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, Department of Psychology, PO Box 15906, 1001 NK Amsterdam, The Netherlands, E-mail: alexandra.sarafoglou@gmail.com

Abstract

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

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

Introduction

The most common way to analyze categorical variables is to conduct either binomial tests, multinomial tests, or chi-square goodness of fit tests. These tests compare the encompassing hypothesis to a null hypothesis that all underlying category proportions are either exactly equal, or follow a specific distribution. Accordingly, these tests are suitable when theories predict either the invariance of all category proportions or specific values. For instance, chi-square goodness of fit tests are commonly used to test Benford's law, which predicts the distribution of leading digits in empirical datasets (Benford, 1938; Newcomb, 1881). Often, however, the predictions that researchers are interested in are of a different kind. Consider for instance the weak-order mixture model of decision-making (Regenwetter & Davis-Stober, 2012). The theory predicts that individuals' choice preferences are weakly ordered at all times, that is, if they prefer choice A over B and B over C then they will also prefer A over C (Regenwetter, Dana, & Davis-Stober, 2011)—a well-constrained prediction of behavior. The theory is, however, silent about the exact values of each choice preference. Hence, the standard tests that compare $\mathcal{H}_e$ to $\mathcal{H}_0$ are unsuited to test the derived predictions. Instead, the predictions need to be translated into an informed hypothesis $\mathcal{H}_r$ that reflects the predicted ordinal relations among the parameters. Only then is it possible to adequately test whether the theory of weakly-ordered preference describes participants' choice behavior. Of course, researchers may be interested in more complex hypotheses, including ones that feature combinations of equality constraints, inequality constraints, and unconstrained category proportions. For instance, Nuijten, Hartgerink, Assen, Epskamp, and Wicherts (2016) hypothesized that articles published in social psychology journals would have higher error rates than articles published in other psychology journals. As in the previous example, the authors had no expectations about the exact error rate distribution across journals. Here, again, the standard tests are inadequate. Generally, by specifying informed hypotheses researchers and practitioners are able to "add theoretical expectations to the traditional alternative hypothesis' ' (Hoijtink, Klugkist, & Boelen, 2008, p. 2) and thus test hypotheses that relate more closely to their theories (Haaf, Klaassen, & Rouder, 2019; Rijkeboer & van den Hout, 2008).

In the Bayesian framework, researchers 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:

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

where the subscript r denotes the informed hypothesis and e denotes the encompassing hypothesis. Several available `R` packages compute Bayes factors for informed hypotheses. For instance, the package **multinomineq** (Heck & Davis-Stober, 2019) evaluates informed hypotheses for multinomial models as well as models that feature independent binomials. The package **BFpack** (Joris Mulder et al., in press) evaluates informed hypotheses for statistical models such as univariate and multivariate normal linear models, generalized linear models, special cases of linear mixed models, survival models, and relational event models. The package **BAIN** (Gu, Hoijtink, Mulder, & Rosseel, 2019) evaluates informed hypotheses for structural equation models. Outside of `R`, the Fortran 90 program **BIEMS** (Joris Mulder, Hoijtink, & de Leeuw, 2012) evaluates informed hypotheses for multivariate linear models such as MANOVA, repeated measures, and multivariate regression. All these packages rely on one of two implementations of the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005; Sedransk, Monahan, & Chiu, 1985) to approximate order constrained Bayes factors: the unconditional encompassing method (Klugkist et al., 2005 ; Hoijtink, 2011; Hoijtink et al., 2008) and the conditional encompassing method (Gu, Mulder, Deković, & Hoijtink, 2014; Laudy, 2006; Joris Mulder, 2014; J. Mulder, 2016; J. Mulder et al., 2009). Even though the encompassing prior approach is currently the most common method to evaluate informed hypotheses, it becomes increasingly unreliable and inefficient as the number of restrictions increases or the parameter space of the restricted model decreases (Sarafoglou et al., in press). For instance, simulation studies conducted by Sarafoglou et al. (in press) have illustrated that the encompassing prior approach is not able to produce Bayes factors when hypotheses with a large number of constrained parameters are considered (i.e., they considered 18 categories). For hypotheses with less number of categories (i.e., 5 or 6), the method worked well when the data were not extreme and provided either weak or moderate evidence in favor of or against the informed hypothesis. However, when the data provided extreme evidence against the predicted constraints, the method again failed to compute Bayes factors.

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

Here we showcase how the proposed bridge sampling routine by Sarafoglou et al. (in press) can be performed with **multibridge**. In the remainder of this article, we will introduce the package and its functionalities and describe the methods used to compute the informed hypotheses in binomial and multinomial models. We will illustrate its core 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 α and β parameters of the binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions then return the estimated Bayes factor for the informed hypothesis relative to the encompassing 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 ordinal relations and equality constraints among parameters (left panel in Figure 2). Additionally, **multibridge** supports the computation of Bayes factors for multiple independent constraints (middle panel in Figure 2), for instance, the simultaneous

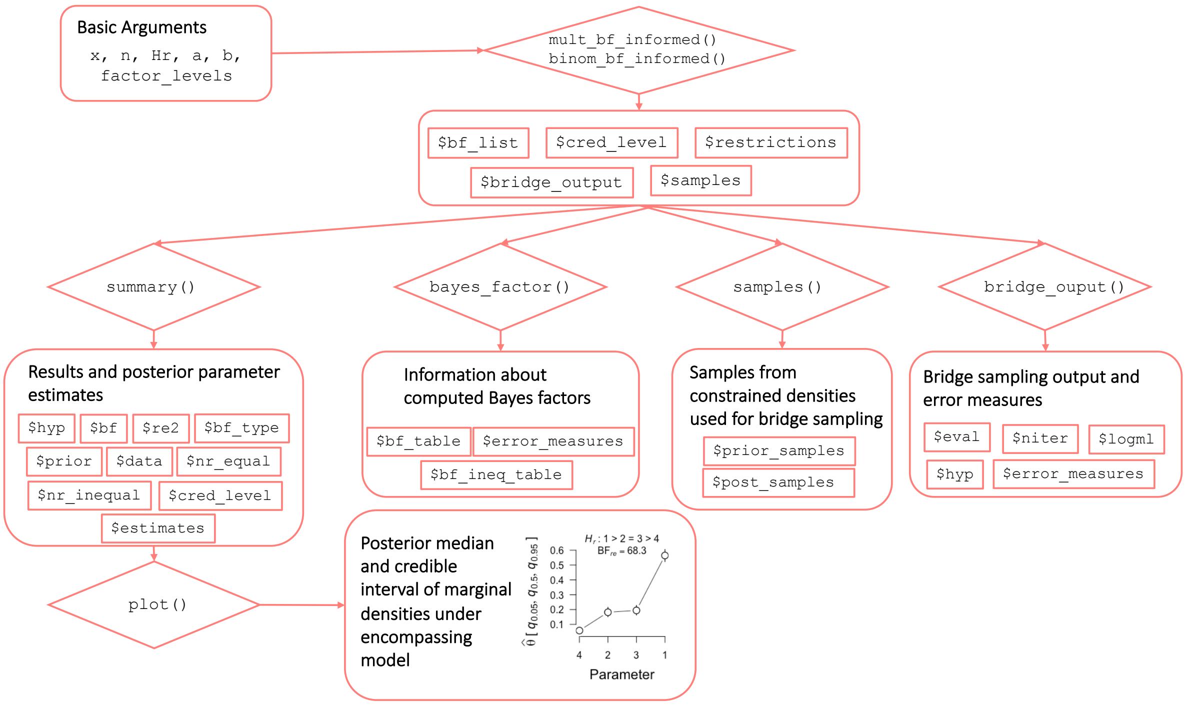


Figure 1. The **multibridge** workflow. The functions `mult_bf_informed` or `binom_bf_informed` return the estimated Bayes factor for the informed hypothesis relative to the encompassing or the null hypothesis. Based on these results different S3 methods can be used to get more detailed information on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

evaluation of inequality constraints on the first three category proportions and an equality constraints on the fifth and sixth category proportion. The package also supports the evaluation of combinations of equality constraints, inequality constraints, and free parameters (right panel in Figure 2). As an example, consider an ordinal hypothesis that identifies a smallest and a largest parameter, and equates the remaining parameters.

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.

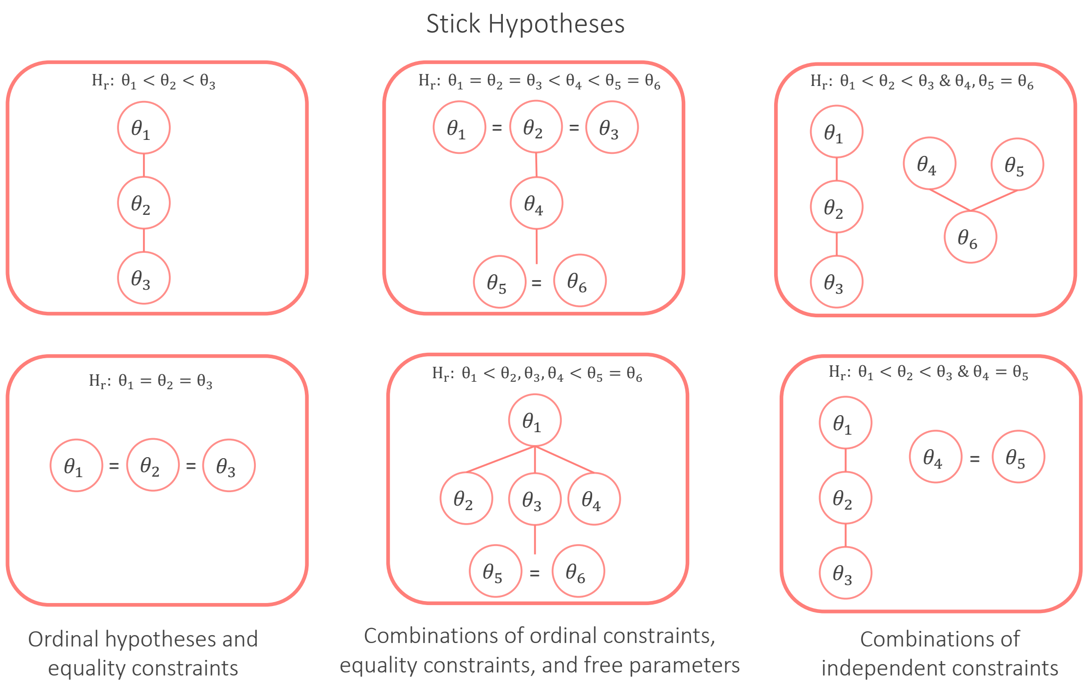


Figure 2. **multibridge** supports informed hypotheses including inequality and equality constraints (left), independent hypotheses (middle), and combinations of inequality and equality constraints and free parameters (right).

An important requirement for the hypotheses supported in **multibridge** is that the constrained parameters share upper and lower bounds. That is, if the constraint was to be drawn as a Hasse diagram or specified as a character vector, the constrained parameters should string together like a chain, ranging from the smallest parameter to the largest. We refer to these hypotheses as "stick-hypotheses". Conversely, "branched-hypotheses", that is, hypotheses that do not share common upper and lower bounds are currently not supported in **multibridge**. Examples for branched-hypotheses are shown in Figure 3. 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).

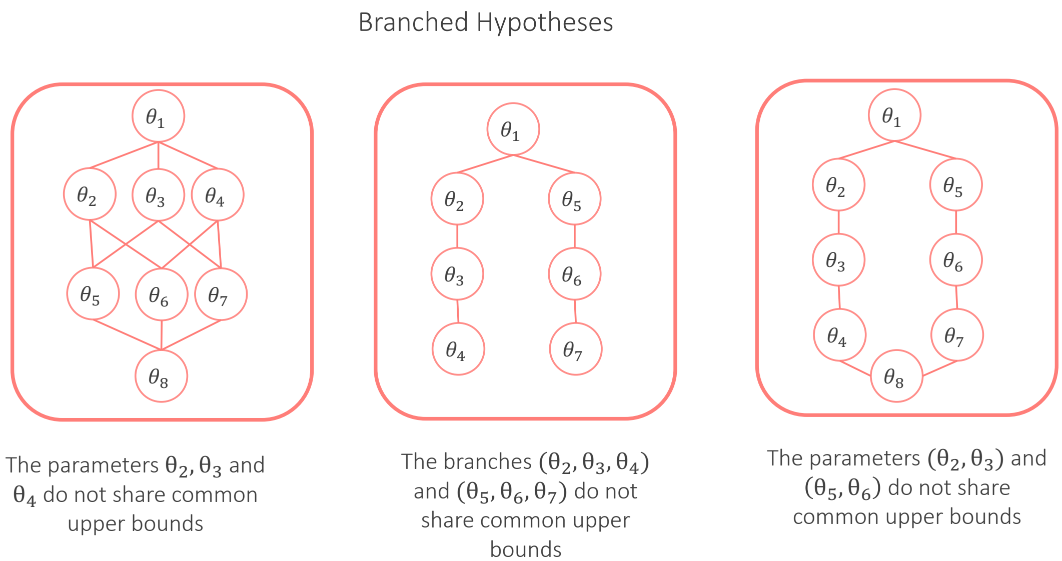


Figure 3. A prerequisite of **multibridge** is that informed hypotheses can be arranged as a stick. Branched hypotheses are currently not supported in the package.

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

Permitted signs to specify informed hypotheses are the ‘"<"‘-sign and ‘">"‘-sign for inequality constraints, the ‘"="‘-sign for equality constraints, and the ‘","‘-sign for parameters to vary freely within a constraint. For instance, ‘"t1 > t2 , t3 , t4"‘ states that `t1` is bigger than (`t2`, `t3`, `t4`) and that no constraints are imposed among `t2`, `t3`, and `t4`, thus they vary freely. Lastly, users can connect multiple independent restrictions using the ‘"&"‘-sign, for instance, ‘"t1 > t2 > t3 & t5 = t6"‘.

When evaluating equality constraints, it should be noted that there is a difference between assuming equality of category proportions and assuming that categories can be merged, that is, the hypothesis $\mathcal{H}_r : \theta_1 = \theta_2 < \theta_3 = \theta_4$ is not the same as $\mathcal{H}_r : \theta_1 + \theta_2 < \theta_3 + \theta_4$. In the first case the hypotheses concerns four categories of which two are expected to have equal category proportions. As a result, we assign priors to each of these four categories. In the second case, the hypothesis concerns only two categories since we assume that θ_1 and θ_2 belong to one group and θ_3 and θ_4 belong to the other. Consequently, one would assign prior to only two categories instead of four. If the goal is to merge observations of different categories, one can combine the counts and use the new data to conduct the analysis on.

Table 2

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

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

multibridge is designed such that the functions `mult_bf_informed` or `binom_bf_informed` combine most supported functionalities of the package. Other available functions compute Bayes factors for hypotheses that postulate only equality or only inequality constraints, and draw from constrained multinomial distributions and distributions of multiple independent binomials. 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
`mult_bf_informed`	Evaluates informed hypotheses on multinomial parameters.
`mult_bf_inequality`	Estimates the marginal likelihood of a constrained prior or posterior Dirichlet distribution.
`mult_bf_equality`	Computes Bayes factor for equality constrained multinomial parameters using the standard Bayesian multinomial test.
`mult_tsampling`	Samples from constrained prior or posterior Dirichlet density.
`lifestresses`, `peas`	Data sets associated with informed hypotheses in multinomial models.
`binom_bf_informed`	Evaluates informed hypotheses on binomial parameters.
`binom_bf_inequality`	Estimates the marginal likelihood of constrained prior or posterior beta distributions.
`binom_bf_equality`	Computes Bayes factor for equality constrained binomial parameters.
`binom_tsampling`	Samples from constrained prior or posterior beta densities.
`journals`	Data set associated with informed hypotheses in binomial models.
`generate_restriction_list`	Encodes the informed hypothesis.

Methodological Background

In this section we provide background information on the methods implemented in **multibridge**. Specifically, this section formalizes multinomial models and models that feature independent binomial probabilities, defines Bayes factors for the Bayesian multinomial and binomial test. Furthermore, the section discusses the influence of priors on the Bayes factors, illustrates how to compute posterior model probabilites and how to compare two informed hypotheses with each other, and provides a non-technical introduction into the bridge sampling routine implemented in **multibridge**. Mathematical details of the methods and principles discussed here can be found in Sarafoglou et al. (in press) and Gronau et al. (2017).

In the binomial model, we assume that the elements in the vector of successes $\mathbf{x}$ and the elements in the vector of total number of observations $\mathbf{n}$ in the K categories follow independent binomial distributions $\mathbf{x} \sim \prod_{k=1}^{K} \text{Binomial}(\theta_k, n_k)$, from which we can derive the likelihood of the data given the parameters:

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

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

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

where $\text{B}(\alpha_k,\, \beta_k)$ is the beta function:

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

The multinomial model constitutes a generalization of the binomial model (for $K \geq 2$). In this model, we assume that the vector of observations $\mathbf{x}$ in the K categories follows a multinomial distribution in which the parameters of interest, $\boldsymbol{\theta}$, represent the underlying category proportions, thus $\mathbf{x} \sim \text{Multinomial}(x_+, \boldsymbol{\theta})$, where $x_+ = \sum_{k=1}^{K} x_k$.

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

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

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

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

Developing Suitable Prior Distributions

In the binomial and multinomial model, the concentration parameters have an intuitive interpretation. In the binomial model, the parameters α_k can be interpreted as vector of *a priori* successes that observations fall within the various categories and β_k can be interpreted as vector of *a priori* failures. Likewise, in the multinomial model, α_k can be interpreted as vector of *a priori* category counts. It follows, that the higher the number of concentration parameters is, the information the prior contains and the more influence it has on parameter estimation and hypothesis testing.

To assign adequate priors for the multiple binomials and multinomial model, we recommend one of the following approaches. If researchers possess no knowledge or expectations about the plausible parameter values, a uniform distribution can be assigned across the parameter space. This prior assumes that before seeing the data, each category contains one observation, that is, all concentration parameters are set to one. A uniform prior distribution, puts equal probability mass on all permitted parameter values, similar to the adjusted priors proposed by Heck and Wagenmakers (2016) (see Figure 4). In contrast to the method proposed in Heck and Wagenmakers (2016) , however, **multibridge** allows priors to be set on the original scale.

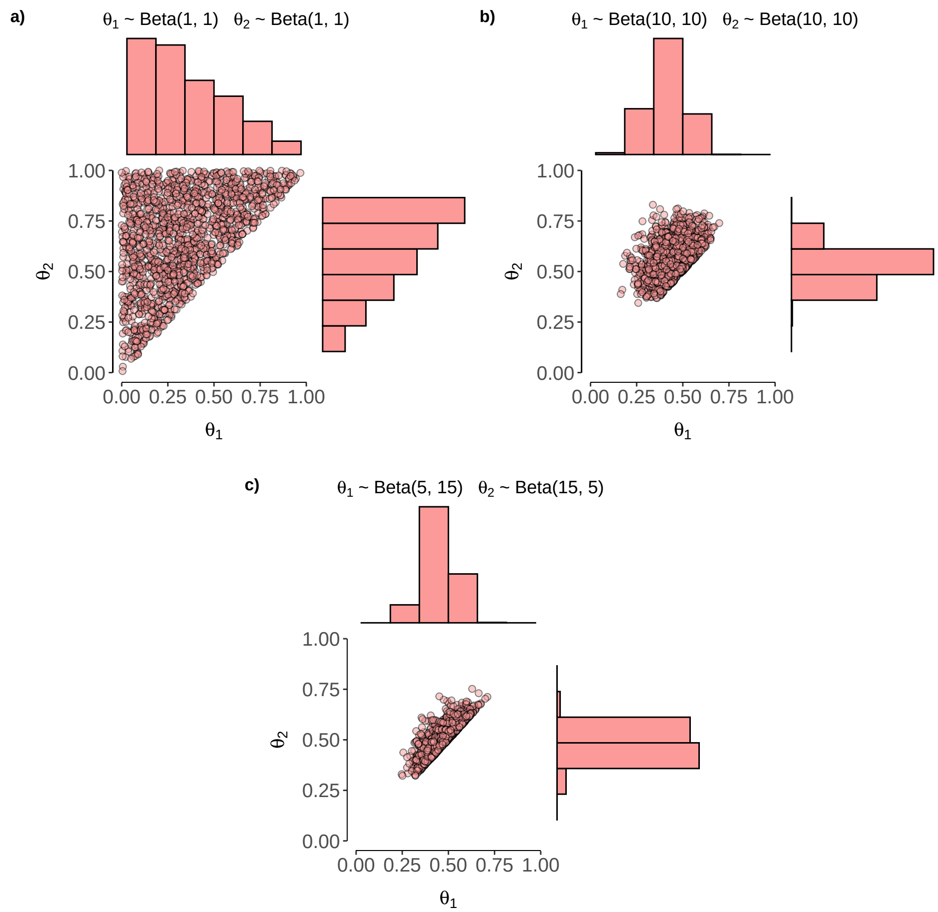


Figure 4. The development of a prior distribution should be accompanied by a visual inspection of the prior predictive. Here we display three prior distributions on two binomial probabilities that are constrained to be $\theta_1 < \theta_2$. The uniform distribution (panel a) assigns equal mass to all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

We recommend incorporating priors knowledge into the models whenever possible. Based on theories, expert knowledge, or informed guesses, researchers often have expectations about plausible and implausible parameter values. In these cases, the prior should match these expectations (Lee & Vanpaemel, 2018). For instance, in the case of informed hypotheses, prior counts can be chosen to match a particular expected ordinal trend. To determine whether the chosen priors are consistent with the theory, researchers can visualize and assess prior predictive distributions, that is, the distribution of the model parameters and data patterns predicted by the priors (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019; Schad, Betancourt, & Vasishth, 2021; Wagenmakers et al., 2021). The developed priors should reflect the theory and make reasonable preductions, but not be too informative to influence on posterior parameter estimates.

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

Bayes factor

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

Testing Equality Constraints. For equality constrained binomial models **multibridge** supports two kinds of null hypotheses, one which states that all parameters are equal and one which states that all parameters are equal and equal to a specific value. Both null hypotheses are tested against an encompassing hypothesis. Under the encompassing hypothesis, we specify a Beta(α_k, β_k) prior on each of the θ_k that yields the following marginal likelihood:

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

Under the first null hypothesis which states that all binomial probabilities are set equal without them being constrained further, we collapse all individual Beta(α_k, β_k) priors and corrects for the change in categories; if K categories are collapsed, $K - 1$ is subtracted from the concentration parameters. A Beta(1, 1) prior on the individual category proportions thus also yields to a Beta(1, 1) prior when all categories are collapsed. Hence, we yield a Beta($\alpha_+ - K - 1$, $\beta_+ - K - 1$) prior on θ, where $\alpha_+ = \sum_{k=1}^{K} \alpha_k$ and $\beta_+ = \sum_{k=1}^{K} \beta_k$. This yields the following marginal likelihood:

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

We can now compute the Bayes factor BF_{01e} as follows:

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

The second null hypothesis states that all binomial probabilities in a model are assumed to be exactly equal *and* equal to a predicted value θ_0. Under this hypothesis, the prior reduces to a single point and the marginal likelihood simplifies to the likelihood:

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

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

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

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)
```

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

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

where θ_{0k} represent the predicted category proportions (see Sarafoglou et al., in press for the derivation). For multinomial models, under the null hypothesis, category probabilities can either all be set equal (i.e., all category probabilities are $\frac{1}{K}$) or can replaced with the 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))
```

Testing Inequality Constraints. For inequality constrained binomial and multinomial models, users can specify informed hypotheses that are either tested against a null hypothesis postulating that all parameters are equal or against the encompassing hypothesis which lets all parameters free to vary. Generally, to obtain the marginal likelihood of the informed hypothesis, it is necessary to integrate over the restricted parameter space, which is difficult to compute. As a solution to the problem of computing marginal likelihood of the informed hypothesis, Klugkist et al. (2005) derived an identity that defines the Bayes factor BF_{re} as the ratio of proportions of posterior and prior parameter space consistent with the restriction. This identity forms the basis of the encompassing prior approach. Recently, Sarafoglou et al. (in press) highlighted that these proportions can be reinterpreted as the marginal likelihoods (i.e., the normalizing constants) of the constrained posterior and constrained prior distribution.}

The constrained prior distribution of the parameters subject to an informed hypothesis $\mathcal{H}_r$ take the following form:

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

The constrained posterior distribution of the parameters under the informed hypothesis can be represented in the same way.

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

where $\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$ is an indicator function that is one for parameter values in the that obey the constrained and zero otherwise. The Klugkist et al. (2005) identity is thus:

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

This reformulation of the Klugkist et al. (2005) identity as a ratio of marginal likelihoods, made it possible to utilize numerical sampling methods such as bridge sampling to compute the Bayes factor. The following section provides a conceptual introduction to bridge sampling how it is used in the context of evaluating informed hypotheses.

Bridge Sampling Routine

The bridge sampling routine implemented in **multibridge** is a numerical method to estimate the marginal likelihood of a target density (cf., Gronau et al., 2017; Overstall & Forster, 2010). The identity used in bridge sampling is displayed in Equation 2; it considers the unnormalized target density, a proposal density with known normalizing constant, and an arbitrary bridge function. The numerator in Equation 2 describes the expected value of the unnormalized target density evaluated with samples from the proposal density. The denominator is the expected value of the proposal density and a bridge function evaluated with samples from the target density. The bridge function serves the purpose of increasing the overlap between the two densities, thus increasing the efficiency and accuracy of the method. The bridge sampling identity can then be expressed as follows:

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

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

It should be noted that the bridge sampling algorithm implemented in **multibridge** is an adapted version of the algorithm implemented in the `R` package **bridgesampling** (Gronau, Singmann, & Wagenmakers, 2020) and allows for the specification of informed hypotheses on probability vectors.[1]

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

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

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

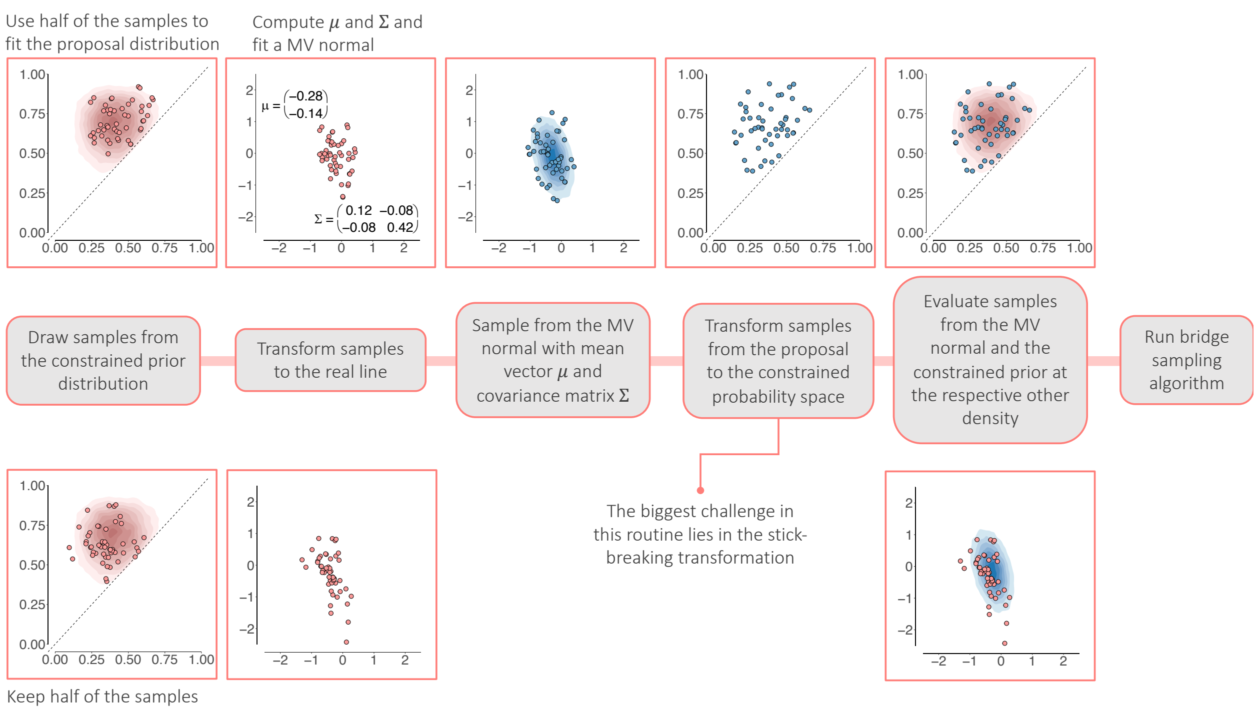


Figure 5. A schematic illustration of the steps taken to estimate the marginal likelihood of the constrained prior distribution of two binomial probabilities under $\mathcal{H}_r : \theta_1 < \theta_2$. As starting point, the routine requires samples from the constrained distribution. The bridge sampling algorithm estimates the marginal likelihood through the identity given in Equation 2 using an iterative scheme.

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

The numerator in Equation 2 evaluates the unnormalized density for the constrained prior distribution with samples from the proposal distribution. The denominator evaluates the normalized proposal distribution with samples from the constrained prior distribution. Using this identity, we obtain the bridge sampling estimator for the marginal likelihood of the constrained prior distribution by applying the iterative scheme proposed by Meng and Wong (1996). **multibridge** then runs the iterative scheme until the tolerance criterion suggested by Gronau et al. (2017) is reached. The sampling from the target and proposal distribution, the transformations and computational steps are performed automatically within the core functions of **multibridge**. The user only needs to provide the functions with the data, a prior and a specification of the informed hypothesis. As part of the standard output of `binom_bf_informed` and `mult_bf_informed`, the functions return the bridge sampling estimate for the log marginal likelihood of the target distribution, its associate relative mean square error and the number of iterations needed to until the bridge sampling estimator reached the tolerance criterion.

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

Stick-Breaking Transformation

The bridge sampling routine in **multibridge** uses the multivariate normal distribution as proposal distribution, which requires moving the target distribution to the real line. Crucially, the transformation needs to retain the ordering of the parameters, that is, it needs to take into account the lower bound and the upper bound of each parameter. To meet these requirements, **multibridge** uses a probit transformation, as proposed in Sarafoglou et al. (in press), and subsequently transforms the elements in the parameter vector, moving from its lowest to its highest value. A schematic illustration of the stick-breaking transformation is given in Figure 6, detailed technical details of the transformation are provided in the appendix.

To perform the transformation from a parameter vector on the real line to an ordered probability vector, we need to determine the lower and upper bound of each parameter. Consider an increasing trend of four parameters, that is, $\theta_1 < \theta_2 < \theta_3 < \theta_4$. The lower bound for for the smallest element in the parameter vector, θ_1, is 0. For θ_2, θ_3, and θ_4 the lower bound is the preceding element in the vector. That is, the lower bound for θ_2 is θ_1, lower bound for θ_3 is θ_2, and the lower bound for θ_4 is θ_3.

This definition holds for both binomial models and multinomial models. Differences in these two models appear only when determining the upper bound for each parameter. For binomial models, the upper bound for each parameter is 1. For multinomial models, due to the sum-to-one constraint the upper bounds need to be computed differently. As proposed in Frigyik, Kapila, and Gupta (2010) and Stan Development Team (2020) we represent $\boldsymbol{\theta}$ as unit-length stick which we subsequently divide into as many elements as there are parameters in the constraint Stan Development Team (2020). In this approach,

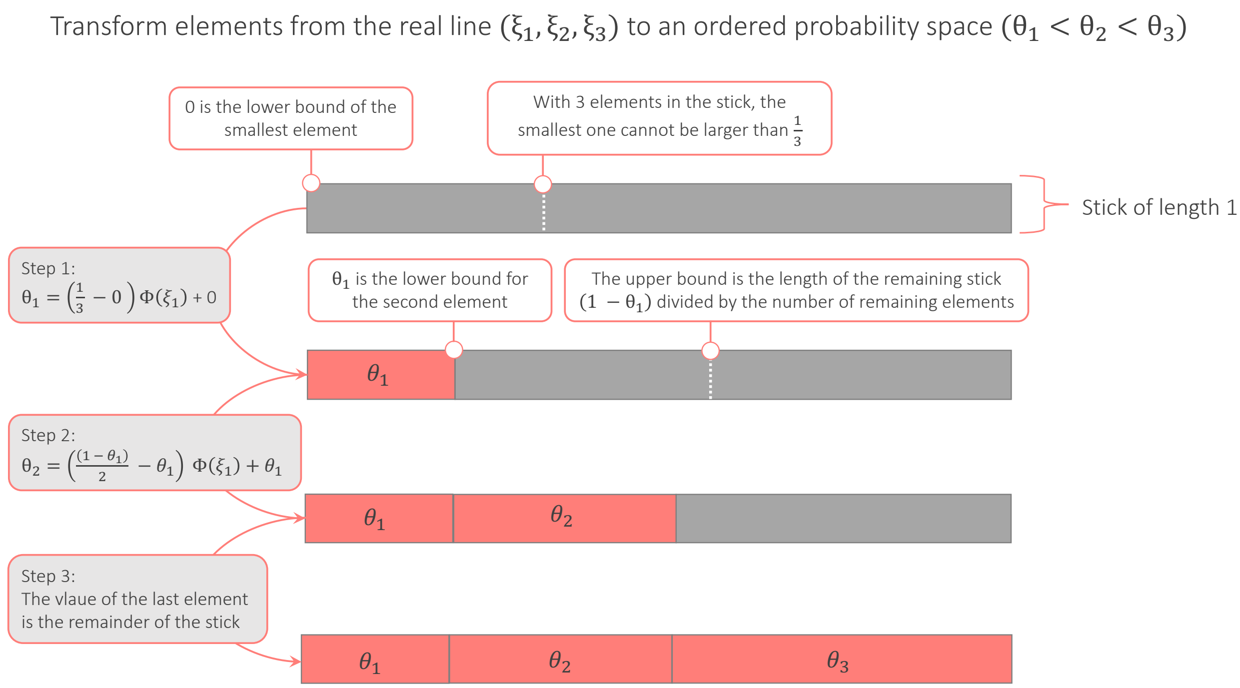


Figure 6. The stick-breaking transformation of elements on the real line to the ordered probability space. The stick-breaking transformation moves from the smallest to the largest value and determines the bounds of each parameter. The elements are transformed as follows: $\theta_k = (u_k - l_k)\Phi(\xi_k) + l_k$, where ξ_k is kth the element on the real line, Φ is the cumulative density function of a standard normal and u_k and l_k are the upper and lower bounds of ξ_k, respectively. The largest element is simply the remainder of the stick.

the upper bounds are derived from on the values of smaller elements as well as on the number of remaining larger parameters in the stick. Concretely, for the smallest element in the parameter vector, θ_1, the upper bound is $\frac{1}{4}$; if this element were larger than that it would be impossible to create a probability vector with increasing values. For θ_2, θ_3, and θ_4 the upper bound is the proportion of the unit-length stick that has not yet been accounted for in the transformation divided by the number of parameters in the remaining stick. For instance, the upper bound for θ_2 is defined as $\frac{1-\theta_1}{3}$. This transformation allows us to effectively transform elements from the real line to an constrained probability space and is therefore a main component of the bridge sampling algorithm.

One drawback of this transformation is, however, that it can only be performed if all parameters in the constraint can be stringed together like a chain, thus, only works for "stick-hypotheses". For hypotheses in which parameters do not share common lower and upper bounds, the assumption is violated that for a given parameter smaller elements and the number of parameters in the remaining stick determine their upper bound.

Poster Model Probabilites, and Bayes Factor Transitivity

Consider a scenario where one has a whole set of hypotheses that they want to compare with each other, for instance, two informed hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ as well as a null hypothesis $\mathcal{H}_0$ and the encompassing hypothesis $\mathcal{H}_e$. An overview of the relative plausibility of all $M = 4$ models simultaneously may be obtained by presenting the posterior model probabilities for all $i = 1, \cdots, 4$ hypotheses, $p(\mathcal{H}_i \,|\, x)$ Berger and Molina (2005). The computation of posterior model probabilities are not automatically computed in **multibridge**, however, after computing the individual Bayes factors, the posterior model probabilities can be derived using the following equation. Denoting the prior model probability for hypothesis $\mathcal{H}_{r1}$ by $p(\mathcal{H}_{r1})$, the posterior model probability $p(\mathcal{H}_{r1} \mid \mathbf{x})$ is given by:

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

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

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

Posterior model probabilities are useful for comparing multiple hypotheses with each other. However, it should be noted that posterior model probabilities are relative quantities and can change depending on which hypotheses are included in the comparison. Thus, hypotheses that describe the data poorly may have high posterior model probabilities if the other hypotheses in the comparison set are even worse descriptions of the data. In order to gain insight into whether a hypothesis describes the data adequately, we therefore include so-called bookend hypotheses in addition to the theory-informed hypotheses, that is, a hypothesis that maximally constrains the parameter space (such as a point-null hypothesis $\mathcal{H}_0$) and the encompassing hypothesis $\mathcal{H}_e$ that does not constrain the parameter space (in this case, that makes no ordinal predictions, Lee & Vanpaemel, 2018). A hypothesis is then considered adequate if it can outperform these bookend models.

Instead of posterior model probabilities, Bayes factors can also be calculated directly between two informed hypotheses. The comparison of any two informed hypotheses with one another follows from the fact that Bayes factors are transitive. For instance, the Bayes factor comparison between two informed hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ can be obtained by first computing BF_{er1} and BF_{er2}, and then dividing out the common hypothesis $\mathcal{H}_e$:

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

Prior Sensitivity

One of the main criticisms of Bayesian hypothesis testing is that the priors exert too much influence on the Bayes factors (e.g., Kass & Raftery, 1995). That is, even if the data are informative enough to overwhelm the prior for parameter estimation, priors can still influence the Bayes factors. The development of suitable priors is thus an important part of Bayesian hypothesis testing.

But even priors that are justified by theory are to a certain degree arbitrary. For instance, if one expects an increasing trend in the data, the parameters in the prior can be chosen to reflect that trend. The exact number of *a priori* category counts, however, is at the discretion of the analyst. It is therefore considered good research practice to conduct a sensitivity analysis on the final results. In a sensitivity analysis, a set of plausible priors are determined in addition to the prior chosen in the main analysis for which the Bayes factors are calculated. The range of Bayes factors then gives an indication of the extend to which the results are fragile or robust to different modeling choices. In general, the prior on which the final analysis is performed as well as the set of priors used to conduct the sensitivity analysis should be determined and preregistered before seeing the data to ensure a fair comparison of the hypotheses of interest.

Usage and Examples

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

The two core functions of **multibridge**—`mult_bf_informed` and the `binom_bf_informed`—can be illustrated schematically as follows:

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

Example 1: Applying A Benford Test to Greek Fiscal Data

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

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

This means that in an empirical data set, numbers with smaller leading digits are more common than numbers with larger leading digits. Specifically, a number has leading digit 1 in 30.1% of the cases, and leading digit 2 in 17.61% of the cases; leading digit 9 is the least frequent digit with an expected proportion of only 4.58% (see Table 4 for an overview of the expected proportions). Empirical data for which this relationship holds include population sizes, death rates, baseball statistics, atomic weights of elements, and physical constants (Benford, 1938). In contrast, 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 and 9 occurred least often which is consistent with the general pattern of Benford's law. However, the proportion for the remaining leading digits were approximately equal. Note that the predicted distribution derived from Hill (1995) is not currently used as a test to detect fraud, however, for the sake of simplicity, we assume that this pattern could be an indication of manipulated auditing data. All hypotheses will be tested against the encompassing hypothesis $\mathcal{H}_e$, which too serves as a bookend comparison hypothesis, and which imposes no constraints on the proportion of leading digits.

Data and Hypothesis. The data we use to illustrate the computation of Bayes factors were originally published by the European statistics agency "Eurostat" and served as basis for reviewing the adherence to the Stability and Growth Pact of EU member states. Rauch, Göttsche, Brähler, and Engel (2011) conducted a Benford test on data related to budget deficit criteria, that is, public deficit, public dept and gross national products. The data used for this example features the proportion of first digits from Greek fiscal data in the years between 1999 and 2010; a total of $N = 1{,}497$ numerical data were included in the analysis. We choose this data, since the Greek government deficit and debt statistics states has been repeatedly criticized by the European Commission in this time span (European Commision, 2004, 2010). In particular, the commission has accused the Greek statistical authorities to have misreported deficit and debt statistics. For further details on the data set see Rauch et al. (2011). The observed and expected proportions are displayed in Table 4; the expected proportions versus the posterior parameter estimates under the encompassing hypothesis are displayed in Figure 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 Proportions: Benford's Law
1	509	0.340	0.301
2	353	0.236	0.176
3	177	0.118	0.125
4	114	0.076	0.097
5	77	0.051	0.079
6	77	0.051	0.067
7	53	0.035	0.058
8	73	0.049	0.051
9	64	0.043	0.046

In this example, the parameter vector of the multinomial model, $\theta_1, \cdots, \theta_K$, reflects the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9. 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:

$$\mathcal{H}_e : \boldsymbol{\theta} \sim \text{Dirichlet}(\mathbf{1})$$
$$\mathcal{H}_0 : \boldsymbol{\theta}_0 = (0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046),$$
$$\mathcal{H}_{r1} : \theta_1 > \theta_2 > \theta_3 > \theta_4 > \theta_5 > \theta_6 > \theta_7 > \theta_8 > \theta_9$$
$$\mathcal{H}_{r2} : \boldsymbol{\theta}_0 = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\right)$$
$$\mathcal{H}_{r3} : \theta_1 > (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7) > (\theta_8,\ \theta_9).$$

Method. Both BF_{0e} and BF_{r2e} may be readily computed by means of a Bayesian multinomial test which is implemented in the function `mult_bf_equality`. This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, and (3) the vector of expected proportions under $\mathcal{H}_0$ and under $\mathcal{H}_{r2}$. In this example, we do not incorporate specific expectations about the distribution of leading digits in the Greek fiscal data and therefore assign a uniform Dirichlet distribution to the proportion of leading digits. That is, we set all concentration parameters under $\mathcal{H}_e$ to 1 (i.e., we assign $\boldsymbol{\theta}$ a uniform Dirichlet prior distribution). This prior supports all possible points equally, meaning that, if the data were completely 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 BF_{r1e} and BF_{r3e}. This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, (3) labels for the categories of interest (i.e., leading digits), and (4) the informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r3}$ (e.g., as a string). In addition to the basic required arguments, we use two additional arguments here. The first argument sets the Bayes factor type, that is, whether the output should print the Bayes factor in favor of the informed hypothesis (i.e., BF_{re}) or in favor of the encompassing hypothesis (i.e., 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:

```
# 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)
# Labels for categories of interest
factor_levels <- 1:9
# Specifying the informed hypotheses as a string
Hr1 <- c('1 > 2 > 3 > 4 > 5 > 6 > 7 > 8 > 9')
Hr3 <- c('1 > 2 = 3 = 4 = 5 = 6 = 7 > 8 , 9')
# Execute the analysis
results_He_Hr1 <- mult_bf_informed(x = x, Hr = Hr1, a = a,
                                   factor_levels = factor_levels,
```

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

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

Table 5

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

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mid \mathbf{x})$	$\log(\text{BF}_{.e})$
$\mathcal{H}_0$	0.2	1.27×10^{-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	5.97×10^{-79}	-172.70
$\mathcal{H}_{r2}$	0.2	2.71×10^{-212}	-479.73

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

The strong Bayes factor support for $\mathcal{H}_{r1}$ translates to a relatively extreme posterior model probability of 0.9994. By comparison, the posterior model probabilities for hypotheses $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$, that is, the bookend null-hypothesis and the hypothesis predicting a data pattern typical of fraud, are only slightly greater than zero. The posterior model probability for $\mathcal{H}_e$ is 0.0006. Thus, hypothesis $\mathcal{H}_{r1}$ can outperform the two bookend hypotheses $\mathcal{H}_{r2}$ and $\mathcal{H}_e$. That $\mathcal{H}_{r1}$ outperforms the unconstrained model $\mathcal{H}_e$} demonstrates how a parsimonious model that makes precise predictions can be favored over a model that is more complex (e.g., Jefferys & Berger, 1992).

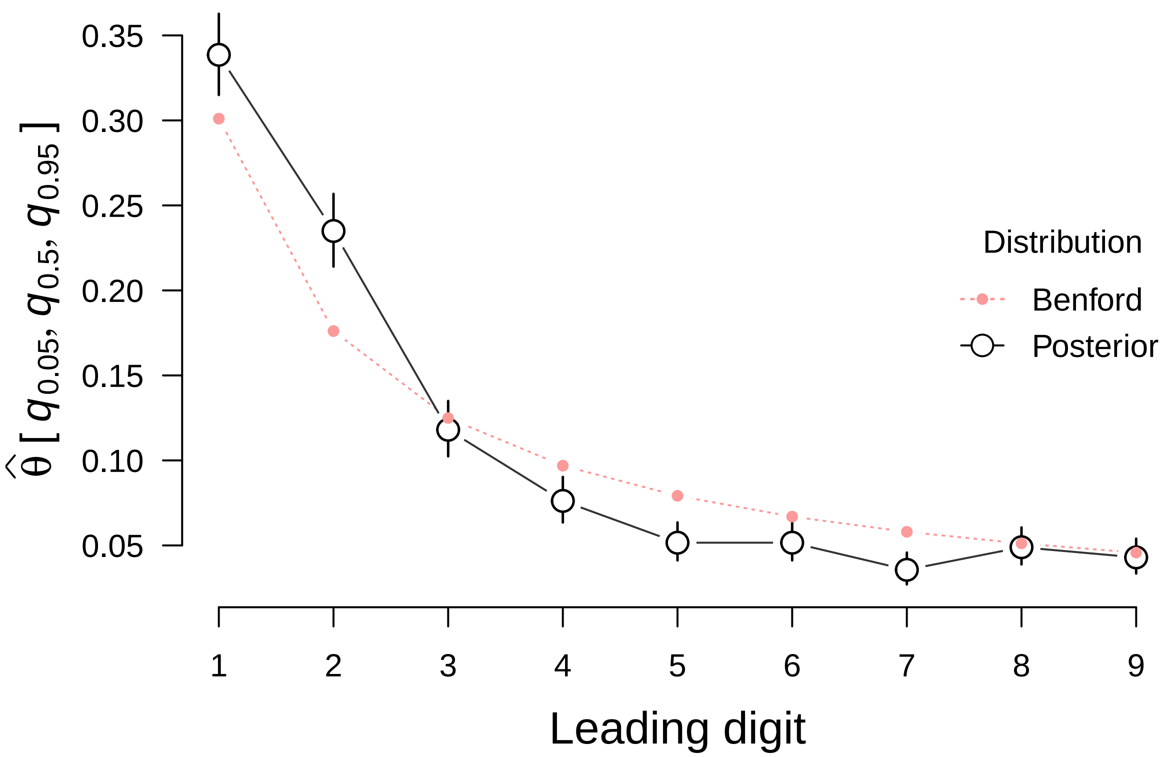


Figure 7. Predictions from Benford's law (in pink) show together with the posterior medians (black circles) for the category proportions estimated under the encompassing model $\mathcal{H}_e$. The circle skewers show the 95% credible intervals. Only three of nine intervals encompass the expected proportions, 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**.

Sensitivity Analysis. In a sensitivity analysis we will determine whether our results are robust against different prior choices. In the main analysis we chose a uniform Dirichlet distribution on the category proportions as prior under $\mathcal{H}_e$. This prior assigns equal probability to all possible parameter values, but alternative prior distributions are also plausible. Experienced audit researchers may argue for the development of more informative and theory-driven priors that resemble, for instance, one of the hypotheses under consideration. The Dirichlet parameters vectors specified below resemble the four 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
```

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

The results of the sensitivity analysis are displayed in Table 6. The general direction of the sensitivity analysis agrees with our conclusions drawn from the main analysis. That is, for the Bayes factors of $\mathcal{H}_{r1}$ compared to $\mathcal{H}_e$, the evidence points towards the informed hypothesis. However, the prior exerts an influence on BF_{r1e}; the evidence in favor for the informed hypothesis ranges from weak to extreme evidence. Specifically, when we choose priors that resemble a decreasing trend for the frequency of leading digits, as we did with $\boldsymbol{\alpha_0}$ and $\boldsymbol{\alpha_1}$, the Bayes factor becomes smaller and the evidence weak (i.e., $(\text{BF}_{r1e} \mid \boldsymbol{\alpha_0}) = 1.87$ on the natural scale) and moderate (i.e., $(\text{BF}_{r1e} \mid \boldsymbol{\alpha_1}) = 4.74$ on the natural scale). However, if the data are contrasted to a prior that makes different predictions, the evidence is very strong or extreme. Thus, a prior that closely resembles the predictive trend reduces to some degree the diagnostic value of the data.

The Bayes factors $\mathcal{H}_0$ compared to $\mathcal{H}_e$, on the other hand, are robust against different prior settings. Here too, the prior changes the Bayes factor estimate but in all cases the data suggests overwhelming evidence in favor of the encompassing hypothesis over 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	$\boldsymbol{\alpha_e} = (1, 1, 1, 1, 1, 1, 1, 1, 1)$	7.42	-17.67
Benford's law	$\boldsymbol{\alpha_0} = (16, 10, 7, 5, 4, 3, 3, 3, 2)$	0.63	-26.00
Montonically decreasing	$\boldsymbol{\alpha_1} = (10, 9, 8, 7, 6, 5, 4, 3, 2)$	1.56	-20.94
Centered on mean	$\boldsymbol{\alpha_2} = (6, 6, 6, 6, 6, 6, 6, 6, 6)$	7.53	-11.35
Fraud pattern	$\boldsymbol{\alpha_3} = (12, 6, 6, 6, 6, 6, 6, 3, 3)$	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 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` package `statcheck`, which scans the PDF of a given scientific article and automatically detects statistical inconsistencies. This package allowed Nuijten et al. (2016) to estimate the prevalence of statistical reporting errors in the field of psychology. In total, the authors investigated a sample of 30,717 articles (which translates to over a quarter of a million *p*-values) published in eight major psychology journals between 1985 to 2013: *Developmental Psychology* (DP), the *Frontiers in Psychology* (FP), the *Journal of Applied Psychology* (JAP), the *Journal of Consulting and Clinical Psychology* (JCCP), *Journal of Experimental Psychology: General* (JEPG), the *Journal of Personality and Social Psychology* (JPSP), the *Public Library of Science* (PLoS), *Psychological Science* (PS).

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

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

```
data(journals)
```

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

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

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

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

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

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

```r
# Since percentages are rounded to two decimal values, we round the
# articles with an error to obtain integer values
x <- round(journals$articles_with_NHST  *
             (journals$perc_articles_with_errors/100))
# Total number of articles
n <- journals$articles_with_NHST
# Prior specification for beta prior distributions under H_e
a <- c(1, 1, 1, 1, 1, 1, 1, 1)
b <- c(1, 1, 1, 1, 1, 1, 1, 1)
# Labels for categories of interest
journal_names <- journals$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)
```

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

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

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

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

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

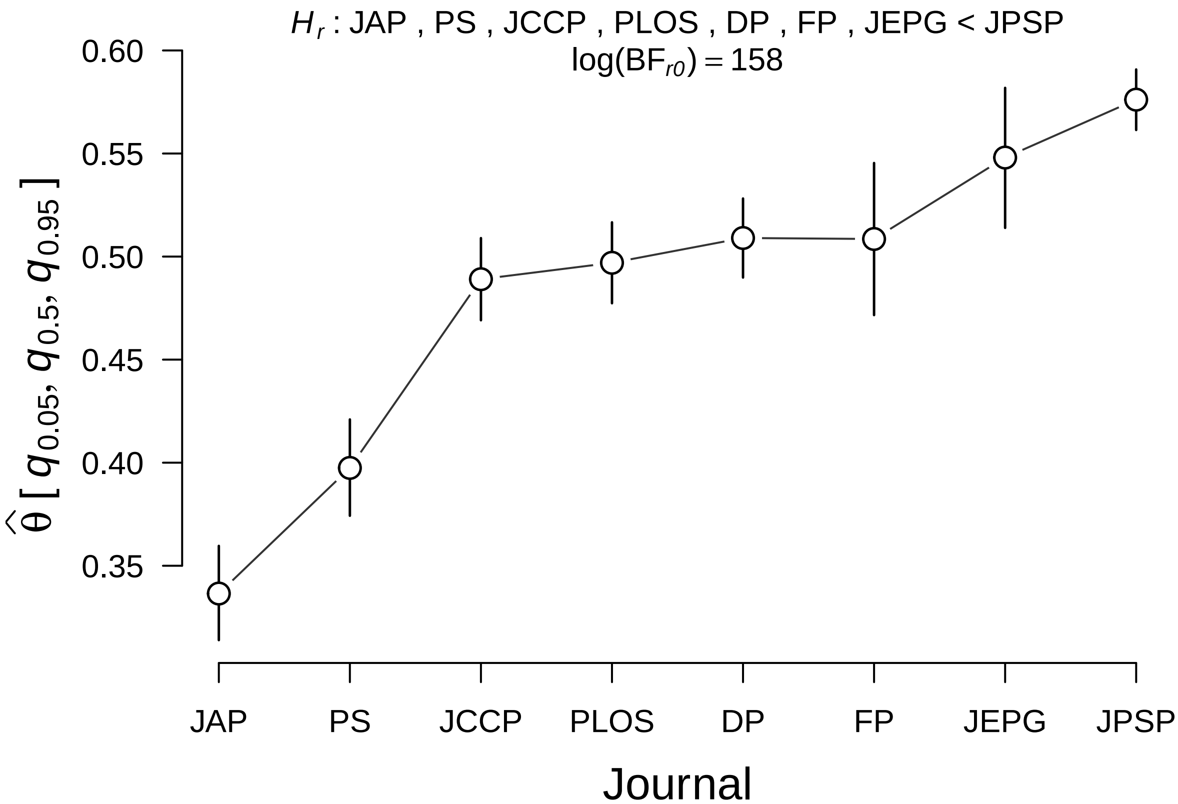


Figure 8. Posterior medians for the prevalence of statistical reporting inconsistencies across eight psychology journals, as obtained using the encompassing model. The circle skewers show the 95% credible intervals. Analysis based on data from Nuijten et al. (2016). This plot was created using the `plot`-S3-method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

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

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

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

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

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

The author analyzed the proportion of stochastic violations for different demographic variables, noting a seemingly ordinal pattern for the probabilities to violate of stochastic dominance for the factors gender (m=male, f=female) and education (1=doctorate, 2=postgraduate degree, 3=bachelor's degree, 4=less than bachelor's degree). In a later study, Myung, Karabatsos, and Iverson (2005) presented a Bayesian inference framework to test decision making axioms (using the "Bayesian p-value") and used Birnbaum's data as an example on how to assess violances of stochastic dominance and their relationship with covariates. Concretely, Myung et al. (2005) reanalyzed the data from Birnbaum (1999) and tested the informed hypothesis that stochastic dominance is violated more frequently in women compared to men and more frequently in lower education levels than higher education levels.

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

```
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))
```

The parameter vector of the binomial success probabilities, $\theta_1, \cdots, \theta_K$, contains the probabilities of observing a value in a particular category; here, it reflects the probabilities of violating stochastic dominance for a particular subgroup (e.g., females with a doctorate). We will compare three inequality-constrained hypotheses $\mathcal{H}_{r1}$, $\mathcal{H}_{r2}$, $\mathcal{H}_{r3}$ formulated by Myung et al. (2005). The first hypothesis $\mathcal{H}_{r1}$ encodes the main effect for gender and states that the probability to violate stochastic dominance is lower for males than for females. The second hypothesis $\mathcal{H}_{r2}$ encodes the main effect of education and states that the probability to violate stochastic dominance is lower for persons with higher education 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
Male		
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
Female		
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

hypothesis against the encompassing hypothesis $\mathcal{H}_e$ without any constraints. In addition, we will include a bookend null-hypothesis $\mathcal{H}_0$ predicting that all probabilities are equal.

$$\mathcal{H}_e : (\theta_{\text{m}1}, \theta_{\text{m}2}, \theta_{\text{m}3}, \theta_{\text{m}4}, \theta_{\text{f}1}, \theta_{\text{f}2}, \theta_{\text{f}3}, \theta_{\text{f}4})$$
$$\mathcal{H}_0 : \boldsymbol{\theta}_0 = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right),$$
$$\mathcal{H}_{r1} : (\theta_{\text{m}1}, \theta_{\text{m}2}, \theta_{\text{m}3}, \theta_{\text{m}4}) < (\theta_{\text{f}1}, \theta_{\text{f}2}, \theta_{\text{f}3}, \theta_{\text{f}4})$$
$$\mathcal{H}_{r2} : (\theta_{\text{m}1}, \theta_{\text{f}1}) < (\theta_{\text{m}2}, \theta_{\text{f}2}) < (\theta_{\text{m}3}, \theta_{\text{f}3}) < (\theta_{\text{m}4}, \theta_{\text{f}4})$$
$$\mathcal{H}_{r3} : \theta_{\text{m}1} < \theta_{\text{f}1} < \theta_{\text{m}2} < \theta_{\text{f}2} < \theta_{\text{m}3} < \theta_{\text{f}3} < \theta_{\text{m}4} < \theta_{\text{f}4}.$$

Method. To evaluate the inequality-constrained hypothesis, we need to specify (1) a vector with observed successes, and (2) a vector containing the total number of observations, (3) the informed hypothesis, (4) a vector with prior parameters alpha for each binomial proportion, (5) a vector with prior parameters beta for each binomial proportion, and (6) the labels of the categories of interest (i.e., gender and education level). As with the previous two example, we assign a uniform Beta prior to the binomial probabilites:

```
# 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 propotion
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)
```

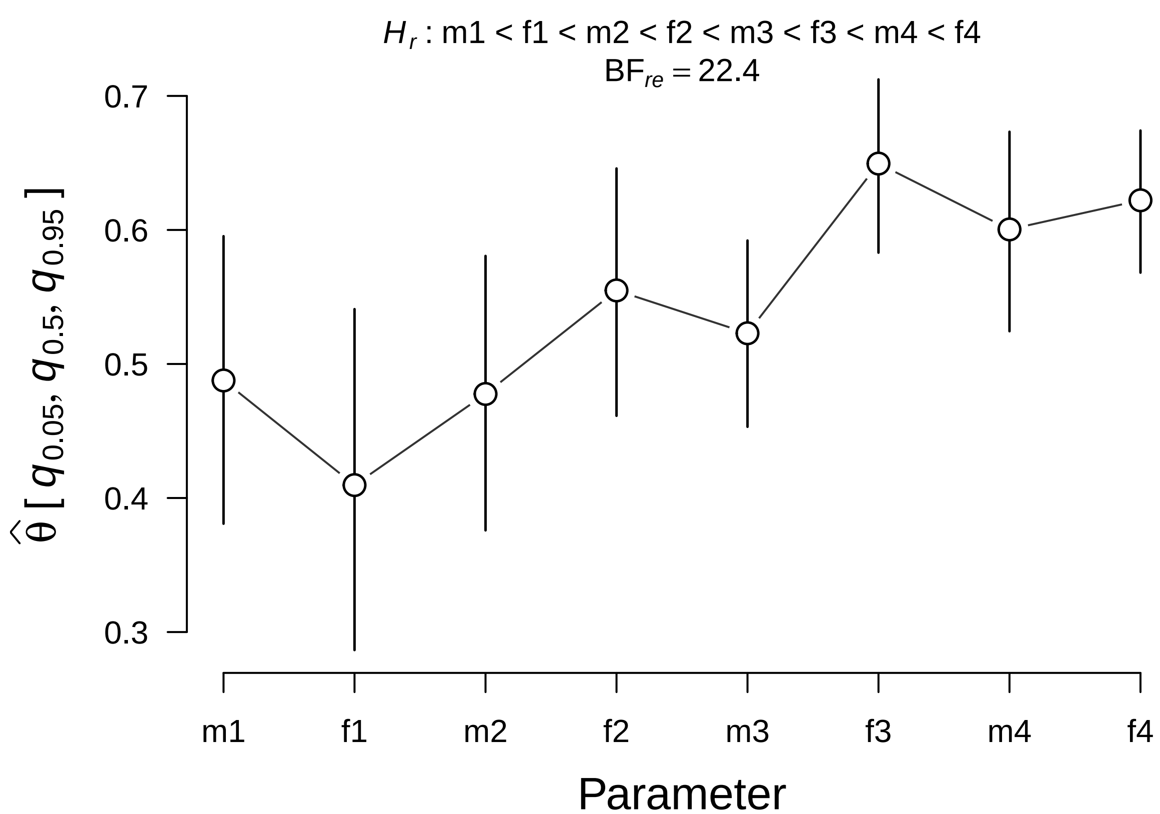
```
gender_edu <- dat$levels
```

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

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


results_Hr1_He <- multibridge::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 <- multibridge::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 <- multibridge::binom_bf_informed(x=x, n=n, Hr=Hr3, a=a, b=b,
                                             factor_levels=gender_edu,
                                             bf_type = 'BFre',
                                             seed = 2020)
```

The results for the analysis are summarized in Table 9. We first inspect the Bayes factors for the three informed hypotheses compared to the encompassing hypothesis. For hypotheses $\mathcal{H}_{r1}$, the data suggest moderate evidence for the encompassing hypothesis than compared to the informed hypothesis, with a Bayes factor of 6.43. This hypothesis predicted a main effect of gender, that is, males should have a lower probability of violating stochastic dominance than females regardless of their education level. For hypotheses $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$, the data suggest strong evidence for the the informed hypothesis compared to the encompassing hypothesis, with Bayes factors of 17.82 and 22.36, respectively. Thus, based on these data, people with lower education levels are more likely to violate stochastic dominance ($\mathcal{H}_{r2}$), and that the factors gender and education level interact with each other ($\mathcal{H}_{r3}$). The data provide strong evidence for both hypotheses. The ordinal constraint predicted by $\mathcal{H}_{r3}$ also becomes apparent, when we plot the posterior estimates.

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



To compare all four hypotheses directly with each other, we computed the posterior model probabilities. The model which predicts only a gender effect performs worse than the baseline model without any restrictions. Hypothesis $\mathcal{H}_{r3}$ outperforms all other models, including the bookend hypotheses, with a posterior model probability of 88 %. These results are in line with the conclusions drawn by Myung et al. (2005) and Birnbaum (1999), that is, that taken into account the complexity of the model, hypothesis $\mathcal{H}_{r3}$ performs the best. That is, there is a combined effect of gender and education with respect to the probability to violate stochastic dominance. With regard to hypothesis $\mathcal{H}_{r1}$, we can conclude that the gender effect only becomes apparent when taking into account the level of education.

```
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)))
```

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 to violate stochastic domaniance.

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mid \mathbf{x})$	$\mathrm{BF}_{.e}$
$\mathcal{H}_e$	0.25	0.0242	1
$\mathcal{H}_0$	0.25	1.34×10^{-53}	5.55×10^{-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

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

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

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

Furthermore, to make the method accessible to a larger audience of users and students, **multibridge** will be made available in future versions of the software package JASP (JASP Team, 2022). JASP offers an intuitive graphical user interface and does not require extensive knowledge in programming. A first prototype of the **multibridge** module can be seen in Figure 9:

In addition, we plan to expand the types of hypotheses that can be evaluated in future versions of this package, Currently, **multibridge** only supports informed hypotheses which are "stick-hypotheses", that is, hypotheses in which all parameters shared common

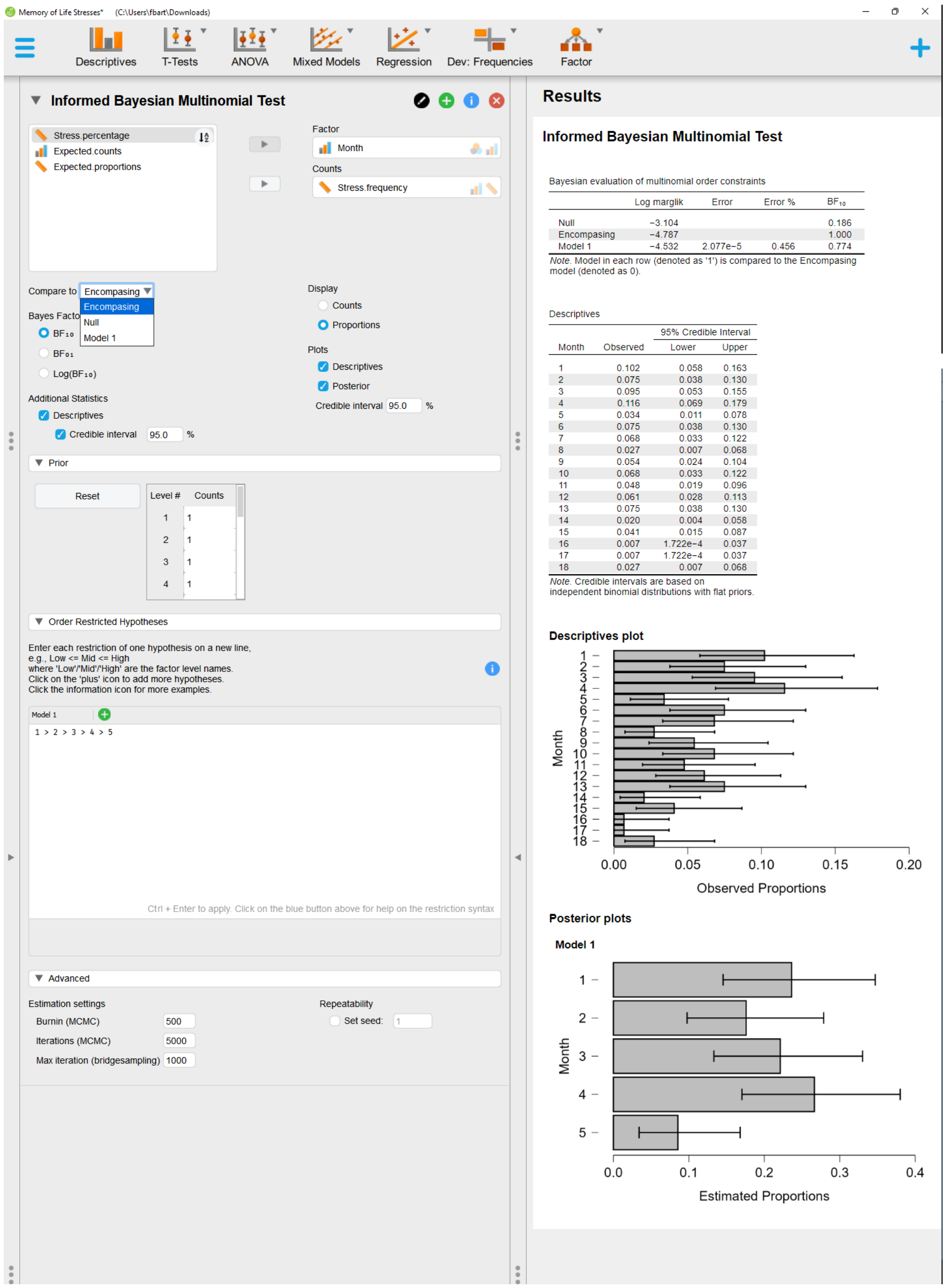


Figure 9. A first prototype of the implementation of the **multibridge** module in JASP.

upper and lower bounds. 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 constrains 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, 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 combines equality-constrained categories but tracks their initial number to effectively sample from the constrained parameter space and when transforming 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.

Declarations

Availability of data and code

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

Funding

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

Author contributions

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

Conflicts of interest

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

Ethical Approval

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.

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

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

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

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

European Commision. (2010). *Report on Greek government deficit and debt statistics* [Eurostat Report]. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/COM_2010_REPORT_GREEK.

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

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*, 389–402.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., & Hothorn, F. S. T. (2020). *Mvtnorm: Multivariate normal and t distributions.* Retrieved from http://CRAN.R-project.org/package=mvtnorm

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Gronau, Q. F., Singmann, H., & Wagenmakers, E. –J. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*, 1–29.

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*, 1526–1553.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511–527.

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261.

Haaf, J. M., Klaassen, F., & Rouder, J. (2019). Capturing ordinal theoretical constraint in psychological science. *PsyArXiv.* Retrieved from https://doi.org/10.31234/osf.io/a4xu9

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87.

Heck, D. W., & Wagenmakers, E.-J. (2016). Adjusted priors for Bayes factors involving reparameterized order constraints. *Journal of Mathematical Psychology*, *73*, 110–116.

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

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

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

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*, 267–285.

JASP Team. (2022). *JASP (Version 0.16.3.0) [Computer software]*. https://jasp-stats.org/.

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

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

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

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*, 281–299.

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

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

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

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

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

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

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

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

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

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

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

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

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

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126.

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

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

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahnik, S., … Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.

Transforming an Ordered Probability Vector to the Real Line

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

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

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

The inverse transformation is given by

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

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

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

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

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

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

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

The transformations outlined above are suitable only for ordered probability vectors, that is, for informed hypotheses in binomial and multinomial models that only feature inequality constraints. However, when informed hypotheses also feature equality constrained parameters, as well as parameters that are free to vary we need to modify the formula. Specifically, to determine the lower bounds for any θ_k, we need to take into account how many parameters were set equal to it (denoted as e_k) and how many parameters were set equal to its preceding value θ_{k-1} (denoted as e_{k-1}):

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

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

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

The elements in the remaining stick are then computed as follows

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

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

Equality Constrained Parameters

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

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

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

When transforming the samples from the collapsed distributions, we need to account for the fact that the inequality constraints imposed under the original parameter values might not hold for the collapsed parameters. Consider, for instance, a multinomial model in which we specify the following informed hypothesis

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

where samples from the encompassing distribution take the values $(0.05, 0.15, 0.15, 0.15, 0.23, 0.27)$. For these parameter values the inequality constraints hold since 0.05 is smaller than 0.15, 0.23, and 0.27. However, the same constraint does not hold when we collapse the categories θ_2, θ_3, and θ_4 into θ_*. That is, the collapsed parameter $\theta_* = 0.15 + 0.15 + 0.15 = 0.45$ is now larger than 0.23 and 0.27. In general, to determine the lower bound for a given parameter θ_k we thus need to take into account both the number of collapsed categories in the preceding parameter e_{k-1} as well as the number of collapsed categories in the current parameter e_k. Thus, lower bounds for the parameters need to be adjusted as follows:

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

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

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

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

Corrections for Free Parameters

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

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

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

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

where f_k represents the number of free parameters that share common bounds with θ_k and that have been not yet been transformed. Here, the number of elements in the remaining stick is defined as the number of all parameters that are larger than θ_k: $ERS = 1 + \sum_{j>k} f_j$. To illustrate this correction, assume that samples from the encompassing distribution take the values $(0.15, 0.29, 0.2, 0.36)$. The upper bound for θ_1 is simply $\frac{1}{4}$. For θ_2, we need to take into account that θ_2 and θ_3 share common bounds. To compute the upper bound for θ_2, we subtract from the length of the remaining stick the lower bound of θ_3: $\dfrac{1 - 0.15 - (1 \times 0.15)}{1+1} = 0.35$.

A further correction is required if a preceding free parameter (i.e., a parameter with common bounds that was transformed already) is larger than the upper bound of the current parameter. For instance, in our example the upper bound for θ_3 would be $\dfrac{1 - 0.44 - 0}{1+1} = 0.28$, which is smaller than the value of the preceding free parameter, which was 0.29. If in this case θ_3 would actually take on the value close to its upper bound, for instance $\theta_3 = 0.275$, then—due to the sum-to-one constraint—θ_4 would violate the constraint (i.e., $0.15 < (0.29\,,\, 0.275) \nless 0.285$). In these cases, the upper bound for the current θ_k needs to be corrected downwards. To do this, we subtract from the current upper bound the difference to the largest preceding free parameter. Thus, if $u_k < \max(\theta_{i<k})$, the upper bound becomes:

$$
\begin{aligned}
u_k &= u_k - (\max(\theta_{i<k}) - u_k) && \text{(C6)} \\
&= 2 \times u_k - \max(\theta_{i<k}). && \text{(C7)}
\end{aligned}
$$

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