides a non-technical introduction into the bridge sampling routine implemented in  
 200 **multibridge**. Mathematical details of the methods and principles discussed here can be  
 201 found in Sarafoglou et al. (2021) and Gronau et al. (2017).

202 In the binomial model, we assume that the elements in the vector of successes  $\mathbf{x}$  and  
 203 the elements in the vector of total number of observations  $\mathbf{n}$  in the  $K$  categories follow  
 204 independent binomial distributions  $\mathbf{x} \sim \prod_{k=1}^K \text{Binomial}(\theta_k, n_k)$ , where  $\theta_k$  is the  $k$ th category  
 205 proportion. From this distribution we can derive the likelihood of the data given the  
 206 parameters:

$$p(\mathbf{x} | \boldsymbol{\theta}) = \prod_{k=1}^K \binom{n_k}{x_k} \theta_k^{x_k} (1 - \theta_k)^{n_k - x_k}.$$

207 The parameter vector of the binomial success probabilities  $\boldsymbol{\theta}$  contains the underlying  
 208 category proportions and assume that categories are independent. Therefore, a suitable  
 209 choice for a prior distribution for  $\boldsymbol{\theta}$  is a vector of independent beta distributions with  
 210 parameters  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ , thus  $\boldsymbol{\theta} \sim \prod_{k=1}^K \text{Beta}(\alpha_k, \beta_k)$ . The prior density is given by:

$$p(\boldsymbol{\theta}) = \prod_{k=1}^K \frac{\theta_k^{\alpha_k-1} (1 - \theta_k)^{\beta_k-1}}{B(\alpha_k, \beta_k)},$$

211 where  $B(\alpha_k, \beta_k)$  is the beta function:

$$B(\alpha_k, \beta_k) = \frac{\Gamma(\alpha_k)\Gamma(\beta_k)}{\Gamma(\alpha_k + \beta_k)}.$$

212 The multinomial model generalizes the binomial model for cases where  $K > 2$ . In this  
 213 model, we assume that the vector of observations  $\mathbf{x}$  in the  $K$  categories follows a

214 multinomial distribution in which the parameters of interest,  $\boldsymbol{\theta}$ , represent the underlying  
215 category proportions, thus  $\mathbf{x} \sim \text{Multinomial}(x_+, \boldsymbol{\theta})$ , where  $x_+ = \sum_{k=1}^K x_k$ .

216 Since the  $K$  categories are dependent, the vector of probability parameters is  
217 constrained to sum to one, such that  $\sum_{k=1}^K (\theta_1, \dots, \theta_K) = 1$ . Therefore, a suitable choice for  
218 a prior distribution for  $\boldsymbol{\theta}$  is the Dirichlet distribution with concentration parameter vector  
219  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ :

$$p(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \theta_k^{\alpha_k - 1},$$

220 where  $B(\boldsymbol{\alpha})$  is the multivariate beta function:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}.$$

221 In `multibridge`, we have deliberately chosen to leave the priors at the original scale  
222 (i.e., in the probability space), because it makes it easier to express ones expectations  
223 about data patterns. Alternative approaches transform the model parameters into the  
224 probit space, which has the advantage that correlations can be specified for hierarchical  
225 models (e.g., as in the latent-trait model for multinomial processing tree models, Klauer,  
226 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015). However, these transformations  
227 make the development of priors more difficult and can lead to unintended consequences; for  
228 instance, a uniform prior on the probit scale does not translate to a uniform prior on the  
229 probability scale (as discussed in Heck & Wagenmakers, 2016).

### 230 Developing Suitable Prior Distributions

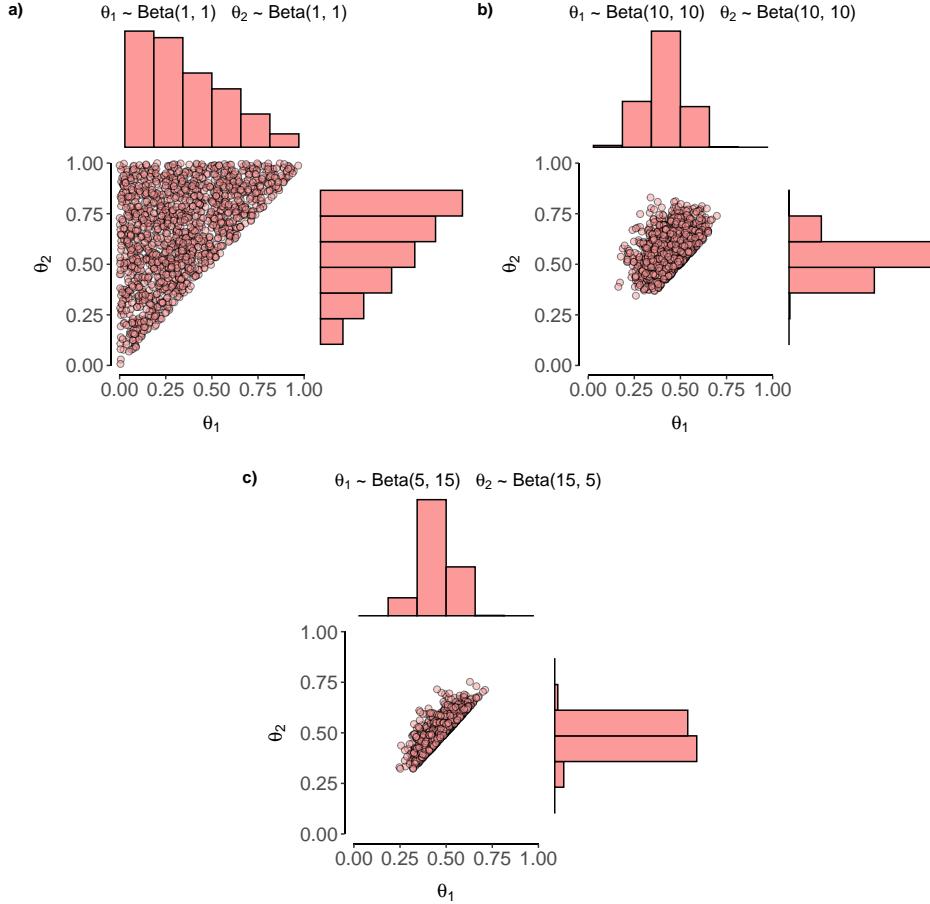
231 In the binomial and multinomial model, the concentration parameters have an  
232 intuitive interpretation. In the binomial model, the parameters  $\alpha_k$  can be interpreted as  
233 vector of *a priori* successes that observations fall within the various categories and  $\beta_k$  can  
234 be interpreted as vector of *a priori* failures. Likewise, in the multinomial model,  $\alpha_k$  can be  
235 interpreted as vector of *a priori* category counts. It follows, that the higher the number of

236 concentration parameters is, the information the prior contains and the more influence it  
237 has on parameter estimation and hypothesis testing.

238 Developing suitable prior distributions for Bayesian inference is a much discussed  
239 topic involving various theoretical and computational considerations (see Consonni,  
240 Fouskakis, Liseo, and Ntzoufras (2018) for a review paper on prior distributions for  
241 objective Bayesian analysis). Therefore, recommending approaches for developing  
242 appropriate prior distributions is, in our view, a difficult undertaking. In this section, we  
243 therefore present a selected subset of approaches that we consider particularly suitable for  
244 assigning adequate priors for the multiple binomials model and the multinomial model.

245 If researchers possess no knowledge or expectations about the plausible parameter  
246 values, a uniform distribution can be assigned across the parameter space. This prior  
247 assumes that before seeing the data, each category contains one observation, that is, all  
248 concentration parameters are set to one. A uniform prior distribution, puts equal  
249 probability mass on all permitted parameter values, similar to the adjusted priors for  
250 reparametrized models proposed by Heck and Wagenmakers (2016) (see Figure 4).  
251 However, **multibridge** allows priors to be set on the original scale.

252 We recommend incorporating prior knowledge into the models whenever possible.  
253 Based on theories, expert knowledge, or informed guesses, researchers often have  
254 expectations about plausible and implausible parameter values. In these cases, the prior  
255 should match these expectations (Lee & Vanpaemel, 2018). For instance, in the case of  
256 informed hypotheses, prior counts can be chosen to match a particular expected ordinal  
257 trend. To determine whether the chosen priors are consistent with the theory, researchers  
258 can visualize and assess prior predictive distributions, that is, the distribution of the model  
259 parameters and data patterns predicted by the priors (Gabry, Simpson, Vehtari,  
260 Betancourt, & Gelman, 2019; Schad, Betancourt, & Vasishth, 2021; Wagenmakers et al.,  
261 2021). The developed priors should reflect expectations about the parameters and make



*Figure 4.* The development of a prior distribution should be accompanied by a visual inspection of the prior predictive. Here we display three prior distributions on two binomial probabilities that are constrained to be  $\theta_1 < \theta_2$ . The uniform distribution (panel a) assigns equal mass to all permissible values of the constrained space. A symmetric prior (panel b) concentrates the mass in the center of the distribution. A prior describing a constraint in the opposite direction (panel c), puts most of the density along the diagonal.

262 sensible predictions. This is particularly important for Bayes factor hypothesis testing;  
263 when the purpose is parameter estimation, however, it may be more informative to assign  
264 prior parameter distributions that are relatively wide (e.g., Doorn et al., 2021).

265 Furthermore, one can choose the observed category counts of previous studies as  
266 priors for the current one, as is often suggested for replication studies and referred to as  
267 “Bayesian learning” (e.g., Verhagen & Wagenmakers, 2014). This approach constructs  
268 highly informative priors; instead of describing the new data as precisely as possible, the  
269 goal with this approach is quantify the additional knowledge gained by the new data.  
270 Finally, priors can be constructed using a fraction of the likelihood of the data while  
271 centering it on the mean of the parameter range (Gu, Mulder, & Hoijtink, 2018; Joris  
272 Mulder, 2014).

### 273 Bayes factor

274 **multibridge** features two different methods to compute Bayes factors: one method  
275 computes Bayes factors for equality constrained parameters (which can be computed  
276 analytically) and one method computes Bayes factors for inequality constrained parameters  
277 (which needs to be approximated). In cases where informed hypotheses feature  
278 combinations between inequality and equality constraints, **multibridge** computes the  
279 overall Bayes factor  $\text{BF}_{re}$  by multiplying the individual Bayes factors for both constraint  
280 types. This is motivated by the fact that the Bayes factor for combinations ( $\text{BF}_{re}$ ) will  
281 factor into a Bayes factor for the equality constraints ( $\text{BF}_{1e}$ ) and a conditional Bayes factor  
282 for the inequality constraints given the equality constraints ( $\text{BF}_{2e|1e}$ ). For instance, to  
283 evaluate the hypothesis  $\mathcal{H}_r : \theta_1 > \theta_2 = \theta_3$ , **multibridge** factors the Bayes factor as follows:

$$\text{BF}_{re} = \underbrace{\frac{p(\theta_1 > \theta_{23} \mid \theta_2 = \theta_3, \mathbf{x}, \mathcal{H}_e)}{p(\theta_1 > \theta_{23} \mid \theta_2 = \theta_3, \mathcal{H}_e)}}_{\text{BF}_{2e|1e}} \times \underbrace{\frac{p(\theta_2 = \theta_3 \mid \mathbf{x}, \mathcal{H}_e)}{p(\theta_2 = \theta_3 \mid \mathcal{H}_e)}}_{\text{BF}_{1e}},$$

284 where the subscript 1 denotes the hypothesis that only features equality constraints,

285 the subscript 2 denotes the hypothesis that only features inequality constraints, and

286  $p(\theta_1 > \theta_{23} \mid \theta_2 = \theta_3)$  refers to a Dirichlet integral, where the category proportions  $\theta_2$  and  $\theta_3$

287 are collapsed. See Sarafoglou et al. (2021) for the proof and a detailed account of this

288 method.

289 **Testing Equality Constraints.** For equality constrained binomial models

290 **multibridge** supports two null hypotheses, one stating that all parameters are equal and

291 one stating that all parameters are equal to a specific value. Both null hypotheses are

292 tested against an encompassing hypothesis. Under the encompassing hypothesis, we specify

293 a  $\text{Beta}(\alpha_k, \beta_k)$  prior on each of the  $\theta_k$  that yields the following marginal likelihood:

$$p(\mathbf{x} \mid \mathcal{H}_e) = \frac{\prod_{k=1}^K \binom{n_k}{x_k} \times \text{B}(x_k + \alpha_k, n_k - x_k + \beta_k)}{\prod_{k=1}^K \text{B}(\alpha_k, \beta_k)}.$$

294 Under the first null hypothesis which states that all binomial probabilities are set

295 equal without a constraint on a specific value, we collapse all individual  $\text{Beta}(\alpha_k, \beta_k)$  priors

296 and correct for the change in categories; if  $K$  categories are collapsed,  $K - 1$  is subtracted

297 from the concentration parameters. The resulting prior is a

298  $\text{Beta}(\alpha_+ - (K - 1), \beta_+ - (K - 1))$  distribution on  $\theta$ , where  $\alpha_+ = \sum_{k=1}^K \alpha_k$  and

299  $\beta_+ = \sum_{k=1}^K \beta_k$ . Hence, a  $\text{Beta}(1, 1)$  prior on each individual category proportion yields

300 again a  $\text{Beta}(1, 1)$  prior on the categories that are collapsed. When the prior is more

301 informative, say a  $\text{Beta}(2, 2)$  prior on three individual category proportions, it would result

302 in a  $\text{Beta}(4, 4)$  prior on  $\theta$  as the information available is added together. The

303 corresponding marginal likelihood takes the following form:

$$p(\mathbf{x} \mid \mathcal{H}_{01}) = \frac{\prod_{k=1}^K \binom{n_k}{x_k} \times \text{B}(x_+ + \alpha_+ - (K - 1), n_+ - x_+ + \beta_+ - (K - 1))}{\text{B}(\alpha_+ - (K - 1), \beta_+ - (K - 1))}.$$

We can now compute the Bayes factor  $\text{BF}_{01e}$  as follows:

$$\begin{aligned}\text{BF}_{0e} &= \frac{p(\mathbf{x} | \mathcal{H}_0)}{p(\mathbf{x} | \mathcal{H}_e)} \\ &= \frac{\frac{\prod_{k=1}^K \binom{n_k}{x_k} \times B(x_+ + \alpha_+ - (K-1), n_+ - x_+ + \beta_+ - (K-1))}{B(\alpha_+ - (K-1), \beta_+ - (K-1))}}{\frac{\prod_{k=1}^K \binom{n_k}{x_k} \times B(x_k + \alpha_k, n_k - x_k + \beta_k)}{\prod_{k=1}^K B(\alpha_k, \beta_k)}} \\ &= \frac{\prod_{k=1}^K B(x_+ + \alpha_+ - (K-1), n_+ - x_+ + \beta_+ - (K-1))}{\prod_{k=1}^K B(x_k + \alpha_k, n_k - x_k + \beta_k)} \times \frac{\prod_{k=1}^K B(\alpha_k, \beta_k)}{B(\alpha_+ - (K-1), \beta_+ - (K-1))}\end{aligned}$$

The second null hypothesis states that all binomial probabilities in a model are

assumed to be exactly equal *and* equal to a predicted value  $\theta_0$ . Under this hypothesis, the

prior reduces to a single point and the marginal likelihood simplifies to the likelihood

function:

$$p(\mathbf{x} | \mathcal{H}_{02}) = \theta_0^{x_+} (1 - \theta_0)^{n_+ - x_+} \times \prod_{k=1}^K \binom{n_k}{x_k}.$$

The Bayes factor for the second null hypothesis is then defined as:

$$\text{BF}_{02e} = \frac{\prod_{k=1}^K B(\alpha_k, \beta_k)}{\prod_{k=1}^K B(\alpha_k + x_k, \beta_k + n_k - x_k)} \times \theta_0^{x_+} (1 - \theta_0)^{n_+ - x_+}.$$

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

binomial probabilities.

```
x <- c(3, 4, 10, 11)
n <- c(15, 12, 12, 12)
a <- c(1, 1, 1, 1)
b <- c(1, 1, 1, 1)

# assuming all binomial proportions are equal
binom_bf_equality(x=x, n=n, a=a, b=b)

# assuming all binomial proportions are equal
# and equal to a predicted value

binom_bf_equality(x=x, n=n, a=a, b=b, p = 0.5)
```

311        The Bayes factor  $\text{BF}_{0e}$  for the multinomial test is defined as:

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

312        where  $\theta_{0k}$  represent the predicted category proportions (see Sarafoglou et al., 2021 for the  
 313        derivation). For multinomial models, under the null hypothesis, category probabilities can  
 314        either all be set equal (i.e., all category probabilities are  $\frac{1}{K}$ ) or can replaced with the  
 315        user-specified predicted values.

```
x <- c(3, 4, 10, 11)
a <- c(1, 1, 1, 1)
# assuming all category proportions are exactly equal
mult_bf_equality(x=x, a=a)
# specifying predicted values
mult_bf_equality(x=x, a=a, p = c(0.1, 0.1, 0.3, 0.5))
```

316        **Testing Inequality Constraints.** For inequality constrained binomial and  
 317        multinomial models, users can specify informed hypotheses that are either tested against a  
 318        null hypothesis postulating that all parameters are equal or against the encompassing  
 319        hypothesis which lets all parameters free to vary. Generally, to obtain the marginal  
 320        likelihood of the informed hypothesis, it is necessary to integrate over the restricted  
 321        parameter space, which is difficult to compute. As a solution to the problem of computing  
 322        marginal likelihood of the informed hypothesis, Klugkist et al. (2005) derived an identity  
 323        that defines the Bayes factor  $\text{BF}_{re}$  as the ratio of proportions of posterior and prior  
 324        parameter space consistent with the restriction. This identity forms the basis of the  
 325        encompassing prior approach. Recently, Sarafoglou et al. (2021) highlighted that these  
 326        proportions can be reinterpreted as the marginal likelihoods (i.e., the normalizing  
 327        constants) of the constrained posterior and constrained prior distribution. The prior  
 328        distribution consistent with the restriction takes the following form:

$$p(\boldsymbol{\theta} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e) \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{\int_{\mathcal{R}_e} p(\boldsymbol{\theta} \mid \mathcal{H}_r) d\boldsymbol{\theta}},$$

329 where  $\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$  is an indicator function that is one for parameter values in the that obey  
 330 the constrained and zero otherwise. The constrained posterior distribution of the  
 331 parameters under the informed hypothesis can be represented in the same way,

$$p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e) \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{\int_{\mathcal{R}_e} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{H}_r) d\boldsymbol{\theta}}.$$

332 The Klugkist identity (Klugkist et al., 2005) can be derived from the marginal likelihoods  
 333 of the two distributions as follows:

$$\text{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}^{\text{Marginal likelihood of constrained posterior distribution}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\text{Marginal likelihood of constrained prior distribution}}}.$$
(1)

334 The Klugkist identity made it possible to utilize numerical sampling methods such as  
 335 bridge sampling to compute the Bayes factor. The following section provides a conceptual  
 336 introduction to bridge sampling and how it is used in the context of evaluating informed  
 337 hypotheses.

### 338 Bridge Sampling Routine

339 The bridge sampling routine implemented in **multibridge** is a numerical method to  
 340 estimate the marginal likelihood of a target density (cf., Gronau et al., 2017; Overstall &  
 341 Forster, 2010). The identity used in bridge sampling is displayed in Equation 2; it considers  
 342 the unnormalized target density, a proposal density with known normalizing constant, and  
 343 an arbitrary bridge function. The numerator in Equation 2 describes the expected value of  
 344 the unnormalized target density evaluated with samples from the proposal density. The

345 denominator is the expected value of the proposal density and a bridge function evaluated  
 346 with samples from the target density. The bridge function serves the purpose of increasing  
 347 the overlap between the two densities, thus increasing the efficiency and accuracy of the  
 348 method. The bridge sampling identity can then be expressed as follows:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})} (p(\boldsymbol{\theta} \mid \mathcal{H}_e) \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r) h(\boldsymbol{\theta}))}{\mathbb{E}_{\text{prior}} (g(\boldsymbol{\theta}) h(\boldsymbol{\theta}))}, \quad (2)$$

349 where the term  $h(\boldsymbol{\theta})$  refers to the bridge function proposed by Meng and Wong (1996),  
 350  $g(\boldsymbol{\theta})$  refers to a proposal density (in this application we choose the multivariate normal  
 351 density), and  $p(\boldsymbol{\theta} \mid \mathcal{H}_e) \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$  is the unnormalized target density; in this case it  
 352 represents the part of the prior parameter space under the encompassing hypothesis that is  
 353 in accordance with the constraint. In the conventional application of bridge sampling, the  
 354 marginal likelihoods of the two competing hypotheses are estimated, that is, the marginal  
 355 likelihood of the informed hypothesis and the marginal likelihood of the encompassing  
 356 hypothesis. But on the basis of Equation 1, the routine implemented in **multibridge**  
 357 estimates the marginal likelihood of the restricted prior and restricted posterior density.

358 It should be noted that the bridge sampling algorithm implemented in **multibridge**  
 359 is an adapted version of the algorithm implemented in the R package **bridgesampling**  
 360 (Gronau, Singmann, & Wagenmakers, 2020) and allows for the specification of informed  
 361 hypotheses on probability vectors.<sup>1</sup>

362 A schematic representation of the bridge sampling routine is displayed in Figure 5.  
 363 To estimate the marginal likelihood, bridge sampling requires samples from the target  
 364 distribution, that is, the constrained Dirichlet distribution for multinomial models and  
 365 constrained beta distributions for binomial models, and samples from the proposal  
 366 distribution which in principle can be any distribution with a known marginal likelihood;

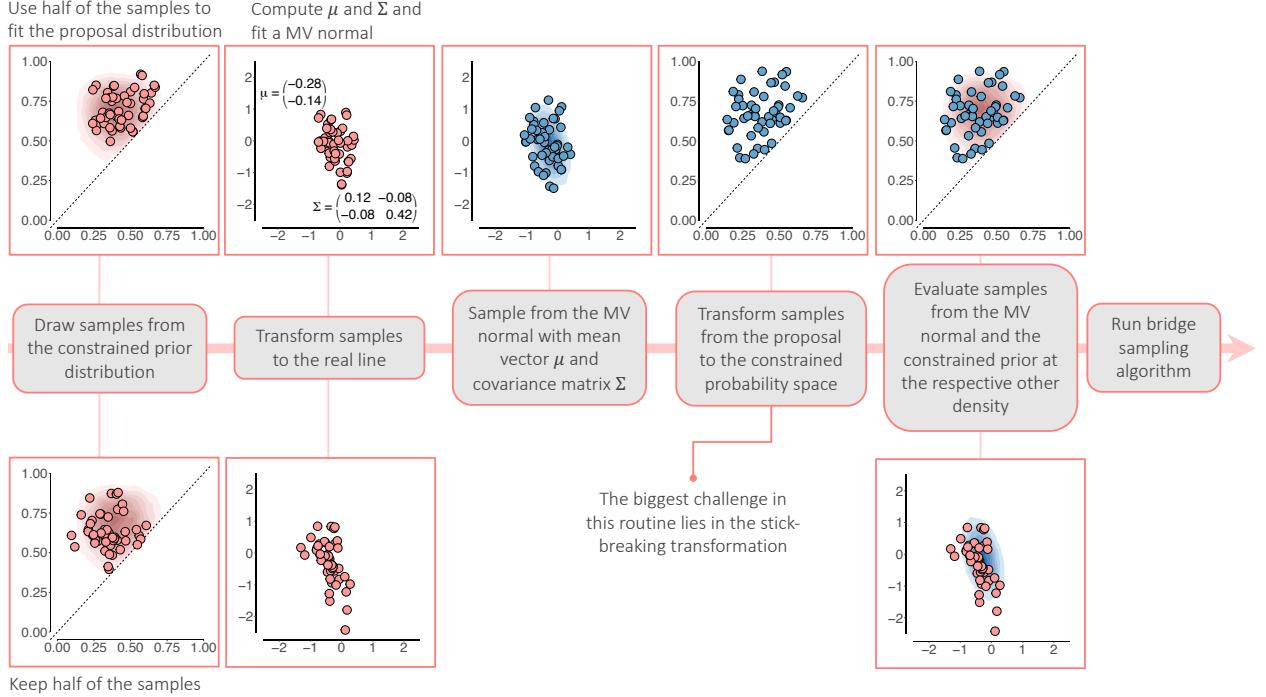
---

<sup>1</sup> 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.

367 in **multibridge** the proposal distribution is the multivariate normal distribution. Samples  
368 from the target distribution are generated using the Gibbs sampling algorithms proposed  
369 by Damien and Walker (2001). For binomial models, we apply the suggested Gibbs  
370 sampling algorithm for constrained beta distributions. In the case of the multinomial  
371 models, we apply an algorithm that simulates values from constrained Gamma  
372 distributions which are then transformed into Dirichlet random variables. To sample  
373 efficiently from these distributions, **multibridge** provides a C++ implementation of this  
374 algorithm. Samples from the proposal distribution are generated using the standard  
375 `rmvnorm`-function from the R package **mvtnorm** (Genz et al., 2020).

376 Despite the bridge function, the efficiency of the bridge sampling method is optimal  
377 only if the target and proposal distribution operate on the same parameter space and have  
378 sufficient overlap. We therefore probit transform the samples of the constrained  
379 distributions to move the samples from the probability space to the entire real line.  
380 Subsequently, we use half of these draws to construct the proposal distribution using the  
381 method of moments. Then, samples are drawn from the proposal density and transformed  
382 back into the probability space, ensuring that the samples correspond to the informed  
383 hypothesis. These transformed samples are then used to evaluate the unnormalized target  
384 density.

385 The numerator in Equation 2 evaluates the unnormalized density for the constrained  
386 prior distribution with samples from the proposal distribution. The denominator evaluates  
387 the normalized proposal distribution with samples from the constrained prior distribution.  
388 Since the optimal bridge function proposed by Meng and Wong (1996) contains the  
389 marginal likelihood of the target density –the quantity we wish to compute– an iterative  
390 scheme is applied to obtain the estimate. **multibridge** then runs the iterative scheme until  
391 the tolerance criterion suggested by Gronau et al. (2017) is reached. The sampling from  
392 the target and proposal distribution, the transformations and computational steps are  
393 performed automatically within the core functions of **multibridge**. The user only needs to



*Figure 5.* A schematic illustration of the steps taken to estimate the marginal likelihood of the constrained prior distribution of two binomial probabilities under  $\mathcal{H}_r : \theta_1 < \theta_2$ . As starting point, the routine requires samples from the constrained prior distribution (red). Following a transformation to the real line, a multivariate normal distribution (blue) is fit to half of the samples. The results from evaluating the samples from the multivariate normal distribution and the constrained prior distribution at the respective other density are needed to compute the expected values displayed in Equation 2. As final step, the bridge sampling algorithm estimates the marginal likelihood of the constrained prior distribution using an iterative scheme.

394 provide the functions with the data, a prior and a specification of the informed hypothesis.  
 395 As part of the standard output of `binom_bf_informed` and `mult_bf_informed`, the  
 396 functions return the bridge sampling estimate for the log marginal likelihood of the target  
 397 distribution, its associate relative mean square error and the number of iterations needed  
 398 to until the bridge sampling estimator reached the tolerance criterion.

399 To summarize, in order to implement the bridge sampling method we only need to be  
 400 able to sample from the constrained densities. Crucially, when using bridge sampling, it  
 401 does not matter how small the constrained parameter space is in proportion to the  
 402 encompassing density. This gives the method a decisive advantage over the encompassing  
 403 prior approach in terms of accuracy and efficiency especially (1) when binomial and  
 404 multinomial models with moderate to high number of categories (i.e.,  $K > 10$ ) are  
 405 evaluated and (2) when relatively little posterior mass falls in the constrained parameter  
 406 space.

## 407 Stick-Breaking Transformation

408 The bridge sampling routine in `multibridge` uses the multivariate normal  
 409 distribution as proposal distribution, which requires moving samples from target  
 410 distribution to the real line and conversely, moving samples from the real line to the  
 411 ordered probability space. Crucially, the transformation needs to retain the ordering of the  
 412 parameters, that is, it needs to take into account the lower bound and the upper bound of  
 413 each parameter. Elements from the real line to the ordered probability space are then  
 414 transformed as follows:

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

415 where  $\xi_k$  is  $k$ th the element on the real line,  $\Phi$  is the cumulative density function of a  
 416 standard normal and  $u_k$  and  $l_k$  are the upper and lower bounds of  $\xi_k$ , respectively. The

<sup>417</sup> largest element is simply the remainder of the stick. The inverse transformation is given by

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

<sup>418</sup> where  $\Phi^{-1}$  denote the inverse cumulative density function. To determine the bounds,

<sup>419</sup> **multibridge** uses a probit transformation, as proposed in Sarafoglou et al. (2021), which  
<sup>420</sup> transforms the elements by moving from the smallest to the largest value. A schematic  
<sup>421</sup> illustration of the stick-breaking transformation is given in Figure 6, detailed technical  
<sup>422</sup> details of the transformation are provided in the appendix.

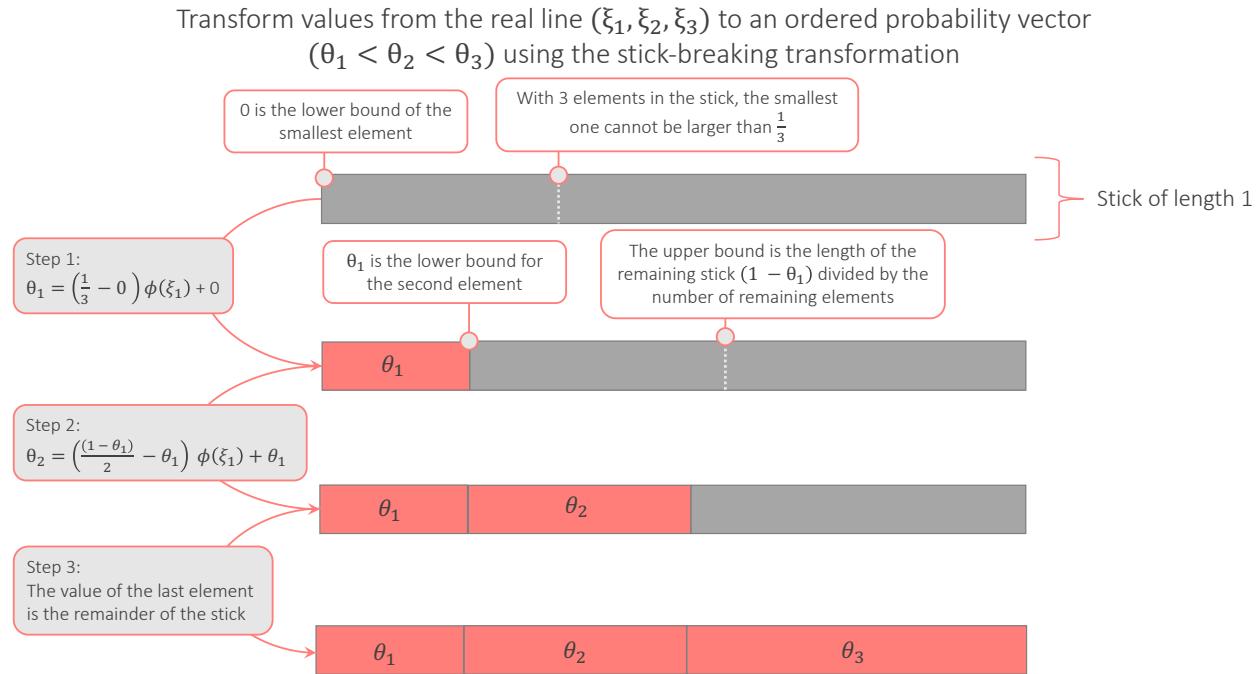


Figure 6. The stick-breaking transformation of elements on the real line to the ordered probability space. The stick-breaking transformation moves from the smallest to the largest element to determine its lower and upper bounds.

<sup>423</sup> To perform the transformation from a parameter vector on the real line to an ordered  
<sup>424</sup> probability vector, we need to determine the lower and upper bound of each parameter.  
<sup>425</sup> Consider an increasing trend of four parameters, that is,  $\theta_1 < \theta_2 < \theta_3$ . The lower bound for  
<sup>426</sup> for the smallest element in the parameter vector,  $\theta_1$ , is 0. For  $\theta_2$  and  $\theta_3$  the lower bound is

427 the preceding element in the vector. That is, the lower bound for  $\theta_2$  is  $\theta_1$ , lower bound for  
428  $\theta_3$  is  $\theta_2$ .

429 This definition holds for both binomial models and multinomial models. Differences  
430 in these two models appear only when determining the upper bound for each parameter.  
431 For binomial models, the upper bound for each parameter is 1. For multinomial models,  
432 due to the sum-to-one constraint the upper bounds need to be computed differently. As  
433 proposed in Frigyik, Kapila, and Gupta (2010) and Stan Development Team (2020) we  
434 represent  $\boldsymbol{\theta}$  as unit-length stick which we subsequently divide into as many elements as  
435 there are parameters in the constraint (Stan Development Team, 2020). In this approach,  
436 the upper bounds are derived from on the values of smaller elements as well as on the  
437 number of remaining larger parameters in the stick. Concretely, for the smallest element in  
438 the parameter vector,  $\theta_1$ , the upper bound is  $\frac{1}{3}$ ; if this element were larger than that it  
439 would be impossible to create a probability vector with increasing values. For  $\theta_2$  and  $\theta_3$  the  
440 upper bound is the proportion of the unit-length stick that has not yet been accounted for  
441 in the transformation divided by the number of parameters in the remaining stick. For  
442 instance, the upper bound for  $\theta_2$  is defined as  $\frac{1-\theta_1}{2}$ . This transformation allows us to  
443 effectively transform elements from the real line to an constrained probability space and is  
444 therefore a main component of the bridge sampling algorithm.

445 The drawback of this transformation is, however, that it can only be performed if all  
446 elements in the constraint are arranged as a linearly ordered set, thus, only works for “stick  
447 hypotheses”. For hypotheses in which elements in a constraint are arranged as a partial  
448 order, the assumption is violated that for a given parameter smaller elements and the  
449 number of parameters in the remaining stick determine their upper bound.

450 **Poster Model Probabilites, and Bayes Factor Transitivity**

451 Consider a scenario where researchers entertain more than two hypotheses that they  
 452 wish to compare. For instance, they may entertain two informed hypotheses  $\mathcal{H}_{r1}$  and  $\mathcal{H}_{r2}$   
 453 as well as a null hypothesis  $\mathcal{H}_0$  and the encompassing hypothesis  $\mathcal{H}_e$ . An overview of the  
 454 relative plausibility of all  $M = 4$  models simultaneously may be obtained by presenting the  
 455 posterior model probabilities for all hypotheses,  $p(\mathcal{H}_i | x)$ ,  $i = 1, \dots, 4$  Berger and Molina  
 456 (2005). Posterior model probabilities are not automatically computed in **multibridge**;  
 457 however, after computing the individual Bayes factors, the posterior model probabilities  
 458 can be obtained easily. Denoting the prior model probability for hypothesis  $\mathcal{H}_{r1}$  by  $p(\mathcal{H}_{r1})$ ,  
 459 the posterior model probability  $p(\mathcal{H}_{r1} | \mathbf{x})$  is given by:

$$p(\mathcal{H}_{r1} | \mathbf{x}) = \frac{\frac{p(\mathbf{x} | \mathcal{H}_{r1})}{p(\mathbf{x} | \mathcal{H}_e)} \times p(\mathcal{H}_{r1})}{\sum_{i=1}^M \frac{p(\mathbf{x} | \mathcal{H}_i)}{p(\mathbf{x} | \mathcal{H}_e)} \times p(\mathcal{H}_i)}.$$

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

$$p(\mathcal{H}_{r1} | \mathbf{x}) = \frac{\text{BF}_{r1e}}{\text{BF}_{r1e} + \text{BF}_{r2e} + \text{BF}_{0e} + \text{BF}_{ee}},$$

461 where  $\text{BF}_{ee}$  equals 1. In R, the posterior model probabilities can be computed as follows:

```
# posterior model probability of Hr1 given three alternative hypotheses
p_Hr1_x <- bfr1e/(bfr1e + bfr2e + bf0e + 1) # bfee = 1
```

462 Posterior model probabilities are useful for comparing multiple hypotheses; however,  
 463 they are relative quantities that change depending on which other hypotheses are included  
 464 in the comparison. Thus, hypotheses that describe the data poorly may have high posterior  
 465 model probabilities if the other hypotheses in the comparison set provide even worse  
 466 descriptions of the data. In order to gain insight into whether a hypothesis describes the  
 467 data adequately, we therefore consider so-called bookend hypotheses along with

<sup>468</sup> theory-informed hypotheses. That is, we include a hypothesis that maximally constrains  
<sup>469</sup> the parameter space (such as a point-null hypothesis  $\mathcal{H}_0$ ) and the encompassing hypothesis  
<sup>470</sup>  $\mathcal{H}_e$  that does not constrain the parameter space (in this case, that makes no ordinal  
<sup>471</sup> predictions, Lee & Vanpaemel, 2018). A hypothesis is then considered adequate if it  
<sup>472</sup> outperforms these two bookend models.

<sup>473</sup> In addition to posterior model probabilities, Bayes factors can also be calculated  
<sup>474</sup> directly between two informed hypotheses. The comparison of any two informed hypotheses  
<sup>475</sup> with one another follows from the fact that Bayes factors are transitive. For instance, the  
<sup>476</sup> Bayes factor comparison between two informed hypotheses  $\mathcal{H}_{r1}$  and  $\mathcal{H}_{r2}$  can be obtained  
<sup>477</sup> by first computing  $\text{BF}_{r1e}$  and  $\text{BF}_{r2e}$ , and then dividing out the common hypothesis  $\mathcal{H}_e$ :

$$\text{BF}_{r1r2} = \frac{\text{BF}_{r1e}}{\text{BF}_{r2e}}.$$

<sup>478</sup> For this comparison to be feasible, the encompassing hypotheses must be identical, that is,  
<sup>479</sup> the same prior distribution must be assigned to the category proportions.

## <sup>480</sup> Prior Sensitivity

<sup>481</sup> Bayesian hypothesis testing has been criticised as the priors exert too much influence  
<sup>482</sup> on the Bayes factors (e.g., Kass & Raftery, 1995). That is, even if the data are informative  
<sup>483</sup> enough to overwhelm the prior for parameter estimation, priors can still influence the  
<sup>484</sup> Bayes factors. The development of suitable priors is thus an important part of Bayesian  
<sup>485</sup> hypothesis testing.

<sup>486</sup> But even priors that are justified by theory are to a certain degree arbitrary. For  
<sup>487</sup> instance, if one expects an increasing trend in the data, the parameters in the prior can be  
<sup>488</sup> chosen to reflect that trend. The exact number of *a priori* category counts, however, is at  
<sup>489</sup> the discretion of the analyst. It is therefore considered good research practice to conduct a  
<sup>490</sup> sensitivity analysis on the final results Lee & Vanpaemel (2018). In a sensitivity analysis, a  
<sup>491</sup> set of plausible priors are determined in addition to the prior chosen in the main analysis

492 for which the Bayes factors are calculated. The range of Bayes factors then gives an  
 493 indication of the extent to which the results are fragile or robust to different modeling  
 494 choices. In general, the prior on which the final analysis is performed as well as the set of  
 495 priors used to conduct the sensitivity analysis should be determined and preregistered  
 496 before seeing the data to ensure a fair comparison of the hypotheses of interest.

497 **Usage and Examples**

498 In the following, we will outline three examples on how to use **multibridge** to  
 499 compare an informed hypothesis to a null or encompassing hypothesis. The first example  
 500 concerns multinomial data and the second and third example concerns independent  
 501 binomial data. Additional examples are available as vignettes (see  
 502 `\texttt{vignette(package = "multibridge")}`).

503 The two core functions of **multibridge**—`mult_bf_informed` and the  
 504 `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)
```

505 **Example 1: Applying A Benford Test to Greek Fiscal Data**

506 The first-digit phenomenon, otherwise known as Benford's law (Benford, 1938;  
 507 Newcomb, 1881) states that the expected proportion of leading digits in empirical data can  
 508 be formalized as follows: for any given leading digit  $d$ ,  $d = (1, \dots, 9)$  the expected  
 509 proportion is approximately equal to

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

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

in 30.1% of the cases, and leading digit 2 in 17.61% of the cases; leading digit 9 is the least frequent digit with an expected proportion of only 4.58% (see Table 4 for an overview of the expected proportions). Empirical data for which this relationship holds include population sizes, death rates, baseball statistics, atomic weights of elements, and physical constants (Benford, 1938). In contrast, artificially generated data, such as telephone numbers, do in general not obey Benford’s law (Hill, 1995). Given that Benford’s law applies to empirical data but not artificially generated data, a so-called Benford test can be used in fields like accounting and auditing to check for indications for poor data quality (for an overview, see e.g., Durtschi, Hillison, & Pacini, 2004; 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 four possible Bayesian adaptations of the Benford test. In a first scenario we simply conduct a Bayesian multinomial test in which we test the point-null hypothesis  $\mathcal{H}_0$  which predicts a Benford distribution. In a second scenario we test the informed hypothesis  $\mathcal{H}_{r1}$ , which predicts a decreasing trend in the proportions of leading digits. The hypothesis  $\mathcal{H}_{r1}$  exerts considerably more constraint than  $\mathcal{H}_e$  and provides a more sensitive test if our primary goal is to test whether data comply with Benford’s law or whether the data follow a similar but different trend. In the next two scenarios, our main goal is to identify fabricated data. The third scenario therefore tests the null hypothesis against the hypothesis that all proportions occur equally often. This hypothesis  $\mathcal{H}_{r2}$  could be considered if it is suspected that the data were generated randomly or could serve as a bookend comparison hypothesis as it maximally constraints the parameter space. In a fourth scenario we test a hypothesis which predicts a trend that is characteristic for manipulated data. This hypothesis, which we denote as  $\mathcal{H}_{r3}$ , could be derived from empirical research on fraud or be based on observed patterns from former fraud cases. For instance, Hill (1995) instructed students to produce a series of random numbers; in the resulting data the proportion of the leading digit 1 occurred most often and the digits 8

539 and 9 occurred least often which is consistent with the general pattern of Benford’s law.  
540 However, the proportion for the remaining leading digits were approximately equal. Note  
541 that the predicted distribution derived from Hill (1995) is not currently used as a test to  
542 detect fraud, however, for the sake of simplicity, we assume that this pattern could be an  
543 indication of manipulated auditing data. All hypotheses will be tested against the  
544 encompassing hypothesis  $\mathcal{H}_e$ , which too serves as a bookend comparison hypothesis, and  
545 which imposes no constraints on the proportion of leading digits.

546       **Data and Hypothesis.** The data we use to illustrate the computation of Bayes

547 factors were originally published by the European statistics agency “Eurostat” and served  
548 as basis for reviewing the adherence to the Stability and Growth Pact of EU member  
549 states. Rauch, Götsche, Brähler, and Engel (2011) conducted a Benford test on data  
550 related to budget deficit criteria, that is, public deficit, public debt and gross national  
551 products. The data used for this example features the proportion of first digits from Greek  
552 fiscal data in the years between 1999 and 2010; a total of  $N = 1,497$  numerical data were  
553 included in the analysis. We choose this data, since the Greek government deficit and debt  
554 statistics states has been repeatedly criticized by the European Commission in this time  
555 span (European Commision, 2004, 2010). In particular, the commission has accused the  
556 Greek statistical authorities to have misreported deficit and debt statistics. For further  
557 details on the data set see Rauch et al. (2011). The observed and expected proportions are  
558 displayed in Table 4; the expected proportions versus the posterior parameter estimates  
559 under the encompassing hypothesis are displayed in Figure 7.

Table 4

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

Leading digit	Observed Counts	Observed Proportions	Expected	Proportion- tions: 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,  $\theta_1, \dots, \theta_K$ , reflects the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9. Each of the hypotheses above will be tested against the encompassing hypothesis  $\mathcal{H}_e$  which imposes no constraints on the parameters. The hypotheses introduced above can then be

formalized as follows:

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

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

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

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

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

560       **Method.** Both  $\text{BF}_{0e}$  and  $\text{BF}_{r2e}$  may be readily computed by means of a Bayesian  
 561 multinomial test which is implemented in the function `mult_bf_equality`. This function  
 562 requires (1) a vector with observed counts, (2) a vector with concentration parameters of  
 563 the Dirichlet prior distribution under  $\mathcal{H}_e$ , and (3) the vector of expected proportions under  
 564  $\mathcal{H}_0$  and under  $\mathcal{H}_{r2}$ . In this example, we do not incorporate specific expectations about the  
 565 distribution of leading digits in the Greek fiscal data and therefore assign a uniform  
 566 Dirichlet distribution to the proportion of leading digits. That is, we set all concentration  
 567 parameters under  $\mathcal{H}_e$  to 1 (i.e., we assign  $\boldsymbol{\theta}$  a uniform Dirichlet prior distribution). This  
 568 prior supports all possible points equally, meaning that, if the data were completely  
 569 random, none of the hypotheses under consideration should be favored over the other.

```
# Observed counts
x <- c(509, 353, 177, 114, 77, 77, 53, 73, 64)

# Prior specification for Dirichlet prior distribution under H_e
a <- c(1, 1, 1, 1, 1, 1, 1, 1, 1)

# Expected proportions for H_0 and H_r2
p0 <- log10((1:9 + 1)/1:9)
pr2 <- c(1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9)

# Execute the analysis
results_H0_He <- mult_bf_equality(x = x, a = a, p = p0)
```

```

results_Hr2_He <- mult_bf_equality(x = x, a = a, p = pr2)

logBFe0 <- results_H0_He$bf$LogBFe0

logBFer2 <- results_Hr2_He$bf$LogBFe0

```

The hypotheses  $\mathcal{H}_{r1}$  and  $\mathcal{H}_{r3}$  contain inequality constraints, and this necessitates the use of the function `mult_bf_informed` to compute the Bayes factors  $\text{BF}_{r1e}$  and  $\text{BF}_{r3e}$ . This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under  $\mathcal{H}_e$ , (3) labels for the categories of interest (i.e., leading digits), and (4) the informed hypothesis  $\mathcal{H}_{r1}$  or  $\mathcal{H}_{r3}$  (e.g., as a string). In addition to the basic required arguments, we use two additional arguments here. The first argument sets the Bayes factor type, that is, whether the output should print the Bayes factor in favor of the informed hypothesis (i.e.,  $\text{BF}_{re}$ ) or in favor of the encompassing hypothesis (i.e.,  $\text{BF}_{er}$ ). It is also possible to compute the log Bayes factor in favor of the hypothesis, which is the setting we choose for this example. The purpose of the second argument `seed` is to make the results reproducible:

```

bf_type = 'LogBFer', seed = 2020)

results_He_Hr3 <- mult_bf_informed(x = x, Hr = Hr3, a = a,
                                     factor_levels = factor_levels,
                                     bf_type = 'LogBFer', seed = 2020)

logBFer1 <- summary(results_He_Hr1)$bf

logBFer3 <- summary(results_He_Hr3)$bf

```

581 We also compute the posterior model probabilities for all hypotheses. The results are  
 582 shown in Table 5.

Table 5

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

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_.   \mathbf{x})$	$\log(\text{BF}_{.e})$
$\mathcal{H}_0$	0.2	$1.27 \times 10^{-11}$	-17.67
$\mathcal{H}_{r1}$	0.2	0.9994	7.42
$\mathcal{H}_e$	0.2	0.0006	0
$\mathcal{H}_{r3}$	0.2	$5.97 \times 10^{-79}$	-172.70
$\mathcal{H}_{r2}$	0.2	$2.71 \times 10^{-212}$	-479.73

583 The results indicate strong support for  $\mathcal{H}_{r1}$  –the model in which the proportions are  
 584 assumed to decrease monotonically– over all other models. The log Bayes factor of  $\mathcal{H}_{r1}$   
 585 against the encompassing hypothesis  $\mathcal{H}_e$  is 7.42, which equates to a Bayes factor of 1,664  
 586 on a natural scale.

587 The strong Bayes factor support for  $\mathcal{H}_{r1}$  translates to a relatively extreme posterior  
 588 model probability of 0.9994. By comparison, the posterior model probabilities for  
 589 hypotheses  $\mathcal{H}_{r2}$  and  $\mathcal{H}_{r3}$ , that is, the bookend null-hypothesis and the hypothesis  
 590 predicting a data pattern typical of fraud, are only slightly greater than zero. The

591 posterior model probability for  $\mathcal{H}_e$  is 0.0006. Thus, hypothesis  $\mathcal{H}_{r1}$  can outperform the two  
 592 bookend hypotheses  $\mathcal{H}_{r2}$  and  $\mathcal{H}_e$ . That  $\mathcal{H}_{r1}$  outperforms the unconstrained model  $\mathcal{H}_e$   
 593 demonstrates how a parsimonious model that makes precise predictions can be favored over  
 594 a model that is more complex (e.g., Jefferys & Berger, 1992).

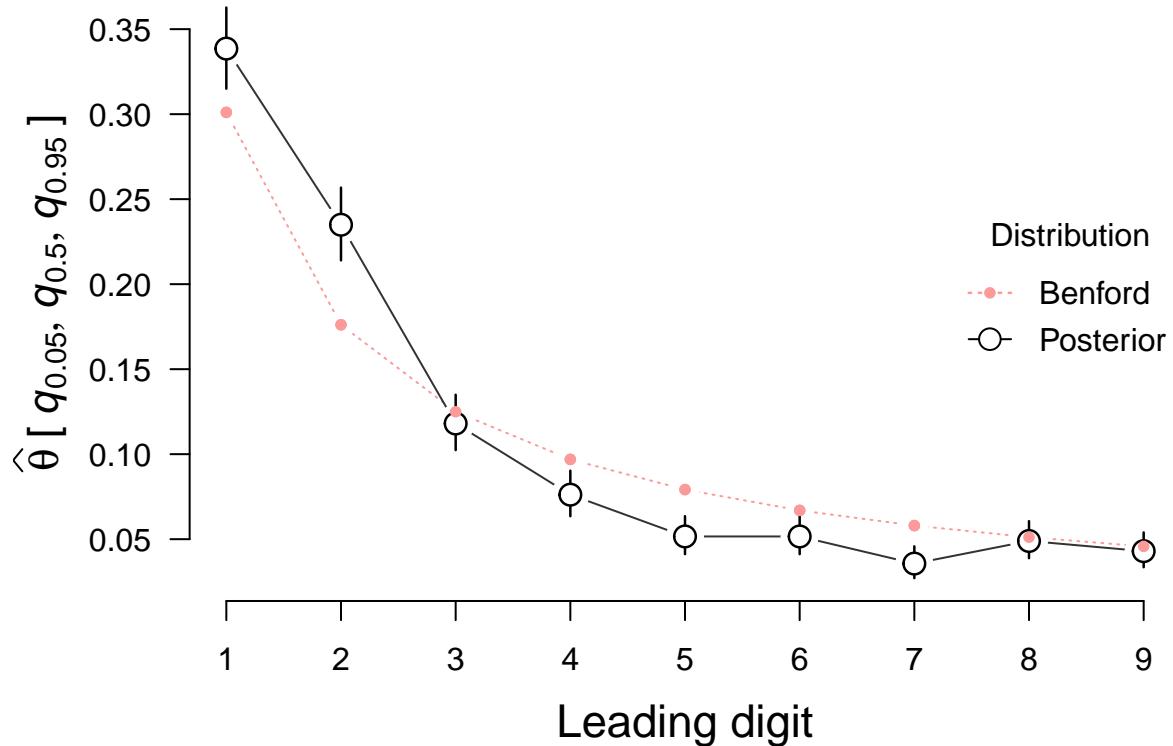


Figure 7. Predictions from Benford’s law (in pink) show together with the posterior medians (black circles) for the category proportions estimated under the encompassing model  $\mathcal{H}_e$ . The circle skewers show the 95% credible intervals. Only three of nine intervals encompass the expected proportions, suggesting that the data do not follow Benford’s law. This plot was created using the `plot-S3-method` for `summary.bmult` objects in `multibridge`.

595 **Sensitivity Analysis.** In a sensitivity analysis we will determine whether our  
 596 results are robust against different prior choices. In the main analysis we chose a uniform  
 597 Dirichlet distribution on the category proportions as prior under  $\mathcal{H}_e$ . This prior assigns

598 equal probability to all possible parameter values, but alternative prior distributions are  
 599 seem also conceivable. Audit researchers may argue for the development of more  
 600 informative and theory-driven priors that resemble one of the hypotheses under  
 601 consideration. The Dirichlet parameters vectors specified below resemble the four  
 602 hypotheses, assuming  $N = 54$  prior observations.

```
# Alternative prior specifications

a0 <- c(16, 10, 7, 5, 4, 3, 3, 3, 2) # Benford's law
a1 <- c(10, 9, 8, 7, 6, 5, 4, 3, 2) # Monotonically decreasing trend
a2 <- c(6, 6, 6, 6, 6, 6, 6, 6, 6) # Equal proportions
a3 <- c(12, 6, 6, 6, 6, 6, 6, 3, 3) # Fraud pattern
```

603 The sensitivity analysis is then carried out for each prior choice and will be compared  
 604 to the main results. For this analysis, we are particularly interested in the Bayes factors of  
 605 the hypothesis postulating a decreasing trend  $\mathcal{H}_{r1}$  and Benford's law  $\mathcal{H}_0$  to the  
 606 encompassing hypothesis  $\mathcal{H}_e$ .

```
# Sensitivity analysis for log(BFe_r1)

sensitivity0 <- mult_bf_informed(x = x, Hr = Hr1, a = a0,
                                    factor_levels = factor_levels,
                                    bf_type = 'LogBfer', seed = 2020)

sensitivity1 <- mult_bf_informed(x = x, Hr = Hr1, a = a1,
                                    factor_levels = factor_levels,
                                    bf_type = 'LogBfer', seed = 2020)

sensitivity2 <- mult_bf_informed(x = x, Hr = Hr1, a = a2,
                                    factor_levels = factor_levels,
                                    bf_type = 'LogBfer', seed = 2020)

sensitivity3 <- mult_bf_informed(x = x, Hr = Hr1, a = a3,
```

```

        factor_levels = factor_levels,
        bf_type = 'LogBFer', seed = 2020)

# Sensitivity analysis for log(BFe_0)
sensitivity4 <- mult_bf_equality(x = x, a = a0, p = p0)
sensitivity5 <- mult_bf_equality(x = x, a = a1, p = p0)
sensitivity6 <- mult_bf_equality(x = x, a = a2, p = p0)
sensitivity7 <- mult_bf_equality(x = x, a = a3, p = p0)

```

607       The results of the sensitivity analysis are displayed in Table 6. The general direction  
 608       of the sensitivity analysis agrees with our conclusions drawn from the main analysis. That  
 609       is, for the Bayes factors of  $\mathcal{H}_{r1}$  compared to  $\mathcal{H}_e$ , the evidence points towards the informed  
 610       hypothesis. However, the prior exerts an influence on  $\text{BF}_{r1e}$ ; the evidence in favor for the  
 611       informed hypothesis ranges from weak to extreme evidence. Specifically, when we choose  
 612       priors that resemble a decreasing trend for the frequency of leading digits, as we did with  
 613        $\boldsymbol{\alpha}_0$  and  $\boldsymbol{\alpha}_1$ , the Bayes factor becomes smaller and the evidence weak (i.e.,  $(\text{BF}_{r1e} | \boldsymbol{\alpha}_0) =$   
 614       1.87 on the natural scale) and moderate (i.e.,  $(\text{BF}_{r1e} | \boldsymbol{\alpha}_1) = 4.74$  on the natural scale).  
 615       When the prior contrasts the data, the evidence becomes very strong or extreme. Thus, a  
 616       prior that closely resembles the predictive trend reduces to some degree the diagnostic  
 617       value of the data.

618       By contrast, the Bayes factors for  $\mathcal{H}_0$  compared to  $\mathcal{H}_e$  are robust against different  
 619       prior settings. Here too, the prior changes the Bayes factor estimate but in all cases the  
 620       data suggests overwhelming evidence in favor of the encompassing hypothesis over  
 621       Benford's law.

Table 6

*Results of a sensitivity analysis for the Greek fiscal data set.*

Description	Prior	$\log(\text{BF}_{r1e})$	$\log(\text{BF}_{0e})$
Uniform	$\alpha_e = (1, 1, 1, 1, 1, 1, 1, 1, 1)$	7.42	-17.67
Benford's law	$\alpha_0 = (16, 10, 7, 5, 4, 3, 3, 3, 2)$	0.63	-26.00
Monotonically decreasing	$\alpha_1 = (10, 9, 8, 7, 6, 5, 4, 3, 2)$	1.56	-20.94
Centered on mean	$\alpha_2 = (6, 6, 6, 6, 6, 6, 6, 6, 6)$	7.53	-11.35
Fraud pattern	$\alpha_3 = (12, 6, 6, 6, 6, 6, 6, 3, 3)$	3.93	-18.62

To summarize, the data offer overwhelming support for hypothesis  $\mathcal{H}_{r1}$ , which postulates a decreasing trend in the digit proportions. This model outperformed both simpler models (e.g., the Benford model and the bookend null-hypothesis) and a more complex model in which the proportions were free to vary. The results are sensitive to our prior choices as a sensitivity analysis showed: for moderately informative priors which resemble the predicted decreasing trend, the  $\mathcal{H}_{r1}$  cannot outperform the encompassing model. On the other hand, the conclusion that Benford's law does not offer a good description of the data was robust to different prior settings. Detailed follow-up analyses are needed to discover why the Greek fiscal data fail to adhere to Benford's law (Nigrini, 2019).

### Example 2: Prevalence of Statistical Reporting Errors

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

639 package **statcheck**, which scans the PDF of a given scientific article and automatically  
 640 detects statistical inconsistencies. This package allowed Nuijten et al. (2016) to estimate  
 641 the prevalence of statistical reporting errors in the field of psychology. In total, the authors  
 642 investigated a sample of 30,717 articles (which translates to over a quarter of a million  
 643 *p*-values) published in eight major psychology journals between 1985 to 2013:  
 644 *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).

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

652         **Data and Hypothesis.** Here we reuse the original data published by Nuijten et al.  
 653 (2016), which we also distribute with the package **multibridge** under the name **journals**.

```
data(journals)
```

654         The Nuijten et al. (2016) hypothesis of interest,  $\mathcal{H}_r$ , states that the prevalence for  
 655 statistical reporting errors is higher for JPSP than for the other journals.<sup>2</sup> We will consider  
 656 two specific versions of the Nuijten et al. (2016)  $\mathcal{H}_r$  hypothesis. The first hypothesis,  $\mathcal{H}_{r1}$ ,  
 657 stipulates that JPSP has the highest prevalence of reporting inconsistencies, whereas the  
 658 other seven journals share a prevalence that is lower. The second hypothesis,  $\mathcal{H}_{r2}$ , also  
 659 stipulates that JPSP has the highest prevalence of reporting inconsistencies, but does not  
 660 commit to any particular structure on the prevalence for the other seven journals.

---

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

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

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

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

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

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

668        **Method.** To compute the Bayes factor  $\text{BF}_{0r}$  we need to specify (1) a vector with  
 669 observed successes (i.e., the number of articles that contain a statistical inconsistency), (2)  
 670 a vector containing the total number of observations (i.e., the number of articles), (3) a  
 671 vector with prior parameter  $\alpha_k$  for each binomial proportion of the beta prior distribution  
 672 under  $\mathcal{H}_e$ , (4) a vector with prior parameter  $\beta_k$  for each binomial proportion of the beta  
 673 prior distribution under  $\mathcal{H}_e$ , (5) the category labels (i.e., journal names), and (6) the  
 674 informed hypothesis  $\mathcal{H}_{r1}$  or  $\mathcal{H}_{r2}$  (e.g., as a string). We also change the Bayes factor type to  
 675 `LogBFr0` so that the function returns the log Bayes factor in favor for the informed  
 676 hypothesis compared to the null hypothesis. Since we have no specific expectations about  
 677 the distribution of statistical reporting errors in any given journal, we set all parameters  $\alpha_k$   
 678 and  $\beta_k$  to one which corresponds to uniform beta distributions. With this information, we  
 679 can now conduct the analysis with the function `binom_bf_informed`.

```

# Since percentages are rounded to two decimal values, we round the
# articles with an error to obtain integer values

x <- round(journals$articles_with_NHST *
            (journals$perc_articles_with_errors/100))

# Total number of articles

n <- journals$articles_with_NHST

# Prior specification for beta prior distributions under H_e

a <- c(1, 1, 1, 1, 1, 1, 1, 1)

b <- c(1, 1, 1, 1, 1, 1, 1, 1)

# Labels for categories of interest

journal_names <- journals$journal

# Specifying the informed Hypothesis

Hr1 <- c('JAP = PS = JCCP = PLOS = DP = FP = JEPG < JPSP')

Hr2 <- c('JAP , PS , JCCP , PLOS , DP , FP , JEPG < JPSP')

# Execute the analysis for Hr1

results_H0_Hr1 <- binom_bf_informed(x = x, n = n, Hr = Hr1, a = a, b = b,
                                      factor_levels = journal_names,
                                      bf_type = 'LogBFr0', seed = 2020)

# Execute the analysis for Hr2

results_H0_Hr2 <- binom_bf_informed(x = x, n = n, Hr = Hr2, a = a, b = b,
                                      factor_levels = journal_names,
                                      bf_type = 'LogBFr0', seed = 2020)

```

```
LogBFe0 <- results_H0_Hr1$bf_list$bf0_table[['LogBFe0']]
```

```
LogBFr10 <- summary(results_H0_Hr1)$bf
```

```
LogBFr20 <- summary(results_H0_Hr2)$bf
```

Table 7

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

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

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

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

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

687 The posterior medians and 95% credible intervals are returned by the  
 688 `summary`-method and are shown in Figure 8. The figure strongly suggests that the  
 689 prevalence of reporting inconsistencies is not equal across all eight journals. This  
 690 impression may be quantified by comparing the null hypothesis  $\mathcal{H}_0$  to the encompassing  
 691 hypothesis  $\mathcal{H}_e$ . The corresponding Bayes factor equals  $\log(\text{BF}_{e0}) = 156.27$ , which confirms

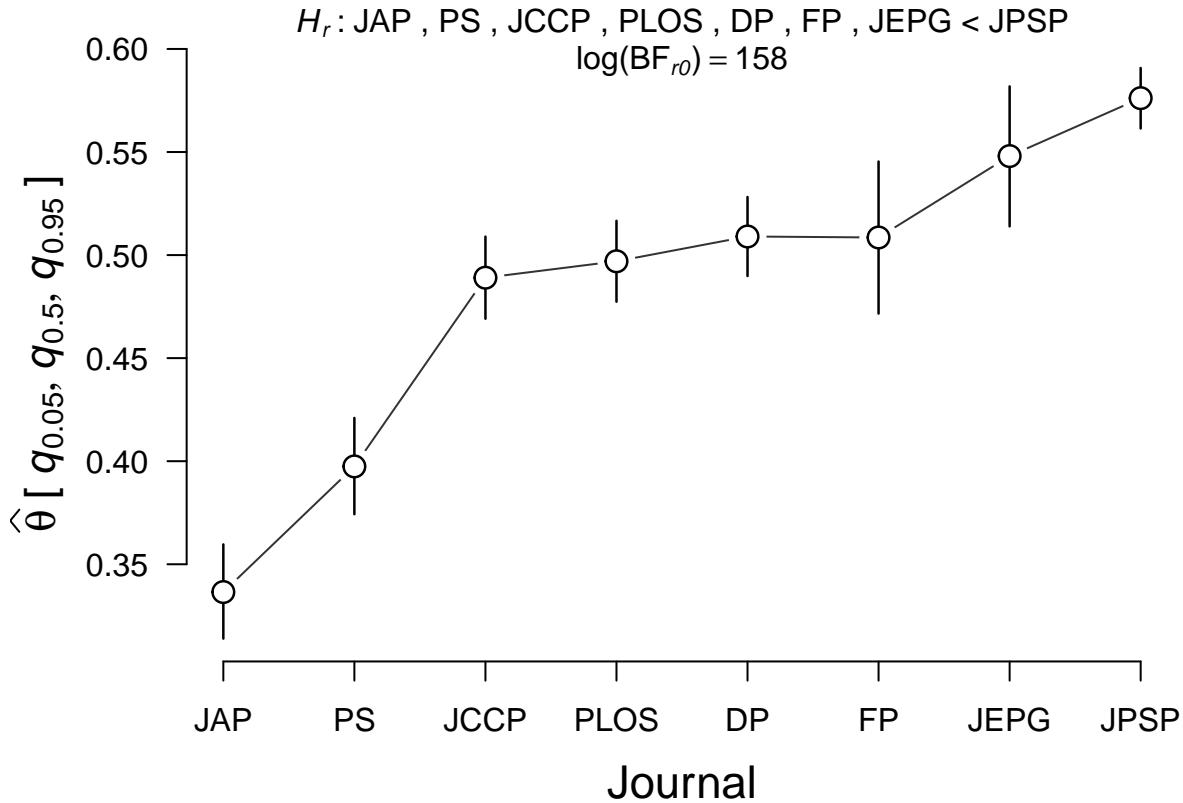


Figure 8. Posterior medians for the prevalence of statistical reporting inconsistencies across eight psychology journals, as obtained using the encompassing model. The circle skewers show the 95% credible intervals. Analysis based on data from Nuijten et al. (2016). This plot was created using the `plot-S3-method` for `summary.bmult` objects.

692 that the data dramatically undercut the null hypothesis that the prevalence of statistical  
 693 reporting inconsistencies is equal across journals.

694 The data offer most support for the Nuijten hypothesis  $\mathcal{H}_{r2}$ , which posits that JPSP  
 695 has the highest prevalence but does not commit to any restriction on the prevalences for  
 696 the remaining seven journals. This hypothesis may be compared to the encompassing  
 697 hypothesis  $\mathcal{H}_e$ , which yields  $\log(\text{BF}_{r2e}) = 2.01$ . This means that the observed data are  
 698  $\exp(2.01) \approx 7.45$  times more likely under  $\mathcal{H}_{r2}$  than under  $\mathcal{H}_e$ ; this is moderate evidence for  
 699 the restriction suggested by Nuijten et al. (2016). Under equal prior probability for the

700 models, this Bayes factor translates to a posterior probability on  $\mathcal{H}_e$  of 0.119, an amount  
701 that researchers may deem too large to discard in an all-or-none fashion.

702 To summarize, the data provide moderate evidence for the hypothesis stated by  
703 Nuijten et al. (2016) that the prevalence of statistical reporting inconsistencies in JPSP is  
704 higher than that in seven other psychology journals.

705 **Example 3: Effects of Gender and Education on the Violation of Stochastic  
706 Dominance**

707 This section illustrates the comparison of four nested hypotheses concerning  
708 independent binomial probabilities. In his study, Birnbaum (1999) presented new  
709 possibilities of online testing for psychological science (in the late 1990s online testing was  
710 still novel and rarely used). To compare data collected from an online research to  
711 traditional lab research, Birnbaum (1999) collected experimental data from 1224  
712 participants online and 124 participants in the lab. In his experiment participants played  
713 20 rounds of a gambling game. In each round, they were presented with two gambles with  
714 different probabilities and monetary values and were asked to indicate which gamble they  
715 would rather play. The gamble chosen by the participants was then played once. Birnbaum  
716 (1999) examined the characteristics of the two samples, for instance, in terms of their risk  
717 aversion and their consistency with decision making axioms, such as stochastic violations,  
718 and correlated them with different demographics.

719 The author analyzed the proportion of stochastic violations for different demographic  
720 variables, noting a seemingly ordinal pattern for the probabilities to violate of stochastic  
721 dominance for the factors gender (m=male, f=female) and education (1=doctorate degree,  
722 2=postgraduate degree, 3=bachelor's degree, 4=less than bachelor's degree). In a later  
723 study, Myung et al. (2005) presented a Bayesian inference framework to test decision  
724 making axioms (using the “Bayesian  $p$ -value”) and used Birnbaum’s data as an example

725 on how to assess violations of stochastic dominance and their relationship with covariates.  
 726 Concretely, Myung et al. (2005) reanalyzed the data from Birnbaum (1999) and tested the  
 727 informed hypothesis that stochastic dominance is violated more frequently in women  
 728 compared to men and more frequently in lower education levels than higher education  
 729 levels.

730 **Data and Hypothesis.** We will use data from Birnbaum (1999) as presented in  
 731 Myung et al. (2005). The data show the stochastic violations of the online sample for one  
 732 of the gambling rounds featuring 1212 valid responses (see Table 8).

```
dat <- data.frame(gender = rep(c('male', 'female'), each = 4),  

                   education = rep(c('1', '2', '3', '4'), 2),  

                   levels = paste0(rep(c('m', 'f'), each = 4), 1:4),  

                   violation = c(0.487, 0.477, 0.523, 0.601,  

                               0.407, 0.555, 0.650, 0.622),  

                   n = c(80, 88, 195, 163,  

                         54, 108, 206, 318),  

                   x = c(39, 42, 102, 98,  

                         22, 60, 134, 198))
```

733 The parameter vector of the binomial success probabilities,  $\theta_1, \dots, \theta_K$ , contains the  
 734 probabilities of observing a value in a particular category; here, it reflects the probabilities  
 735 of violating stochastic dominance for a particular subgroup (e.g., women with a doctorate).  
 736 We will compare three inequality-constrained hypotheses  $\mathcal{H}_{r1}, \mathcal{H}_{r2}, \mathcal{H}_{r3}$  formulated by  
 737 Myung et al. (2005). The first hypothesis  $\mathcal{H}_{r1}$  encodes the main effect for gender and  
 738 states that the probability to violate stochastic dominance is lower for men than for  
 739 women. The second hypothesis  $\mathcal{H}_{r2}$  encodes the main effect of education and states that  
 740 the probability to violate stochastic dominance is lower for persons with higher education  
 741 levels. The third hypothesis  $\mathcal{H}_{r3}$  combines hypotheses  $\mathcal{H}_{r1}$  and  $\mathcal{H}_{r2}$ . We will test this

Table 8

*Observed counts and observed proportions of stochastic dominance violations for the N = 1,212 participants in Birnbaum (1999). The data are split by gender and education level of the participants.*

Education	Observed Counts	Observed Proportions
<b>Male</b>		
Doctorate Degree	39/80	0.49
Postgraduate Degree	42/88	0.48
Bachelor's Degree	102/195	0.52
Less than Bachelor's degree	98/163	0.60
<b>Female</b>		
Doctorate Degree	22/54	0.41
Postgraduate Degree	60/108	0.56
Bachelor's Degree	134/206	0.65
Less than Bachelor's degree	198/318	0.62

<sup>742</sup> hypothesis against the encompassing hypothesis  $\mathcal{H}_e$  without any constraints. In addition,

<sup>743</sup> we will include a bookend null-hypothesis  $\mathcal{H}_0$  predicting that all probabilities are equal.

<sup>744</sup> The set of candidate hypotheses can therefore be written as follows:

$$\mathcal{H}_e : (\theta_{m1}, \theta_{m2}, \theta_{m3}, \theta_{m4}, \theta_{f1}, \theta_{f2}, \theta_{f3}, \theta_{f4})$$

$$\mathcal{H}_0 : \boldsymbol{\theta}_0 = \left( \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8} \right),$$

$$\mathcal{H}_{r1} : (\theta_{m1}, \theta_{m2}, \theta_{m3}, \theta_{m4}) < (\theta_{f1}, \theta_{f2}, \theta_{f3}, \theta_{f4})$$

$$\mathcal{H}_{r2} : (\theta_{m1}, \theta_{f1}) < (\theta_{m2}, \theta_{f2}) < (\theta_{m3}, \theta_{f3}) < (\theta_{m4}, \theta_{f4})$$

$$\mathcal{H}_{r3} : \theta_{m1} < \theta_{f1} < \theta_{m2} < \theta_{f2} < \theta_{m3} < \theta_{f3} < \theta_{m4} < \theta_{f4}.$$

745       **Method.** To evaluate the inequality-constrained hypothesis, we need to specify (1)

746 a vector with observed successes, and (2) a vector containing the total number of

747 observations, (3) the informed hypothesis, (4) a vector with prior parameters alpha for each

748 binomial proportion, (5) a vector with prior parameters beta for each binomial proportion,

749 and (6) the labels of the categories of interest (i.e., gender and education level). As with

750 the previous two example, we assign a uniform Beta prior to the binomial probabilities.

```
# number of violations
x <- dat$x

# total number people in the category
n <- dat$n

# Specifying the informed hypotheses (step 3)

# null hypothesis
p0 <- c(1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)

# informed hypotheses
Hr1 <- c('m1, m2, m3, m4 < f1, f2, f3, f4')
Hr2 <- c('m1, f1 < m2, f2 < m3, f3 < m4, f4')
Hr3 <- c('m1 < f1 < m2 < f2 < m3 < f3 < m4 < f4')

# Prior specification (step 4 and 5)

# We assign a uniform beta distribution to each binomial proportion
a <- c(1, 1, 1, 1, 1, 1, 1, 1)
b <- c(1, 1, 1, 1, 1, 1, 1, 1)

# categories of interest (step 6)
```

```
gender_edu <- dat$levels
```

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

751      With this information, we can now conduct the analysis with the function  
 752      `binom_bf_informed()`. Since we are interested in quantifying evidence in favor of the  
 753      informed hypotheses compared to the encompassing hypothesis, we set the Bayes factor  
 754      type to `BFre`. For reproducibility, we are also setting a seed.

```
results_H0_He <- multibrige::mult_bf_equality(x = x, a = a, p = p0)

results_Hr1_He <- multibrige::binom_bf_informed(x=x, n=n, Hr=Hr1, a=a, b=b,
                                                 factor_levels=gender_edu,
                                                 bf_type = 'BFre',
                                                 seed = 2020)

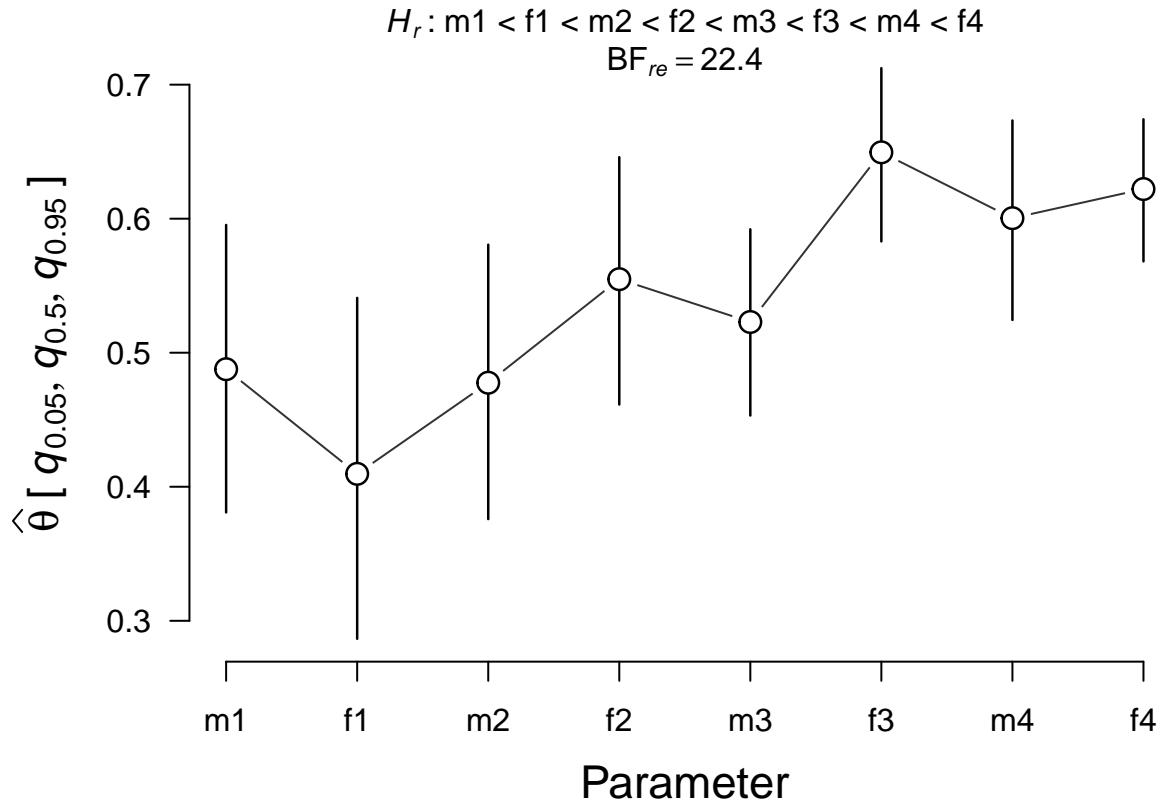
results_Hr2_He <- multibrige::binom_bf_informed(x=x, n=n, Hr=Hr2, a=a, b=b,
                                                 factor_levels=gender_edu,
                                                 bf_type = 'BFre',
                                                 seed = 2020)

results_Hr3_He <- multibrige::binom_bf_informed(x=x, n=n, Hr=Hr3, a=a, b=b,
                                                 factor_levels=gender_edu,
                                                 bf_type = 'BFre',
                                                 seed = 2020)
```

755      The results are summarized in Table 9. We first inspect the Bayes factors for the  
 756      three informed hypotheses compared to the encompassing hypothesis. For hypotheses  $\mathcal{H}_{r1}$ ,

757 the data suggest moderate evidence for the encompassing hypothesis compared to the  
 758 informed hypothesis, with a Bayes factor of 6.43. This hypothesis predicted a main effect  
 759 of gender, that is, men should have a lower probability of violating stochastic dominance  
 760 than women regardless of their education level. For hypotheses  $\mathcal{H}_{r2}$  and  $\mathcal{H}_{r3}$ , the data  
 761 provide strong evidence in favor of the informed hypothesis compared to the encompassing  
 762 hypothesis, with Bayes factors of 17.82 and 22.36, respectively. However, there is no  
 763 predictive advantage of  $\mathcal{H}_{r3}$  over  $\mathcal{H}_{r2}$ ; the Bayes factor directly comparing these hypotheses  
 764 is  $\text{BF}_{r3r2} = \frac{\text{BF}_{r3e}}{\text{BF}_{r2e}} = 1.26$ . The degree to which the data conforms to the predicted pattern  
 765 from  $\mathcal{H}_{r3}$  becomes apparent when we plot the posterior estimates.

```
plot(summary(results_Hr3_He))
```



766

767 To compare all four hypotheses directly with each other, we computed the posterior  
 768 model probabilities:

```

post_probs <- data.frame(
  Hyps = c('p(He | x)', 'p(H0 | x)', 'p(Hr1 | x)', 'p(Hr2 | x)', 'p(Hr3 | x)'),
  Prob = c(1, BF0e, BFr1e, BFr2e, BFr3e)/sum(c(1, BF0e, BFr1e, BFr2e, BFr3e)))

```

769 The model that predicts only a gender effect performs worse than the baseline model  
 770 without any restrictions. Hypothesis  $\mathcal{H}_{r3}$  outperforms all other models, including the  
 771 bookend hypotheses, with a posterior model probability of 54%. Here too, however, the  
 772 posterior probability of hypothesis  $\mathcal{H}_{r2}$  is with 43% almost as high as  $\mathcal{H}_{r3}$ . To sum up,  
 773 even though  $\mathcal{H}_{r3}$  yield the biggest Bayes factor and the highest posterior model probability,  
 774 the difference advantage to  $\mathcal{H}_{r2}$  is slim. Note that Myung et al. (2005) and Birnbaum  
 775 (1999) concluded that hypothesis  $\mathcal{H}_{r3}$  performs the best. In contrast, our analysis  
 776 suggested that here is strong evidence for an effect of education, but it is inconclusive  
 777 whether the effect is moderated by gender.

Table 9

*Prior model probabilities, posterior model probabilities, and Bayes factors for four hypotheses concerning the relationship between gender and education level on the probability of violating stochastic dominance.*

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_.   \mathbf{x})$	$BF_{.e}$
$\mathcal{H}_e$	0.25	0.0242	1
$\mathcal{H}_0$	0.25	$1.34 \times 10^{-53}$	$5.55 \times 10^{-52}$
$\mathcal{H}_{r1}$	0.25	0.0038	0.16
$\mathcal{H}_{r2}$	0.25	0.4310	17.82
$\mathcal{H}_{r3}$	0.25	0.5410	22.36

## Discussion

778 The R package **multibridge** facilitates the estimation of Bayes factors for informed  
 779 hypotheses in both multinomial and independent binomial models. The efficiency gains of  
 780

781 **multibridge** are particularly pronounced when the parameter restrictions are highly  
782 informative or when the number of categories is large.

783       **multibridge** supports the evaluation of informed hypotheses that feature equality  
784 constraints, inequality constraints, and free parameters, as well as combinations between  
785 them. Moreover, users can choose to test the informative hypothesis against an  
786 encompassing hypothesis that lets all parameters vary freely or against the null hypothesis  
787 that states that category proportions are exactly equal. Beyond the core functions  
788 currently implemented in **multibridge**, there are several natural extensions we aim to  
789 include in future versions of this package. For instance, to compare several models with  
790 each other we plan to implement functions that compute the posterior model probabilities.  
791 Another extension is to facilitate the specification of hierarchical binomial and multinomial  
792 models which would allow users to analyze data where responses are nested within a  
793 higher-order structure such as participants, schools, or countries. Hierarchical multinomial  
794 models can be found, for instance, in source memory research where people need to select a  
795 previously studied item from a list (e.g., Arnold, Heck, Bröder, Meiser, & Boywitt, 2019); a  
796 hierarchical binomial model was applied, for instance, in Hoogeveen, Sarafoglou, and  
797 Wagenmakers (2020), to evaluate laypeople's accuracy in predicting replication outcomes  
798 for social science studies.

799       Furthermore, to make the method accessible to a larger audience of users and  
800 students, the informed Bayesian multinomial test and the informed Bayesian test for  
801 multiple binomials will be made available in future versions of the software package JASP  
802 (JASP Team, 2022). JASP offers an intuitive graphical user interface and does not require  
803 extensive knowledge in programming.

804       In addition, we plan to expand the types of hypotheses that can be evaluated in  
805 future versions of this package. Currently, **multibridge** only supports informed  
806 hypotheses which are "stick hypotheses", that is, hypotheses in which all elements within a

constraint are linearly ordered. While the quantity shown in Equation 1 admits in principle any constraint imposed on a vector of category proportions, this requirement is necessary for the bridge sampling routine, in order to transform samples from the real line to the probability space. To be able to evaluate more general ordinal constraints including “branch-hypotheses” with bridge sampling in the future, the stick-breaking transformation needs to be further refined. Arguably, this refinement can be realized more easily for transformations of multiple binomials than for multinomials, since independent binomials live in probability space but are not constrained by the sum-to-one condition.

Finally, we aim to enable the specification of more general informed hypotheses, including hypotheses on the size ratios of the parameters (e.g.,  $\theta_1 < 2 \times \theta_2$ ) or on their odds ratios (e.g.,  $\frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$ ). A framework to evaluate these constraints using the unconditional encompassing approach has already been proposed (Klugkist, Laudy, & Hoijtink, 2010). We believe that the bridge sampling method could also be extended to test these hypotheses as in principle, as all the building blocks are already in place. Specifically, **multibridge** takes size ratios into account when it evaluates hypotheses featuring combinations of equality and inequality constraints. For these hypotheses, **multibridge** first evaluates the equality constraints separately and then evaluates the inequality constraints given the equality constraints hold. To do so, the algorithm merges equality-constrained categories but tracks their initial number to effectively sample from the constrained parameter space and to transform the parameters. For odds ratios, on the other hand, a suitable sampling method and transformation has not yet been developed. To facilitate the evaluation of these hypotheses, alternative methods to sample and transform the parameters are required.

830

## Declarations

831 **Availability of data and code**

832 The source code of the R package is available at:

833 <https://github.com/ASarafoglou/multibridge/>. In addition, readers can access the code for  
834 reproducing all analyses and plots via our project folder on the Open Science Framework:  
835 <https://osf.io/2wf5y/>.836 **Funding**837 This research was supported by a Netherlands Organisation for Scientific Research  
838 (NWO) grant to AS (406-17-568), a Veni grant from the NWO to MM (451-17-017), a Vici  
839 grant from the NWO to EJW (016.Vici.170.083), as well as a European Research Council  
840 (ERC) grant to EJW (283876). This paper was written in Rmarkdown, using the R  
841 package **papaja** (Aust & Barth, 2020).842 **Author contributions**

843 The authors made the following contributions. Alexandra Sarafoglou:

844 Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology,  
845 Project Administration, Software, Validation, Visualization, Writing - Original Draft  
846 Preparation, Writing - Review & Editing; Frederik Aust: Conceptualization, Software,  
847 Supervision, Validation, Visualization, Writing - Original Draft Preparation, Writing -  
848 Review & Editing; Maarten Marsman: Funding Acquisition, Conceptualization,  
849 Methodology, Supervision, Validation, Writing - Review & Editing; Frantisek Bartos:  
850 Software; Eric-Jan Wagenmakers: Funding Acquisition, Methodology, Supervision,  
851 Validation, Writing - Review & Editing; Julia M. Haaf: Conceptualization, Formal  
852 Analysis, Methodology, Software, Supervision, Validation, Writing - Original Draft  
853 Preparation, Writing - Review & Editing.

**854 Conflicts of interest**

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

**857 Ethical Approval**

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

**References**

- Arnold, N. R., Heck, D. W., Bröder, A., Meiser, T., & Boywitt, C. D. (2019). Testing hypotheses about binding in context memory with a hierarchical multinomial modeling approach. *Experimental Psychology*, 66, 239–251.
- Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78, 551–572.
- Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22, 245–268.
- Berger, J. O., & Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59, 3–15.
- Birnbaum, M. H. (1999). Testing critical properties of decision making on the internet. *Psychological Science*, 10, 399–407.
- Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13, 627–679.
- Damien, P., & Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10, 206–215.
- Doorn, J. van Bergh, D. van den Böhm, U., Dablander, F., Derks, K., Draws, T., ... Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3), 813–826.
- 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>

- 886 European Commision. (2004). *Report by Eurostat on the revision of the Greek*  
887 *government deficit and debt figures* [Eurostat Report].  
888 <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/GREECE>.
- 889 European Commision. (2010). *Report on Greek government deficit and debt*  
890 *statistics* [Eurostat Report]. [https://ec.europa.eu/eurostat/web/products-eurostat-news/-/COM\\_2010\\_REPORT\\_GREEK](https://ec.europa.eu/eurostat/web/products-eurostat-news/-/COM_2010_REPORT_GREEK).
- 891 Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet*  
892 *distribution and related processes*. Department of Electrical Engineering,  
893 University of Washington.
- 894 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019).  
895 Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 389–402.
- 896 Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., & Hothorn, F. S. T. (2020).  
897 *Mvtnorm: Multivariate normal and t distributions*. Retrieved from  
898 <http://CRAN.R-project.org/package=mvtnorm>
- 899 Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ...  
900 Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical*  
901 *Psychology*, 81, 80–97.
- 902 Gronau, Q. F., Singmann, H., & Wagenmakers, E. –J. (2020). Bridgesampling: An  
903 R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, 92, 1–29.
- 904 Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for  
905 Bayesian testing of order constrained hypotheses in structural equation models.  
906 *Journal of Statistical Computation and Simulation*, 89, 1526–1553.
- 907 Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of  
908 inequality constrained hypotheses. *Psychological Methods*, 19, 511–527.
- 909 Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes

- 913 factors: A general method for testing informative hypotheses. *British Journal of*  
914 *Mathematical and Statistical Psychology*, 71, 229–261.
- 915 Haaf, J. M., Klaassen, F., & Rouder, J. (2019). Capturing ordinal theoretical  
916 constraint in psychological science. *PsyArXiv*. Retrieved from  
917 <https://doi.org/10.31234/osf.io/a4xu9>
- 918 Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear  
919 inequality constraints: Overview and improvements of computational methods  
920 for Bayesian inference. *Journal of Mathematical Psychology*, 91, 70–87.
- 921 Heck, D. W., & Wagenmakers, E.-J. (2016). Adjusted priors for Bayes factors  
922 involving reparameterized order constraints. *Journal of Mathematical*  
923 *Psychology*, 73, 110–116.
- 924 Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical*  
925 *Science*, 10, 354–363.
- 926 Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and*  
927 *social scientists*. Boca Raton, FL: Chapman & Hall/CRC.
- 928 Hoijtink, H., Klugkist, I., & Boelen, P. (Eds.). (2008). *Bayesian evaluation of*  
929 *informative hypotheses*. New York: Springer Verlag.
- 930 Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople can predict  
931 which social-science studies will be replicated successfully. *Advances in Methods*  
932 *and Practices in Psychological Science*, 3, 267–285.
- 933 JASP Team. (2022). *JASP (Version 0.16.3.0) [Computer software]*.  
934 <https://jasp-stats.org/>.
- 935 Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis.  
936 *American Scientist*, 80, 64–72.
- 937 Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability.  
938 *Proceedings of the Cambridge Philosophy Society*, 31, 203–222.
- 939 Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American*

- 940                   *Statistical Association*, 90, 773–795.
- 941                   Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A  
942                   latent-trait approach. *Psychometrika*, 75, 70–98.
- 943                   Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using  
944                   encompassing priors. *Statistica Neerlandica*, 59, 57–69.
- 945                   Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and  
946                   equality constrained hypotheses for contingency tables. *Psychological Methods*,  
947                   15, 281–299.
- 948                   Laudy, O. (2006). *Bayesian inequality constrained models for categorical data* (PhD  
949                   thesis). Utrecht University.
- 950                   Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive  
951                   models. *Psychonomic Bulletin & Review*, 25, 114–127.
- 952                   Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015).  
953                   Bayesian estimation of multinomial processing tree models with heterogeneity in  
954                   participants and items. *Psychometrika*, 80, 205–235.
- 955                   Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via  
956                   a simple identity: A theoretical exploration. *Statistica Sinica*, 6, 831–860.
- 957                   Mulder, Joris. (2014). Prior adjusted default Bayes factors for testing (in) equality  
958                   constrained hypotheses. *Computational Statistics & Data Analysis*, 71, 448–463.
- 959                   Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on  
960                   correlations. *Journal of Mathematical Psychology*, 72, 104–115.
- 961                   Mulder, Joris, Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F.,  
962                   Hoijtink, H., ... van Lissa, C. (2021). BFpack: Flexible Bayes factor testing of  
963                   scientific theories in R. *Journal of Statistical Software*, 2–63, 239–251.
- 964                   Mulder, Joris, Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program  
965                   for calculating Bayes factors for inequality and equality constrained models.  
966                   *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.
- Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, 49, 205–225.
- 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. J. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.
- Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud schemes. *Managerial Auditing Journal*, 34, 602–622.
- Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford's law as an aid in analytical procedures. *Auditing*, 16, 52–67.
- Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48, 1205–1226.
- Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54, 3269–3288.
- Rauch, B., Götsche, 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.

- 994 Rijkeboer, M., & van den Hout, M. (2008). A psychologists's view on Bayesian  
995 evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. A. Boelen  
996 (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 299–309). Berlin:  
997 Springer Verlag.
- 998 Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E. –J., &  
999 Marsman, M. (2021). Evaluating multinomial order restrictions with bridge  
1000 sampling. *Psychological Methods*.
- 1001 Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian  
1002 workflow in cognitive science. *Psychological Methods*, 26(1), 103–126.
- 1003 Sedransk, J., Monahan, J., & Chiu, H. (1985). Bayesian estimation of finite  
1004 population parameters in categorical data models incorporating order  
1005 restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*,  
1006 47, 519–527.
- 1007 Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior  
1008 distributions. *The American Statistician*, 56, 196–201.
- 1009 Stan Development Team. (2020). *Stan modeling language user's guide and reference  
1010 manual, version 2.23.0*. R Foundation for Statistical Computing. Retrieved from  
1011 <http://mc-stan.org/>
- 1012 Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the  
1013 Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.
- 1014 Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of  
1015 a replication attempt. *Journal of Experimental Psychology: General*, 143,  
1016 1457–1475.
- 1017 Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahnik,  
1018 S., . . . Aczel, B. (2021). Seven steps toward more transparency in statistical  
1019 practice. *Nature Human Behaviour*, 5, 1473–1480.

1020

## Appendix

### 1021 Transforming an Ordered Probability Vector to the Real Line

1022 The bridge sampling routine in **multibridge** uses the multivariate normal  
 1023 distribution as proposal distribution, which requires moving the target distribution  $\boldsymbol{\theta}$  to  
 1024 the real line. Crucially, the transformation needs to retain the ordering of the parameters,  
 1025 that is, it needs to take into account the lower bound  $l_k$  and the upper bound  $u_k$  of each  $\theta_k$ .  
 1026 To meet these requirements, **multibridge** uses a probit transformation, as proposed in  
 1027 Sarafoglou et al. (2021), and subsequently transforms the elements in  $\boldsymbol{\theta}$ , moving from its  
 1028 lowest to its highest value. In the binomial model, we move all elements in  $\boldsymbol{\theta}$  to the real  
 1029 line and thus construct a new vector  $\mathbf{y} \in \mathbb{R}^K$ . For multinomial models it follows from the  
 1030 sum-to-one constraint that the vector  $\boldsymbol{\theta}$  is completely determined by its first  $K - 1$   
 1031 elements, where  $\theta_K$  is defined as  $1 - \sum_{k=1}^{K-1} \theta_k$ . Hence, for multinomial models we will only  
 1032 consider the first  $K - 1$  elements of  $\boldsymbol{\theta}$  and we will transform them to  $K - 1$  elements of a  
 1033 new vector  $\mathbf{y} \in \mathbb{R}^{K-1}$ .

1034 Let  $\phi$  denote the density of a normal variable with a mean of zero and a variance of  
 1035 one,  $\Phi$  denote its cumulative density function, and  $\Phi^{-1}$  denote the inverse cumulative  
 1036 density function. Then for each element  $\theta_k$ , the transformation is

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

1037 The inverse transformation is given by

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

1038 To perform the transformations, we need to determine the lower bound  $l_k$  and the  
 1039 upper bound  $u_k$  of each  $\theta_k$ . Assuming  $\theta_{k-1} < \theta_k$  for  $k \in \{2, \dots, K\}$  the lower bound for any  
 1040 element in  $\boldsymbol{\theta}$  is defined as

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

1041 This definition holds for both binomial models and multinomial models. Differences  
 1042 in these two models appear only when determining the upper bound for each parameter.  
 1043 For binomial models, the upper bound for each  $\theta_k$  is simply 1. For multinomial models,  
 1044 however, due to the sum-to-one constraint the upper bounds depend on the values of  
 1045 smaller elements as well as on the number of remaining larger elements in  $\boldsymbol{\theta}$ . To be able to  
 1046 determine the upper bounds, we represent  $\boldsymbol{\theta}$  as unit-length stick which we subsequently  
 1047 divide into  $K$  elements Stan Development Team (2020). By using this so-called  
 1048 stick-breaking method we can define the upper bound for any  $\theta_k$  as follows:

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

1049 where  $1 - \sum_{i < k} \theta_i$  represents the length of the remaining stick, that is, the proportion of  
 1050 the unit-length stick that has not yet been accounted for in the transformation. The  
 1051 elements in the remaining stick are denoted as  $ERS$ , and are computed as follows:

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

1052 The transformations outlined above are suitable only for ordered probability vectors,  
 1053 that is, for informed hypotheses in binomial and multinomial models that only feature  
 1054 inequality constraints. However, when informed hypotheses also feature equality  
 1055 constrained parameters, as well as parameters that are free to vary we need to modify the  
 1056 formula. Specifically, to determine the lower bounds for any  $\theta_k$ , we need to take into

1057 account how many parameters were set equal to it (denoted as  $e_k$ ) and how many  
 1058 parameters were set equal to its preceding value  $\theta_{k-1}$  (denoted as  $e_{k-1}$ ):

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

1059 The upper bound for parameters in the binomial models still remains 1. To determine the  
 1060 upper bound for multinomial models we must, additionally for each element  $\theta_k$ , take into  
 1061 account the number of free parameters that share common upper and lower bounds  
 1062 (denoted with  $f_k$ ). The upper bound is then defined as:

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

1063 The elements in the remaining stick are then computed as follows

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

1064 The rationale behind these modifications will be described in more detail in the following  
 1065 sections. In **multibridge**, information that is relevant for the transformation of the  
 1066 parameter vectors is stored in the generated `restriction_list` which is returned by the  
 1067 main functions `binom_bf_informed` and `mult_bf_informed` but can also be generated  
 1068 separately with the function `generate_restriction_list`. This restriction list features  
 1069 the sublist `inequality_constraints` which encodes the number of equality constraints

1070 collapsed in each parameter in `nr_mult_equal`. Similarly the number of free parameters  
1071 that share common bounds are encoded under `nr_mult_free`.

1072 **Equality Constrained Parameters.** In cases where informed hypotheses feature  
1073 a mix of equality and inequality constrained parameters, we compute the Bayes factor  
1074  $\text{BF}_{re}$ , by multiplying the individual Bayes factors for both constraint types with each other:

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

1075 where the subscript 1 denotes the hypothesis that only features equality constraints and  
1076 the subscript 2 denotes the hypothesis that only features inequality constraints. To receive  
1077  $\text{BF}_{2e} \mid \text{BF}_{1e}$ , we collapse all equality constrained parameters in the constrained prior and  
1078 posterior distributions into one category. This collapse has implications on the performed  
1079 transformations.

1080 When transforming the samples from the collapsed distributions, we need to account  
1081 for the fact that the inequality constraints imposed under the original parameter values  
1082 might not hold for the collapsed parameters. Consider, for instance, a multinomial model  
1083 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  $\theta_2$ ,  $\theta_3$ , and  $\theta_4$  into  $\theta_*$ . 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  $\theta_k$  we thus need to take into account both the number of collapsed categories in the preceding parameter  $e_{k-1}$  as well as the number of collapsed categories in the current parameter  $e_k$ . Thus, lower bounds for the parameters

need to be adjusted as follows:

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

which leads to Equation C2. In this equation,  $e_{k-1}$  and  $e_k$  refer to the number of equality constrained parameters that are collapsed in  $\theta_{k-1}$  and  $\theta_k$ , respectively. In the example above, this means that to determine the lower bound for  $\theta_*$  we multiply the preceding value  $\theta_1$  by three, such that the lower bound is  $(\frac{0.05}{1}) \times 3 = 0.15$ . In addition, to determine the lower bound of  $\theta_5$  we divide the preceding value  $\theta_*$  by three, that is,  $(\frac{0.45}{3}) \times 1 = 0.15$ . Similarly, to determine the upper bound for a given parameter value  $\theta_k$ , we need to multiple the upper bound by the number of parameters that are collapsed within it:

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

where  $1 - \sum_{i < k} \theta_i$  represents the length of the remaining stick and the number of elements in the remaining stick are computed as follows:  $ERS = \sum_k^K e_k$ . For the example above, the upper bound for  $\theta_*$  is  $\frac{1 - 0.05}{5} \times 3 = 0.57$ . The upper bound for  $\theta_5$  is then  $\frac{(1 - 0.05 - 0.45)}{2} \times 1 = 0.25$ .

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

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

This hypothesis specifies that  $\theta_2$  and  $\theta_3$  have the shared lower bound  $\theta_1$  and the shared upper bound  $\theta_4$ , however,  $\theta_2$  can be larger than  $\theta_3$  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  $\theta_k$ , we need to subtract from the length of the remaining stick the lower bound from the parameters that are free to vary. However, only those parameters are included in this calculation that have not yet been transformed:

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

where  $f_k$  represents the number of free parameters that share common bounds with  $\theta_k$  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  $\theta_k$ :

$ERS = 1 + \sum_{j > k} f_j$ . 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  $\theta_1$  is simply  $\frac{1}{4}$ . For  $\theta_2$ , we need to take into account that  $\theta_2$  and  $\theta_3$  share common bounds. To compute the upper bound for  $\theta_2$ , we subtract from the length of the remaining stick the lower bound of  $\theta_3$ :  $\frac{1 - 0.15 - (1 \times 0.15)}{1 + 1} = 0.35$ .

A further correction is required if a preceding free parameter (i.e., a parameter with common bounds that was transformed already) is larger than the upper bound of the current parameter. For instance, in our example the upper bound for  $\theta_3$  would be

$\frac{1 - 0.44 - 0}{1 + 1} = 0.28$ , which is smaller than the value of the preceding free parameter, which was 0.29. If in this case  $\theta_3$  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— $\theta_4$  would violate the constraint (i.e.,  $0.15 < (0.29, 0.275) \not\propto 0.285$ ). In these cases, the upper bound for the current  $\theta_k$  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:

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

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

1106 For our example the corrected upper bound for  $\theta_3$  would become  $2 \times 0.28 - 0.29 = 0.27$   
 1107 which secures the proper ordering for the remainder of the parameters. If in this case  $\theta_3$   
 1108 would take on the value close to its upper bound, for instance  $\theta_3 = 0.265$ ,  $\theta_4$ —due to the  
 1109 sum-to-one constraint—would take on the value 0.295 which would be in accordance with  
 1110 the constraint (i.e.,  $0.15 < (0.29, 0.265) < 0.295$ ).