multibridge: An R Package To Evaluate Multinomial Order Constraints

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

[1] University of Amsterdam

9   Abstract

10   The **multibridge** package has been developed to efficiently compute Bayes factors for

11   binomial and multinomial models, that feature inequality constraints, equality constraints,

12   free parameters and mixtures between them. By using the bridge sampling algorithm to

13   compute the Bayes factor, **multibridge** facilitates the evaluation of large models with many

14   constraints and models with small parameter spaces. The package was developed in the R

15   programming language and is freely available from the Comprehensive R Archive Network

16   (CRAN). We illustrate the functions based on two empirical examples.

₁₇   multibridge: An R Package To Evaluate Multinomial Order Constraints

₁₈   **Introduction**

₁₉   We present **multibridge**, an `R` package to evaluate informed hypotheses in

₂₀ multinomial models and models featuring independent binomials using Bayesian inference.

₂₁ This package allows users to specify constraints on the underlying category proportions

₂₂ including inequality constraints, equality constraints, free parameters and mixtures between

₂₃ them. This package is available from the Comprehensive `R` Archive Network (CRAN) at

₂₄ https://CRAN.R-project.org/package=multibridge. Here we introduce the methodology

₂₅ used to evaluate informed hypotheses on categorical variables and show how to use the

₂₆ implementations provided in **multibridge** through fully reproducible examples.

₂₇   The most common way to analyze categorical variables is to test whether the

₂₈ underlying category proportions are exactly equal or whether they are fixed and follow a

₂₉ predicted pattern (what is generally known as either chi-square goodness of fit tests, or

₃₀ binomial or multinomial tests). Often, however, this analytic approach cannot or only

₃₁ insufficiently test the specific predictions derived by scientific theories researchers are

₃₂ interested in. This is the case, for instance, when theories predict ordinal relations among

₃₃ the underlying category proportions, such as increasing or decreasing trends. For instance, to

₃₄ check for irregularities in audit data, one could test whether the leading digits in the data

₃₅ are distributed according to an expected Benford distribution or whether they deviate from

₃₆ it, for example, by showing a general decreasing trend. Here, the Benford distribution can be

₃₇ tested with standard methods, however, the general decreasing trend cannot be tested, since

₃₈ we cannot derive fixed underlying proportions for the leading digits. Theories can also

₃₉ generate more complex predictions, including ones that feature combinations of equality and

₄₀ inequality constraints, as well as predictions that let some category proportions free to vary.

₄₁ In the following, we will denote such predictions as informed hypotheses. Such an informed

hypothesis was expressed, for instance, by Nuijten, Hartgerink, Assen, Epskamp, and Wicherts (2016) who studied the prevalence of statistical reporting errors in articles published in different areas of psychological science. Nuijten et al. (2016) hypothesized that articles published in social psychology journals would have higher error rates than articles published in other psychological journals while not expressing expectations about the error rate distribution among the other journals. Here again no fixed proportions can be derived, such that this hypothesis cannot not be evaluated using standard tests. Generally, if researchers can utilize statistical methods for testing informed hypotheses, they are able to specify hypotheses that relate more closely to their theories.

In the Bayesian framework, researchers can compare models that instantiate the hypotheses of interest by means of Bayes factors (Jeffreys, 1935; Kass & Raftery, 1995). Bayes factors compare the relative evidence of two hypotheses in the light of the data. It is defined as the ratio of marginal likelihoods of the respective hypotheses. For instance, the Bayes factor for the informed hypothesis versus a hypothesis that lets all parameters free to vary is defined as:

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

where the subscript r denotes the informed hypothesis and e denotes a hypothesis which predicts that all parameters free to vary. For the Bayesian evaluation of informed hypotheses several R packages are available. For instance, with the package **multinomineq** (Heck & Davis-Stober, 2019) users can specify inequality constrained hypotheses but also more general linear inequality constraints for multinomial models as well as models that feature independent binomials. The **BAIN** package (Gu, Hoijtink, Mulder, & Rosseel, 2019)

63 allows for the evaluation of inequality constraints in structural equation models. Outside of

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

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

66 measures, and multivariate regression. All these packages rely on one of two methods to

67 approximate order constrained Bayes factors: the encompassing prior approach (Gu, Mulder,

68 Deković, & Hoijtink, 2014; Hoijtink, 2011; Hoijtink, Klugkist, & Boelen, 2008; Klugkist,

69 Kato, & Hoijtink, 2005) and the conditioning method (Mulder, 2014, 2016; Mulder et al.,

70 2009). However, even though these methods are currently widely used, they are known to

71 become increasingly unreliable and inefficient as the number of constraints increases or when

72 the parameter space of the constrained model is small (Sarafoglou et al., 2020).

73   In contrast to these available packages, **multibridge** uses a bridge sampling routine

74 that enables users to compute Bayes factors for informed hypotheses more reliably and

75 efficiently (Bennett, 1976; Meng & Wong, 1996; Sarafoglou et al., 2020). The workhorse for

76 this analysis, the bridge sampling algorithm, constitutes a special case of the algorithm

77 implemented in the R package **bridgesampling** (Gronau, Singmann, & Wagenmakers,

78 2020). The **bridgesampling** package, allows users to estimate the marginal likelihood for a

79 wide variety of models, including models implemented in Stan (Stan Development Team,

80 2020). However, the algorithm implemented in **bridgesampling** is not suitable for models

81 that include constraints on probability vectors and hence is unsuitable for the analysis of

82 categorical data. Therefore, in **multibridge**, we tailored the bridge sampling algorithm such

83 that it accommodates the specification of informed hypotheses on probability vectors. The

84 package then produces an estimate for the Bayes factor in favor of or against the informed

85 hypothesis. The resulting Bayes factor compares the evidence for the informed hypotheses to

86 the encompassing hypothesis that imposes no constraints on the underlying category

87 proportions. Alternatively, the informed hypothesis can be tested against the null hypothesis

88 that all underlying category proportions are exactly equal. Given this result, users can then

89 either receive a visualization of the posterior parameter estimates under the encompassing

90 hypothesis using the `plot`-method, or get more detailed information on how the Bayes factor

91 is composed using the `summary`-method. For hypotheses that include mixtures between

92 equality and inequality informed hypotheses the `bayes_factor` method shows the

93 conditional Bayes factor for the inequality constraints given the equality constraints and a

94 Bayes factor for the equality constraints. The general workflow of **multibridge** is illustrated

95 in Figure 1. Table 1 summarizes all S3 methods currently available in **multibridge**.
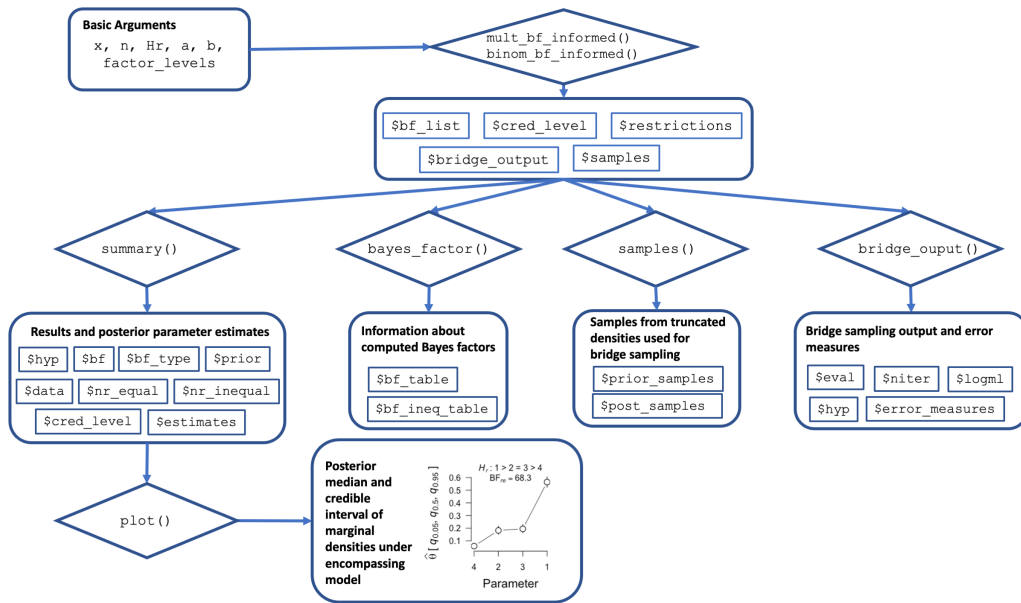

Figure 1. The **multibridge** workflow. The user specifies the data values (`x` and `n` for binomial
models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and
β parameters of the Binomial prior distributions (`a` and `b`) or the concentration parameters
for the Dirichlet prior distribution (`a`), respectively, and the factor levels (`factor_levels`).
The functions `mult_bf_informed` and `binom_bf_informed` then produce an estimate for
the Bayes factor of the informed hypothesis versus the encompassing or the null hypothesis.
Based on these results different S3 methods can be used to get more detailed information
on the individual components of the analysis (e.g., `summary`, `bayes_factor`), and parameter
estimates of the encompassing distribution (`plot`).

96   The remainder of this article is organized as follows: In the methods section, we

describe the Bayes factor identity for informed hypotheses in binomial and multinomial models, and present the bridge sampling routine implemented in the **multibridge** package including details of the necessary transformations required for this routine. In Section 3, we will schematically introduce the most relevant functions in **multibridge** and their arguments. Section 4 illustrates how to use the **multibridge** package to estimate parameters, and compute Bayes factors using two examples.

Methods

In this section we formalize multinomial models and models that feature independent binomial probabilities. In the multinomial model, we assume that the vector of observations $x_1, \cdots, x_K$ in the K categories follow a multinomial distribution. The parameter vector of the multinomial model, $\theta_1, \cdots, \theta_K$, contains the probabilities of observing a value in a particular category. In **multibridge** we assume that the parameter vector $\theta_1, \cdots, \theta_K$ is drawn from a Dirichlet distribution with concentration parameters $\alpha_1, \cdots, \alpha_K$:

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

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

In the binomial model, we assume that the elements in the vector of successes $x_1, \cdots, x_K$ and the elements in the vector of total number of observations $n_1, \cdots, n_K$ in the K categories follow independent binomial distributions. As in the multinomial model, the parameter vector of the binomial success probabilities, $\theta_1, \cdots, \theta_K$, contains the probabilities of observing a value in a particular category. In **multibridge** we assume that the parameter values $\theta_1, \cdots, \theta_K$ are drawn from independent beta distributions with parameters $\alpha_1, \cdots, \alpha_K$ and $\beta_1, \cdots, \beta_K$:

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

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

Bayes factor

When evaluating informed hypotheses that feature mixtures between inequality and equality constraints it is important to realize that the Bayes factor, further denoted as BF_{re}, factors as follows:

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

where the subscript $<$ denotes the hypothesis that only features inequality constraints and the subscript $=$ denotes the hypothesis that only features equality constraints. A Bayes factor for mixtures thus factors into a Bayes factor for the equality constraints, $\text{BF}_{=e}$, and a conditional Bayes factor for the inequality constraints given the equality constraints $\text{BF}_{<e} \mid \text{BF}_{=e}$ (for the proof, see Sarafoglou et al., 2020).

The Bayes Factor For Equality Constraints

For binomial models, the (marginal) Bayes factor for a null hypothesis versus a hypothesis in which all parameters are free to vary can be computed analytically with the function `binom_bf_equality`. Assuming that the first i binomial probabilities in a model are equality constrained, the Bayes factor BF_{0e} is defined as:

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

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

128 $n_+ = \sum_{i<k} n_i$. The latter factor introduces a correction for marginalizing which stems from

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

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

131 parameters in the corresponding Binomial distribution.

For multinomial models, the (marginal) Bayes factor for for a null hypothesis versus a

hypothesis in which all parameters are free to vary can be computed analytically with the

function `multBayes_bf_equality`. Assuming again that the first i category probabilities in

a model are equality constraint, the Bayes factor BF_{0e} is defined as:

$$\mathrm{BF}_{e0} = \frac{\mathrm{B}(\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha} + \mathbf{x})} \left(\frac{1}{i}\right)^{\sum_{i<k} x_i} \frac{\mathrm{B}\left(\sum_{i<k} \alpha_i + x_i - i + 1, \, \alpha_k + x_k, \, \ldots, \, \alpha_K + x_K\right)}{\mathrm{B}\left(\sum_{i<k} \alpha_i - i + 1, \, \alpha_k, \, \ldots, \, \alpha_K\right)},$$

132 **The Bayes Factor For Inequality Constraints**

133 For informed hypotheses, Klugkist et al. (2005) derived the following identity of the

134 Bayes factor BF_{re}:

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

135 Recently Sarafoglou et al. (2020) showed that the Bayes factor BF_{re} can also be

136 interpreted as ratio of two marginal likelihoods:

$$\text{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}^{\substack{\text{Marginal likelihood of} \\ \text{posterior distribution}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Marginal likelihood of} \\ \text{prior distribution}}}}. \tag{6}$$

In this identity, $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$ denotes the marginal likelihood of the constrained posterior distribution and $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ denotes the marginal likelihood of the constrained prior distribution. Even though both identities are mathematically equivalent, the methods to estimate these identities are very different. In the first case, for instance, the number of samples from the encompassing distribution in accordance with the inequality constrained hypothesis, serve as an estimate for the proportion of prior parameter space consistent with the restriction. On the flip side, however, this means that the accuracy of this estimate is strongly dependent on the number of the constrained parameters in the model and the size of the constrained parameter space. That is, as the constraints become stronger, the constrained parameter space decreases. As a result it becomes less likely that draws from the encompassing distribution will fall into the constrained region, so that in some cases the estimation of the Bayes factor becomes practically impossible (Sarafoglou et al., 2020).

However, when we interpret the Bayes factor BF_{re} as ratio of marginal likelihoods and we are able to sample from the constrained prior and posterior distributions, we can utilize numerical sampling methods such as bridge sampling to obtain the estimates. Crucially, in this approach, 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 relatively high number of categories (i.e., $K > 10$) are evaluated and (2) when relatively little posterior mass falls in the constrained parameter space.

The Bridge Sampling Method

Bridge sampling is a method to estimate the ratio of two marginal likelihoods which yield the Bayes factor (Bennett, 1976; Meng & Wong, 1996). In the **multibridge** package we implemented a version of bridge sampling that estimates one marginal likelihood at the time since it increases the accuracy of the method without considerably increasing its computational efficiency (Overstall & Forster, 2010). Specifically, we subsequently estimate the marginal likelihood for the constrained prior distribution and the marginal likelihood of the constrained posterior distribution.

When applying this modified version of the bridge sampling method, we estimate a marginal likelihood by means of a so-called proposal distribution. In **multibridge** this proposal distribution is the multivariate normal distribution. To estimate the marginal likelihood, bridge sampling only requires samples from the distribution of interest—the so-called target distribution—and samples from the proposal distribution. In **multibridge**, the samples from the target distribution—that is the constrained prior and posterior Dirichlet distribution for multinomial models and constrained prior and posterior beta distributions for binomial models—are drawn through 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, however, we apply an algorithm that simulates values from constrained Gamma distributions which are then transformed into Dirichlet random variables (for details, see Appendix C in Sarafoglou et al. (2020)). To sample efficiently from these distributions, **multibridge** uses a `C++` routine for this algorithm.

The efficiency of the bridge sampling method is guaranteed only if the target and proposal distribution (1) operate on the same parameter space and (2) have sufficient overlap. To meet these requirements, **multibridge** applies the appropriate probit

183 transformations on the samples of the constrained distributions to move the samples from

184 the probability space to the entire real line. Details on these transformations are provided in

185 the appendix. To ensure sufficient overlap, half of the draws are then used to construct the

186 proposal distribution using the method of moments. Samples from the proposal distribution

187 can be generated using the standard `rmvnorm`-function from the R package **stats**. For the

188 marginal likelihood of the constrained prior distribution, the modified bridge sampling

189 identity is then defined as:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\text{prior}}\left(g(\boldsymbol{\theta})h(\boldsymbol{\theta})\right)}, \tag{7}$$

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

191 minimized the relative mean square error of the estimate and $g(\boldsymbol{\theta})$ refers to the proposal

192 distribution. The numerator evaluates the unnormalized density for the constrained prior

193 distribution with samples from the proposal distribution. The denominator evaluates the

194 normalized proposal distribution with samples from the constrained prior distribution. The

195 expression for the marginal likelihood for the constrained posterior distribution can be

196 described in a similar way. As final step, we apply the iterative scheme proposed by Meng

197 and Wong (1996) to receive the bridge sampling estimator:

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

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

199 $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$, N_2 denotes the number of samples drawn from the proposal distribution, that

200 is $\tilde{\boldsymbol{\theta}} \sim g(\boldsymbol{\theta})$, $s_1 = \frac{N_1}{N_2+N_1}$, and $s_2 = \frac{N_2}{N_2+N_1}$. The quantities $\ell_{1,n}$ and $\ell_{2,m}$ are defined as follows:

$$\ell_{1,n} = \frac{q_{1,1}}{q_{1,2}} = \frac{p(\boldsymbol{\theta_n^*} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta_n^*} \in \mathcal{R}_r)}{g(\boldsymbol{\xi_n}^*)}, \tag{8}$$

$$\ell_{2,m} = \frac{q_{2,1}}{q_{2,2}} = \frac{p(\tilde{\boldsymbol{\theta}}_m \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_m \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\xi}}_m)}, \tag{9}$$

where $\boldsymbol{\xi_n}^* = \Phi^{-1}\left(\dfrac{\boldsymbol{\theta_n^*} - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_m = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_m) + \mathbf{l}) \, |J|)$. The quantity $q_{1,1}$ refers to the

evaluations of the constrained distribution for constrained samples and $q_{1,2}$ refers to the

proposal evaluations for constrained samples, respectively. The quantities $q_{2,1}$ refers to

evaluations of the constrained distribution for samples from the proposal and $q_{2,2}$ refers to

the proposal evaluations for samples from the proposal, respectively. Note that the quantities

$\ell_{1,n}$ and $\ell_{2,m}$ have been adjusted to account for the necessary parameter transformations to

create overlap between the constrained distributions and the proposal distribution.

multibridge runs the iterative scheme until the tolerance criterion suggested by Gronau et

al. (2017) is reached, that is, $\dfrac{\mid \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} - \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)} \mid}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)}} \leq 10^{-10}.$

The bridge sampling estimate for the log marginal likelihood of the constrained

distribution and its associate relative mean square error, the number of iterations, and the

quantities $q_{1,2}$, $q_{1,2}$, $q_{1,2}$, and $q_{1,2}$ are included in the standard output in **multibridge**. The

function to compute the relative mean square error was taken from the R package

bridgesampling.

> Is this important enough to mention it here?

Usage and Examples

The **multibridge** package can be installed from the Comprehensive R Archive

Network (CRAN) at https://CRAN.R-project.org/package=multibridge:

```
install.packages('multibridge')

library('multibridge')
```

²¹⁹ A list of all currently available functions and datasets is given in Table 3. Additional

²²⁰ examples are available as vignettes (see https://cran.r-project.org/package=multibridge, or

²²¹ `vignette(package = "multibridge")`). The two core functions of **multibridge**—the

²²² `mult_bf_informed`-function and the `binom_bf_informed`-function—can be illustrated

²²³ schematically as follows:

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

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

²²⁴ The basic required arguments for these functions are listed in Table 2. In the following,

²²⁵ we will outline two examples on how to use **multibridge** to compare an informed hypothesis

²²⁶ to a null or encompassing hypothesis. In addition, the first example shows how two informed

²²⁷ hypotheses can be compared to each other.

Example 1: Applying A Benford Test to Greek Fiscal Data

²²⁹ The first digit phenomenon, otherwise known as Benford's law (Benford, 1938;

²³⁰ Newcomb, 1881) states that the expected proportion of leading digits in empirical data can

²³¹ be formalized as follows: for any given leading digit $d, d = (1, \cdots, 9)$ the expected proportion

²³² is approximately equal to

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

²³³ This means that a number in a empirical dataset has leading digit 1 in 30.1% of the cases,

²³⁴ and leading digit 2 in 17.61% of the cases; leading digit 9 is the least frequent digit with an

²³⁵ expected proportion of only 4.58% (see Table 4 for an overview of the expected proportions).

²³⁶ Empirical data includes, for instance, data on population sizes, death rates, baseball

statistics, atomic weights of elements, and physical constants (Benford, 1938). In contrast,

generated data, such as telephone numbers, do in general not obey Benford's law (Hill, 1995).

Since Benford's law proved to be highly suitable to discriminate between empirical data and

generated data, a so-called Benford test can be used to check whether certain observed

frequencies of first digits, obey Benford's law and therefore can be considered an empirical

dataset. Benford's tests are used in fields like accounting and auditing to check for

indications for poor data quality, for instance, in fiscal statements [for an overview,]see e.g.,

Durtschi, Hillison, and Pacini (2004), Nigrini and Mittermaier (1997), Nigrini (2012)]. Data

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

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

In the following, we discuss three possible Bayesian adaptations of the Benford's test.

In a first scenario we simply conduct Bayesian multinomial test in which we test the

point-null hypothesis $\mathcal{H}_0$ which predicts a Benford distribution against the encompassing

hypothesis $\mathcal{H}_e$ which leaves all model parameters free to vary. Testing against the

encompassing hypothesis is considered standard practice, yet, it leads to an unfair

comparison to the detriment of the null hypothesis. In general, if we are dealing with a

high-dimensional parameter space and the competing hypotheses differ largely in their

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

the data follow the predicted trend of the more complex hypothesis considerably well. In a

second scenario we therefore test the null hypothesis against an alternative hypothesis,

denoted as $\mathcal{H}_{r1}$, which predicts a decreasing trend in the proportions of leading digits. The

hypothesis $\mathcal{H}_{r1}$ implies considerably more constraints than $\mathcal{H}_e$ and is a suitable choice if our

primary goal is to distinguish whether data comply with Benford's law or whether the data

only follow a similar trend. In a third scenario, where the main goal is to identify fabricated

data, we could test the null hypothesis against an hypothesis, which predicts a trend that is

characteristic for manipulated data. This hypothesis, which we denote as $\mathcal{H}_{r2}$, could be

derived from empirical research on fraud or be based on observed patterns from former fraud

264   cases. For instance, Hill (1988) instructed students to produce a series of random numbers;

265   in the resulting data the proportion of the leading digit 1 occurred most often and the digits

266   8 and 9 occurred least often which is consistent with the general pattern of Benford's law.

267   However, the proportion for the remaining leading digits were approximately equal. We do

268   want to note that the predicted distribution derived from Hill (1988) is not currently used as

269   a test to detect fraud. However, for the sake of simplicity, if we assume that this pattern

270   could be an indication for fabricated auditing data, the Bayes factor could quantify the

271   evidence of whether the proportion of first digits resemble authentic or fabricated data.

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

273   factors were originally published by the European statistics agency "Eurostat" and served as

274   basis for reviewing the adherence to the Stability and Growth Pact of EU member states.

275   Rauch, Göttsche, Brähler, and Engel (2011) conducted a Benford test on data related to

276   budget deficit criteria, that is, public deficit, public dept and gross national products. This

277   data used for this example contains fiscal data from Greece related in the years between 1999

278   and 2010; a total of $N = 1,497$ numerical data were included in the analysis. We choose this

279   data, since the Greek government deficit and debt statistics states has been repeatedly

280   criticized by the European Commission in this timespan (European Commision, 2004, 2010).

281   In particular, the commission has accused the Greek statistical authorities to have

282   misreported deficit and debt statistics. For further details on the dataset see Rauch et al.

283   (2011). The observed proportions are displayed in Table 4, the figure displaying the observed

284   versus the expected proportions are displayed in Figure 2.

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

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

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

288   proportions of first digits obeys Benford's law:

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

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

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

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

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

In cases, in which we are interested in computing two informed hypotheses with each other, we need to make use of the transitivity property of the Bayes factor. For instance, if we would like to compare the two informed hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ with each other, we would first compute BF_{er1} and BF_{er2} and then yield BF_{r1r2} as follows:

$$\mathrm{BF}_{r1e} \times \mathrm{BF}_{er2} = \mathrm{BF}_{r1r2}.$$

Method. We can compare $\mathcal{H}_0$ and $\mathcal{H}_e$ by means of a Bayesian multinomial test, that is, we stipulate equality constraints on the entire parameter vector $\boldsymbol{\theta}$. The corresponding Bayes factor is thus computationally straightforward; we can calculate BF_{0e} by applying the function `mult_bf_equality`. To evaluate $\mathcal{H}_0$, we only need to specify (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution, and (3) the vector of predicted proportions. Since we have no specific expectations about the distribution of leading digits in the Greek fiscal data, we set all concentration parameters to one which corresponds to a uniform Dirichlet distribution.

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

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

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

```
# Labels for categories of interest
factor_levels <- 1:9
# Specifying the informed Hypothesis
Hr1 <- c('1 > 2 > 3 > 4 > 5 > 6 > 7 > 8 > 9')
Hr2 <- c('1 > 2 = 3 = 4 = 5 = 6 = 7 > 8 > 9')
# Execute the analysis
results_He_Hr1 <- mult_bf_informed(x = x, Hr = Hr1, a = a,
                                   factor_levels = factor_levels,
                                   seed = 2020)
results_He_Hr2 <- mult_bf_informed(x = x, Hr = Hr2, a = a,
                                   factor_levels = factor_levels,
                                   seed = 2020)
```

```
logbf <- c(log(results_H0_He$bf$BF0e),
           log(results_H0_He$bf$BF0e * results_He_Hr1$bf_list$bf$BFer),
           log(results_H0_He$bf$BF0e * results_He_Hr2$bf_list$bf$BFer))
bayes_factor_table <- data.frame(
   BFType = c('LogBF0e', 'LogBF0r1', 'LogBF0r2'),
   LogBF  = logbf)
```

```
bayes_factor_table
```

```
## 307     BFType     LogBF
## 308 1  LogBF0e -17.67150
## 309 2 LogBF0r1 -25.08832
## 310 3 LogBF0r2 154.57114
```

As the evidence is extreme in all three cases, we report all Bayes factors on the log scale. The log Bayes factor $\log(\mathrm{BF}_{0e})$ suggests extreme evidence against the hypothesis that the first digits in the Greek fiscal data follow a Benford's distribution; $\log(\mathrm{BF}_{0e})$ = -17.67. The log Bayes factor $\log(\mathrm{BF}_{0r1})$ indicates extreme evidence in favor for a decreasing trend, $\log(\mathrm{BF}_{0r1})$ = -25.09. Even though the Bayes factor suggests extreme evidence against the hypothesis that the Greek fiscal data are an empirical dataset, there is no support the hypothesis that the data are fabricated. The log Bayes factor $\log(\mathrm{BF}_{0r1})$ indicates extreme evidence against $\mathcal{H}_{r2}$ with $\log(\mathrm{BF}_{0r2})$ = 154.57. The proportions of leading digits is best characterized by a monotonously decreasing trend, compared to all parameters varying freely $(\log(\mathrm{BF}_{r1e})$ = 7.42), and compared to a distribution that one could expect from fabricated data $(\log(\mathrm{BF}_{r1r2})$ = 180).

Discussion.   In this example we tested the data quality of Greek fiscal data in the years 1999 to 2009 by conducting three variations of a Bayesian Benford test. More precisely, we evaluated the null hypothesis thatß Greek fiscal data conform to Benfords law. We tested this hypothesis against three alternatives. The first alternative hypothesis, $\mathcal{H}_e$ relaxed the constraints imposed by the null hypothesis and left all model parameters free to vary. The second alternative hypothesis, $\mathcal{H}_{r1}$ predicted a decreasing trend in the proportion of leading digits. The third alternative hypothesis $\mathcal{H}_{r2}$ predicted a trend that Hill (1988) observed when humans tried to generate random numbers. Our result suggest that the leading digits in the fiscal statistics do not follow a Benford distribution; in fact, we collected extreme
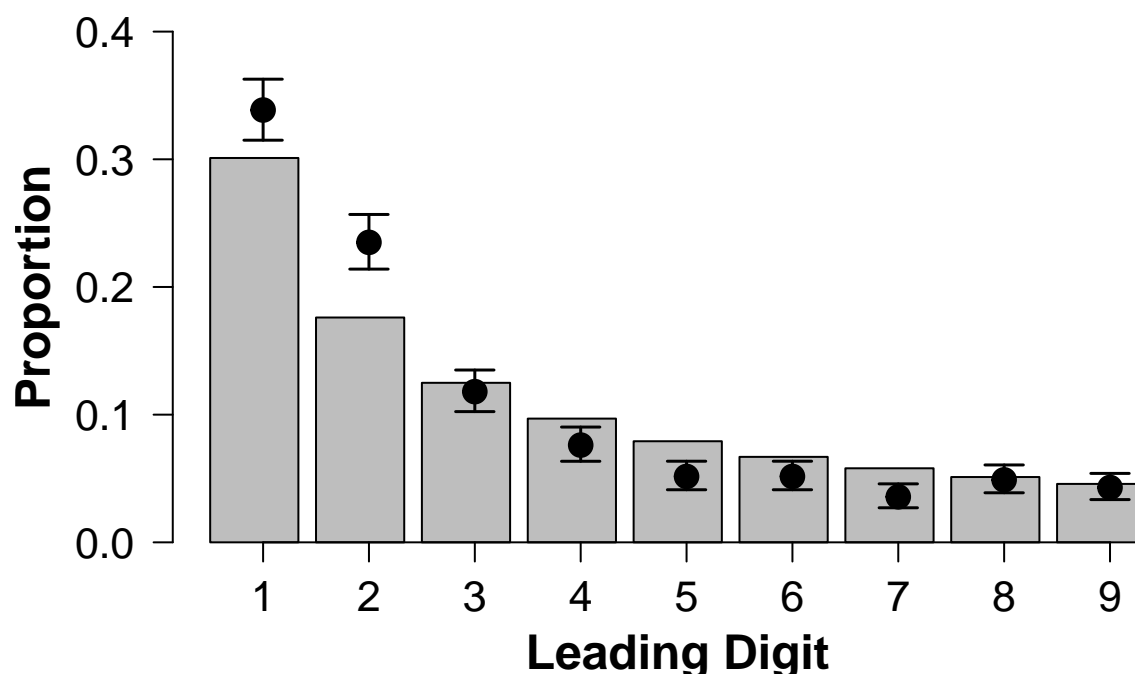
Figure 2. The bargraph displays the expected proportions of leading digits according to Benford's law. The black dots indicate for the actual fiscal statistics from Greece the posterior estimates for the proportion of leading digits and the corresponding 95% credible intervals based on the encompassing model. Only three out of nine estimates cover the expected proportions.

331 evidence against Benford's law compared to two out of three of the alternative hypotheses.

332 When comparing the alternative hypotheses directly to each other, the data show most

333 evidence in favor for a decreasing trend. A Benford test of fiscal statements can be a helpful

334 tool to detect poor data quality and suspicious numbers. In follow-up checks of these

335 numbers, it could then be examined for instance, whether financial statements were actually

336 materially misstated, for isntance, by rounding up or down numbers, avoiding certain

337 thresholds etc., Nigrini (2019).

Example 2: Prevalence of Statistical Reporting Errors

In any scientific article that uses null hypothesis significance testing, there is a chance that the reported test statistic and degrees of freedom do not match the reported p-value. In most cases this is because researchers copy the relevant test statistics by hand into their articles and there are no automatic checks to detect these mistakes. Therefore, Epskamp and Nuijten (2014) developed the R package `statcheck`, which only requires the PDF of a given scientific article to detect these reporting errors automatically and efficiently. This package allowed Nuijten et al. (2016) to get an overview about the prevalence of statistical reporting errors in the field of psychology. In total, the authors investigated a sample of 30,717 articles (which translates to over a quarter of a million p-values) published in eight major psychological journals between 1985 to 2013: *Developmental Psychology* (DP), the *Frontiers in Psychology* (FP), the *Journal of Applied Psychology* (JAP), the *Journal of Consulting and Clinical Psychology* (JCCP), *Journal of Experimental Psychology: General* (JEPG), the *Journal of Personality and Social Psychology* (JPSP), the *Public Library of Science* (PLoS), *Psychological Science* (PS).

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

364 practices (John, Loewenstein, & Prelec, 2012). Specifically, John et al. (2012) found that

365 researchers from the area of social psychology assessed questionable research practices both

366 as more defensible and more applicable for their research compared to other research areas in

367 psychology.

368   **Data and Hypothesis.**   Here, we are reusing the original data published by Nuijten

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

370 `journals`.

```
data(journals)
```

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

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

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

374 (2016) did not make use of inferential statistics since their sample included the entire

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

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

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

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

379 probabilities, $\boldsymbol{\theta}$, reflects the probabilities of a statistical reporting error in one of the 8

380 journals. Thus, we can formalize the discussed hypotheses as follows:

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

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

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

382 observed successes (i.e., number of articles that contain a statistical reporting error), and (2)

383 a vector containing the total number of observations, (3) the informed hypothesis, (4) a

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

385 parameter β_i for each binomial proportion, and (6) the categories of interest (i.e., journal

386 names). Since we have no specific expectations about the distribution of statistical reporting

387 errors across journals, we set all parameters α_i and β_i to one which corresponds to uniform

388 Beta distributions. With this information, we can now conduct the analysis with the

389 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
# We assign a uniform beta distribution to each binomial proportion
a <- rep(1, 8)
b <- rep(1, 8)


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


# Category labels
journal_names <- journals$journal


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

```
                                factor_levels = journal_names,

                                bf_type = 'BF0r', seed = 2020)


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

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

BFr0 <- results_H0_Hr$bf_list$bf['BFr0']
```

390   The Bayes factor BF_{r0} suggests extreme evidence for the informed hypothesis that the

391 social psychology journal JPSP has the highest prevalence for statistical reporting errors

392 compared to the null hypothesis that the statistical reporting errors are equal across journals;

393 $\text{BF}_{r0} = -158$. When taking a closer look at the Bayes factors, we also see that the data

394 suggest that the null hypothesis that the statistical reporting errors are equal across journals

395 is highly unlikely compared to the encompassing hypothesis; with BF_{e0} of 7.38×10^{67}. In

396 addition, the results suggest that the data are 7.43 more likely under the informed

397 hypothesis than under the hypothesis that the ordering of the journals can vary freely.

398   In order to get a clearer picture about the ordering of the journals, we can investigate

399 the posterior estimates under the encompassing model as the next step. The posterior

400 median and 95% credible interval are returned the `summary`-method and can be plotted, too:

401   **Discussion.** In this example, we tested whether the prevalence for statistical

402 reporting errors for articles published in social psychology journals (i.e., JPSP) is higher

403 than for articles published in other journals. We tested this hypothesis against the null

404 hypothesis that the prevalence for statistical reporting errors is equal across all journals. The

405 resulting Bayes factor of $\text{BF}_{r0} = 1.82 \times 10^{-69}$ provides extreme evidence for the informed

406 hypothesis. However, this result should be interpreted with caution and be considered more

407 differentiated. It seems that the result is above all an indication that the null hypothesis is

408 highly misspecified and that the prevalence for a statistical reporting error varies greatly
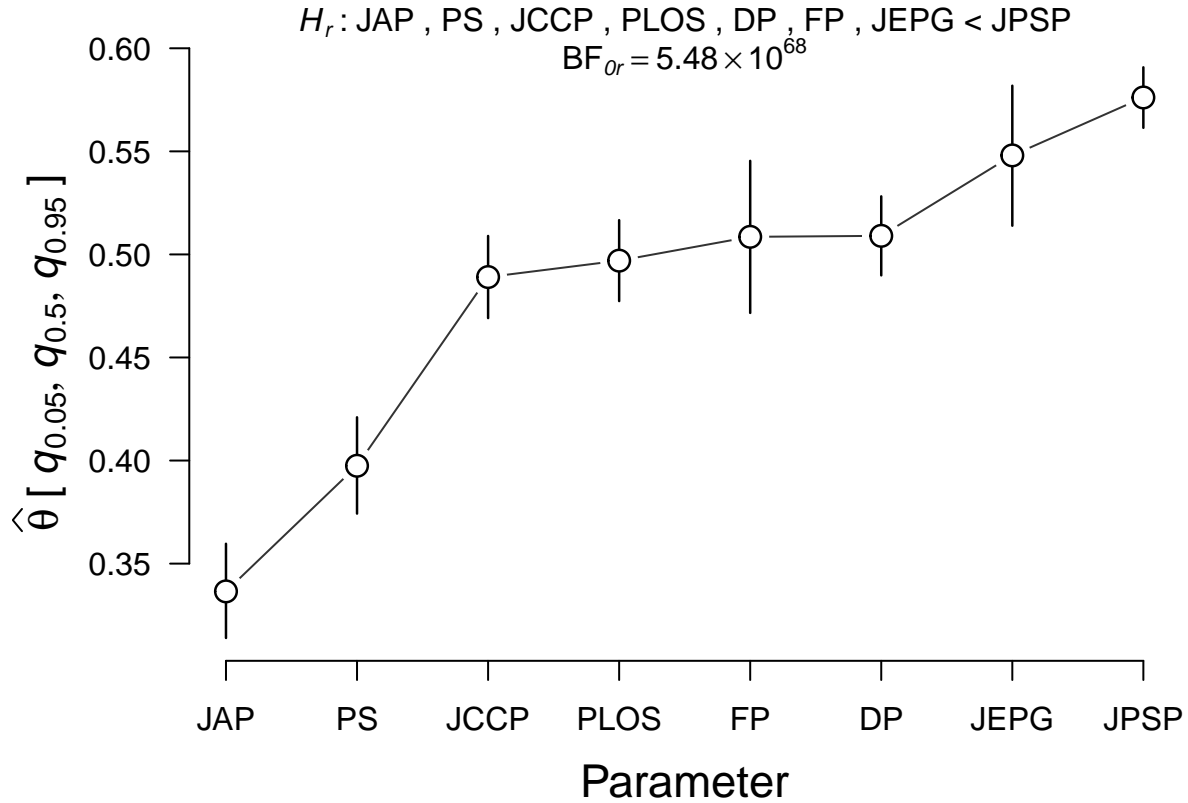
Figure 3. The figure displays for each journal the posterior estimates for the prevalence that an article includes a statistical reporting error and the corresponding 95% credible intervals based on the encompassing model. It appears that all journals show a relatively similar prevalence for statistical reporting errors, with the exception of the *Journal of Applied Psychology* (JAP) and *Psychological Science* (PS), whose prevalence is much lower.

from journal to journal. Evidence that JPSP stands out and has a higher prevalence than the other journals is relatively small; the data provided only moderate evidence against the encompassing hypotheses.

Summary

The R package **multibridge** facilitates the computation of Bayes factors for informed hypotheses in multinomial models. The underlying algorithm is based on a recently developed bridge sampling routine that is more efficient and reliable than available methods.

₄₁₆ **multibridge** can evaluate hypotheses that feature equality constraints, inequality

₄₁₇ constraints, and free parameters as well as mixtures between them. The core functions of the

₄₁₈ software package were illustrated with two empirical examples. The **multibridge** package is

₄₁₉ under continuous development. In the future, we aim to implement methods that extend the

₄₂₀ functionality of the package to hierarchical binomial and multinomial models. In addition,

₄₂₁ we want to enable users to specify order constraints that are more complex, including

₄₂₂ hypotheses on the size ratios of the parameters of interest or the difference between category

₄₂₃ proportions.

₄₂₄
> reiterate benefits over existing packages

Table 1

S3 methods available in **multibridge**

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

Table 2

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

Argument	Description
x	a vector with data (for multinomial models) or a vector of counts of successes, or a two-dimensional table (or matrix) with 2 columns, giving the counts of successes and failures, respectively (for binomial models)
n	numeric. Vector of counts of trials. Must be the same length as x. Ignored if x is a matrix or a table
Hr	string or character. Encodes the user specified informed hypothesis. Users can either use the specified `factor_levels` or indexes to refer to parameters.
a	numeric. Vector with concentration parameters of Dirichlet distribution (for multinomial models) or α parameters for independent beta distributions (for binomial models). Default sets all parameters to 1
b	numeric. Vector with β parameters. Must be the same length as x. Default sets all β parameters to 1
factor_levels	character. Vector with category names. Must be the same length as x

Table 3

*Core functions available in **multibridge***

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

Table 4

The Table shows the Observed Counts, Observed Proportions, and Expected Proportions of first digits in Greece governmental data. The total sample size was N = 1,497 observations. Note that the observed proportions and counts deviate slightly from those reported in Rauch et al. (2011) (probably due to rounding errors).

Leading digit	Observed Counts	Observed Proportions	Expected Proportions: Benford's Law
1	509	0.340	0.301
2	353	0.236	0.176
3	177	0.118	0.125
4	114	0.076	0.097
5	77	0.051	0.079
6	77	0.051	0.067
7	53	0.035	0.058
8	73	0.049	0.051
9	64	0.043	0.046

References

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*, 245–268.

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

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

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

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

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

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.

447   Gronau, Q. F., Singmann, H., & Wagenmakers, E. (2020). Bridgesampling: An R

448   package for estimating normalizing constants. *Journal of Statistical Software,*

449   *Articles, 92*(10), 1–29.

450   Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for bayesian

451   testing of order constrained hypotheses in structural equation models. *Journal of*

452   *Statistical Computation and Simulation, 89,* 1526–1553.

453   Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of

454   inequality constrained hypotheses. *Psychological Methods, 19,* 511–527.

455   Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear

456   inequality constraints: Overview and improvements of computational methods for

457   Bayesian inference. *Journal of Mathematical Psychology, 91,* 70–87.

458   Hill, T. P. (1988). Random-number guessing and the first digit phenomenon.

459   *Psychological Reports, 62,* 967–971.

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

461   *Science,* 354–363.

462   Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and*

463   *social scientists.* Boca Raton, FL: Chapman & Hall/CRC.

464   Hoijtink, H., Klugkist, I., & Boelen, P. (Eds.). (2008). *Bayesian evaluation of*

465   *informative hypotheses.* New York: Springer Verlag.

466   Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability.

467   *Proceedings of the Cambridge Philosophy Society, 31,* 203–222.

468   John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of

469   questionable research practices with incentives for truth telling. *Psychological*

Science, *23*, 524–532.

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

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

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

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

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

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

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

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*, 39–40.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.

Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud schemes. *Managerial Auditing Journal, 34*, 602–622.

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of benford's law as an aid in analytical procedures. *Auditing, 16*, 52.

Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226.

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis, 54*, 3269–3288.

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

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

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

Appendix

Transforming An Ordered Probability Vector To The Real Line

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

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

511 transformation needs to retain the inequality constraints imposed on the parameters, that is,

512 it needs to take into account the lower bound l_k and the upper bound u_k of each θ_k. To

513 achieve this goal, **multibridge** uses a probit transformation which subsequently transforms

514 the elements in $\boldsymbol{\theta}$ moving from its lowest to its highest value. In the binomial model, we

515 move all elements in $\boldsymbol{\theta}$ to the real line and thus construct a new vector $\boldsymbol{y} \in \mathbb{R}^K$. For

516 multinomial models, however, it follows from the unit constraint that the vector $\boldsymbol{\theta}$ is

517 completely determined by its first $K - 1$ elements of $\boldsymbol{\theta} : \theta_1 \leq \theta_2 \leq \cdots \leq 1 - \sum_{k=1}^{K} \theta_k$. Hence,

518 for the transformation we will only consider the first $K - 1$ elements of $\boldsymbol{\theta}$ and we will

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

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

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

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

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

523 The inverse transformation is given by

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

524 The Jacobian of this transformation is:

$$|J| = \prod_{k=1}^{K-1} (u_k - l_k)\, \phi(\xi_k).$$

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

526 upper bound u_k of each θ_k. Assuming $\theta_{k-1} < \theta_k$ for $k \in \{1 \cdots, K\}$ the lower bound for any

527 element in $\boldsymbol{\theta}$ is defined as

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

528   This definition holds for both binomial models and multinomial models. Differences in

529   these two models appear only when determining the upper bound for each parameter, since

530   parameters in a multinomial models are unit constrained. For binomial models, the upper

531   bound for each θ_k is simply 1. For multinomial models, however, the upper bound for each

532   θ_k depends on the size of smaller elements as well as on the number of remaining larger

533   elements in $\boldsymbol{\theta}$. To determine the upper bound for multinomial parameters we are using a

534   stick-breaking method (Frigyik, Kapila, & Gupta, 2010; Stan Development Team, 2020).

535   The stick-breaking approach represents $\boldsymbol{\theta}$ as unit-length stick which we subsequently divide

536   into K elements. By this definition, the upper bound or any θ_k is:

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{10}$$

537   where $1 - \sum_{i<k} \theta_i$ represents the length of the remaining stick, that is, the proportion of the

538   unit-length stick that still needs to be divided among the remaining elements in $\boldsymbol{\theta}$. The

539   elements in the remaining stick are denoted as ERS, and are computed as follows:

$$ERS = K - 1 + k$$

540   .

541   The transformations outlined above are suitable for binomial and multinomial models

542   featuring hypotheses in which all parameters are inequality constrained. However, when

543   hypotheses feature a combination of equality and inequality constrained parameters, as well

$_{544}$ as parameters that are free to vary we need to modify the formula to compute the upper and

$_{545}$ lower bounds:

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\[2ex] \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K, \end{cases} \tag{11}$$

$_{546}$ where e_{k-1} refers to the number of equality constrained parameters that are collapsed in θ_{k-1}.

$_{547}$ The upper bound for parameters in the binomial models still remains 1. For

$_{548}$ multinomial models, the upper bound is then defined as:

$$u_k = \begin{cases} \dfrac{1}{K} - (f_k \times l_k) & \text{if } k = 1 \\[3ex] \left(\dfrac{1 - \sum_{i<k} \theta_i}{ERS} - (f_k \times l_k) \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k \geq \max(\theta_{i<k}), \\[3ex] \left(2 \times \left(\dfrac{1 - \sum_{i<k} \theta_i}{ERS} - (f_k \times l_k) \right) - \max(\theta_{i<k}) \right) \times e_k & \text{if } 1 < k < K \text{ and } u_k < \max(\theta_{i<k}), \end{cases} \tag{12}$$

$_{549}$ where f_k represents the number of free parameters that share common upper and lower

$_{550}$ bounds. The elements in the remaining stick are then computed as follows

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

$_{551}$ The rationale behind these modifications will be described in more detail in the following

$_{552}$ sections. In **multibridge**, information that is relevant for the transformation of the

$_{553}$ parameter vectors is stored in the generated `restriction_list` which is returned by the

$_{554}$ main functions \texttt{binom_bf_informed} and \texttt{mult_bf_informed} but can also

$_{555}$ be generated separately with the function \texttt{generate_restriction_list}. This

556 restriction list features the sublist `inequality_constraints` which encodes the number of

557 equality constraints collapsed in each parameter in `nr_mult_equal`. Similarly the number of

558 free parameters that share a common bounds are encoded under `nr_mult_free`.

Equality Constrained Parameters. When informed hypotheses feature a mix of

equality and inequality constrained parameters, we collapse in the constrained prior and

posterior distributions all equality constrained parameters into one category. When

transforming the samples from these distributions, we need to account for the fact that

inequality constraints on the collapsed parameters might not hold even though the constraint

is valid under the original parameter values. For instance, for $\theta_1 = \theta_2 = \theta_3 \leq \theta_4 \leq \theta_5$, where

the elements in $\boldsymbol{\theta}$ take the values $(0.15, 0.15, 0.15, 0.25, 0.3)$, the inequality constraint does

not hold for the collapsed parameters (i.e., $\theta^* \not\leq \theta_4 \leq \theta_5$ since $0.45 \not\leq 0.25 \leq 0.3$). For these

cases, the upper and lower bounds for the parameters need to be adjusted as follows:

$$
l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K, \end{cases} \tag{13}
$$

559 where e_{k-1} and e_k refer to the number of equality constrained parameters that are collapsed

560 in θ_{k-1} and θ_k, respectively. The upper bound is defined as

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

561 where $1 - \sum_{i<k} \theta_i$ represents the length of the remaining stick and the number of elements in

562 the remaining stick are computed as follows: $ERS = \sum_k^K e_k$. The upper bound is then

563 multiplied by the number of equality constrained parameters within the current constraint.

Concretely, for the constraint above, that is $\theta^* \leq \theta_4$, the lower bound for θ^* would be 0.

565 The upper bound is computed by taking into account the number of equality constrained

566 parameters, such that $u_k = {}^1/_5 \times 3 = 0.6$. For θ_4 the lower bound is ${}^{\theta^*}/_3 = 0.15$, since 3

567 parameters are collapsed in θ^*. The upper bound for θ_4 is then $\dfrac{(1 - \theta^*)}{2} = 0.275$ and θ_5 is

568 $1 - \theta^* - \theta_4 = 1 - 0.45 - 0.25 = 0.3$.

569 ***Corrections for Free Parameters.*** Different adjustments are required for a

570 sequence of inequality constrained parameters that have shared upper and lower bounds, but

571 can vary freely within certain upper and lower bounds. For instance, the hypothesis

$$\mathcal{H}_r : \theta_1 < \theta_2, \theta_3$$

specifies that θ_2 and θ_3 have the shared lower bound θ_1 and the shared upper bound 1,

however, θ_2 can be larger than θ_3 or vice versa. To integrate these cases within the

stick-breaking approach one must account for these potential changes of order. For these

cases, the lower bounds for the parameters remain unchanged, however the upper bounds

need to be adjusted as follows:

$$u_k = \begin{cases} \dfrac{1}{K} - (f_k \times l_k) & \text{if } k = 1 \\[3mm] \dfrac{1 - \sum_{i<k} \theta_i}{ERS} - (f_k \times l_k) & \text{if } 1 < k < K, \end{cases} \tag{15}$$

where f_k represents the number of free parameters that share common upper and lower

bounds. Here, the number of elements in the remaining stick are computed as follows:

$ERS = 1 + \sum_{j>k} f_j$. Subtracting the lower bound for the remaining free parameters from

the upper bound of the current parameter secures a minimum stick-length for the remaining

free parameters to comply with the constraint. A further correction is required, if a

preceding free parameter (i.e., a free parameter that was already accounted for in the stick)

is larger than the upper bound of the current parameter. In that case, we need we subtract

the difference between the largest preceding free parameter in the sequence with the current

upper bound. Thus if $u_k < \max(\theta_{i<k})$, the upper bound becomes:

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

$$= 2 \times u_k - \max(\theta_{i<k}). \tag{17}$$

To outline when such a correction is necessary, consider the constraint $\theta_1 \leq \theta_2, \theta_3 \leq \theta_4$, where the elements in $\boldsymbol{\theta}$ take on the values $(0.1, 0.35, 0.15, 0.40)$. When transforming the parameters, the lower bound for θ_1 is 0, the upper bound $1/4$. The parameters θ_2 and θ_3 share the same lower bound, which is, $\theta_1 = 0.1$. The upper bound for θ_2, is the length of the remaining stick divided by the elements of the remaining stick, that is, $0.9/2 = 0.45$. From the resulting upper bound, we subtract the lower bound for the remaining free parameters of the sequence, which yields an upper bound for θ_2 of $0.45 - 0.1 = 0.35$. Since θ_2 is the first free parameter in the sequence that is evaluated an additional downward correction is not necessary. The upper bound for θ_3 is $(1 - 0.1 - 0.35)/2 = 0.275$. However, if θ_3 would actually take on the value 0.275, θ_4 would need to be 0.275 too, which would violate the constraint (i.e., $0.1 \leq 0.35, 0.275 \nleq 0.275$). Therefore, we apply the additional correction, such that

$$u_k = 2 \times u_k - \max(\theta_{i<k}) \tag{18}$$

$$= 2 \times 0.275 - 0.35 \tag{19}$$

$$= 0.2, \tag{20}$$

which secures the proper ordering for the remainder of the parameters, since $\theta_4 = 0.2$ would yield $0.1 \leq 0.35, 0.2 \leq 0.35$.

References

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

578   Stan Development Team. (2020). *Stan modeling language user's guide and reference*

579   *manual, version 2.23.0.* R Foundation for Statistical Computing. Retrieved from

580   http://mc-stan.org/