bayesanova: An R package for Bayesian Inference in the Analysis of Variance via Markov Chain Monte Carlo in Gaussian Mixture Models

by Riko Kelter

Abstract This paper introduces the R package **bayesanova**, which performs Bayesian inference in the analysis of variance (ANOVA). Traditional ANOVA based on null hypothesis significance testing (NHST) is prone to overestimating effects and stating effects if none are present. Bayesian ANOVAs developed so far are based on Bayes factors (BF), which also enforce a hypothesis testing stance. Instead, the Bayesian ANOVA implemented in **bayesanova** focusses on effect size estimation and is based on a Gaussian mixture with known allocations, for which full posterior inference for the component parameters is implemented via Markov-Chain-Monte-Carlo (MCMC). Inference for the difference in means, standard deviations and effect sizes between each of the groups is obtained automatically. Estimation of the parameters instead of hypothesis testing is embraced via the region of practical equivalence (ROPE), and helper functions provide checks of the model assumptions and visualization of the results.

Introduction

This article introduces **bayesanova**, an R package for conducting a Bayesian analysis of variance (ANOVA) via Markov Chain Monte Carlo (MCMC) in a Gaussian mixture model. Classic frequentist analysis of variance is based on null hypothesis significance testing (NHST), which recently has been shown to produce serious problems regarding the reproducibility and reliability of scientific results (Benjamin et al., 2018; Colquhoun, 2017, 2019; Wasserstein et al., 2019; Wasserstein and Lazar, 2016). NHST is based on test statistics, p-values and significance levels α, which are designed to control the long-term false-positive rate. Still, in a multitude of settings these approaches do in fact lead to an inflated rate of false-positive results, undermining the validity and progress of science. Examples include optional stopping of participant recruiting in studies (Carlin and Louis, 2009) or the necessary testing for violations of distributional assumptions which some frequentist hypothesis tests make (Rochon et al., 2012).

As a solution to these problems, Bayesian methods have been proposed recently and are since gaining popularity in a wide range of scientific domains (McElreath and Smaldino, 2015; Kruschke, 2013, 2015). The Bayesian philosophy proceeds by combining the model likelihood $f(x|\theta)$ with the available prior information $p(\theta)$ to obtain the posterior distribution $f(\theta|x)$ through the use of Bayes' theorem:

$$f(\theta|x) \propto f(x|\theta)f(\theta) \tag{1}$$

While the Bayesian philosophy thus allows for flexible modeling, inference for the posterior distribution $f(\theta|x)$ can be complicated in practice. Therefore, Markov chain Monte Carlo techniques have been developed, which make use of the facts that (1) constructing a Markov chain which has the posterior distribution $f(\theta|x)$ as its stationary distribution, and (2) drawing samples from this Markov chain to approximate the posterior $f(\theta|x)$ can be used to obtain the posterior numerically.

The **bayesanova** package is designed as a Bayesian alternative to the frequentist analysis of variance. By using a Gaussian mixture model and implementing a Markov Chain Monte Carlo algorithm for this model, full posterior inference can be obtained. This allows for explicit hypothesis testing between groups as in the frequentist ANOVA, or for estimation of parameters under uncertainty. The focus in **bayesanova** is on the latter perspective and avoids explicit hypothesis testing. While Bayesian versions of the analysis of variance have been proposed recently by Rouder et al. (2012) and Bergh et al. (2019), these implementations focus on the Bayes factor as a measure of evidence (van Doorn et al., 2019; JASP Team, 2019). As the Bayes factor suffers from multiple problems, one of which is its strong dependence on the used priors – see Kamary et al. (2014) and Robert (2016) – the implementation in **bayesanova** avoids the Bayes factor and uses a different posterior index, the region of practical equivalence (ROPE) (Kruschke, 2018), which has lately been shown to have some desirable properties, in particular in contrast to the Bayes factor (Makowski et al., 2019b).

The plan of the paper is as follows: The next section introduces the analysis of variance in a frequentist and Bayesian fashion and gives an overview about packages implementing these methods.

The following section then introduces the novel approach implemented in **bayesanova**. The details on the mixture representation of the Bayesian analysis of variance are discussed and scenarios where **bayesanova** is designed to be used are detailed. The section thereafter outlines the structure of the package and details the included functions. The following section presents a variety of examples and illustrations using real datasets from biomedical and psychological research as well as synthetic datasets. The last section then provides a summary of the benefits and drawbacks of the used implementation, as well as future plans for the package.

Frequentist and Bayesian analysis of variance

Traditional ANOVA models using NHST via the F-statistic

In applied statistics, the one-way analysis of variance is a method which can be used to compare means of two or more samples (typically three). The one-way ANOVA assumes numerical (response) data in each group and (usually) categorical input data like a group indicator in a randomized clinical trial (RCT). Interpreting the ANOVA as a linear model, one obtains for data $y_{i,j}$, where $i = 1, ..., n$ is an index over the experimental units (patients, participants) and $j = 1, ..., k$ an index over treatment groups

$$y_{i,j} = \mu_j + \varepsilon_{i,j} \tag{2}$$

if the experiment is completely randomized. Here, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ so that $\varepsilon_{i,j}$ are normally distributed zero-mean residuals. μ_j is the mean of treatment group j and $y_{i,j}$ the response variable which is measured in the experiment.

The one-way ANOVA then tests the null hypothesis H_0 that all samples are drawn from populations with identical means. To do this, (1) two estimates of the population variance are obtained which rely on various assumptions and (2) an F-statistic is produced by the ANOVA, which is the ratio of variance calculated among the means to the variance within the samples. The intuition here is that if group means are drawn from populations with identical means, the variance of the group means should be smaller than the variance of samples and a high ratio thereby indicates differing means. Mathematical details on computing the F-statistic can be found in the Appendix.

The one-way ANOVA as detailed above makes several assumptions, the most important of which are: (1) variances of populations are equal; (2) responses for a given group are independent and identical distributed random variables; (3) response variable residuals are normally distributed, that is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

While Monte Carlo studies have shown that the ANOVA is quite robust to small to medium violations of these assumptions (Donaldson, 1966), severe violations of assumptions (1)-(3) will result in inflated rates of false positives and and thereby unreliable results (Tiku, 1971).

Bayesian ANOVA models

Bayesian models for the ANOVA have been developed recently to solve some of the problems of NHST. The developed models can be categorized broadly into two approaches: The first approach relies on the Bayes factor as a measure of relative evidence and was developed by Rouder et al. (2012). The second approach is based on MCMC algorithms like Gibbs sampling in JAGS (Plummer, 2003) or Hamiltonian Monte Carlo (HMC) in Stan (Carpenter et al., 2017; Stan Development Team, 2020). This approach was popularized by Kruschke (2015). Here the region of practical equivalence (ROPE) as introduced by Kruschke (2015) is used for measuring the evidence given the data. Also, an explicit hypothesis testing stance is avoided.

The approach of Rouder et al. (2012) can be summarized as follows: An independent Cauchy prior is considered

$$p(\theta) = \prod_{i=1}^{p} \frac{1}{(1 + \theta_i^2)\pi} \tag{3}$$

for the vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ of the p effects between different groups. For example, in a three-group setting there would be three effects θ_1, θ_2 and θ_3 corresponding to the effects between the first and second, first and third, and second and third group. In the case of $k = 4$ groups, there are $p = 6$ effects and so on. The ANOVA is then rewritten as a linear model

$$\boldsymbol{y} = \boldsymbol{\mu}\mathbf{1} + \sigma \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{4}$$

where $\boldsymbol{\mu}$ is the grand mean parameter, $\mathbf{1}$ a column vector of length n with entries equal to 1, $\boldsymbol{\theta}$ a column

vector of the standardized effect size parameters of length p, and $\boldsymbol{X}$ is the $n \times p$ design matrix. The factor σ in $\sigma\boldsymbol{X}\boldsymbol{\theta}$ is attributed to the reparameterization according to Jeffreys: Following Jeffreys (1961) by reparameterizing $\delta = \mu/\sigma$, where δ is the effect size of Cohen (1988), Rouder et al. (2012) rewrote the observed data sampling distribution as

$$y_i \sim \mathcal{N}(\sigma\delta, \sigma^2) \tag{5}$$

The residuals $\boldsymbol{\varepsilon}$ in Equation (4) are defined to be

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}) \tag{6}$$

with $\boldsymbol{I}$ being the identity matrix of size n and $\mathbf{0}$ a column vector of zeros of size n.

Putting a Jeffreys prior $p(\mu, \sigma^2) = 1/\sigma^2$ on the mean and variance, and assuming the following g-prior structure

$$\boldsymbol{\theta}|\boldsymbol{G} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G}) \tag{7}$$

which is based on Zellner (1980), where $\boldsymbol{G}$ is a $p \times p$ diagonal matrix, the only open aspect remaining is putting a prior on the diagonal elements g_l of $\boldsymbol{G}$ for $l = 1, ..., p$. (Rouder et al., 2012) chose

$$g_l \sim \text{Inverse-}\chi^2_1 \tag{8}$$

so that the marginal prior on the effect size parameter vector $\boldsymbol{\theta}$ results in the independent Cauchy distribution given in Equation (3). Rouder et al. (2012) then showed that the resulting BF_{10} can be written as

$$BF_{10} = \int_g K(\boldsymbol{n}, g) \left(\frac{\sum_i \sum_j (y_{ij} - \bar{y})^2 + \frac{1}{g}(\sum_i c_i \bar{y}_i^2 - (\sum_i c_i \bar{y}_i)^2 / (\sum_i c_i))}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \right)^{-(N-1)/2} p(g) dg \tag{9}$$

if a balanced one-way design is used (equal sample sizes in each group). Here, $\boldsymbol{n} = (n_1, ..., n_p)'$ is the vector of sample sizes for each effect $1, ..., p$, $n = \sum_i n_i$ is the full sample size, $c_i = n_i/(n_i + 1/g)$ and

$$K(\boldsymbol{n}, g) = \sqrt{N} \left(\frac{\prod_i 1/(1 + g n_i)}{\sum_i n_i/(1 + g n_i)} \right)^{1/2} \tag{10}$$

In summary, this Bayes factor of Rouder et al. (2012) can be computed via Gaussian quadrature, as it constitutes a one-dimensional integral after inserting the necessary quantities.

The second approach of a Bayesian ANOVA model can be credited to Kruschke (2015), who uses the MCMC sampler JAGS (Plummer, 2003) to obtain full posterior inference in his model instead of relying on the Bayes factor. The reasons for avoiding the Bayes factor as a measure of evidence are that (1) it depends strongly on the selected prior modeling (Kamary et al., 2014); (2) the Bayes factor states only relative evidence for the alternative to the null hypothesis (or vice versa) so that even a large Bayes factor does not indicate that either one of both hypotheses is a good fit for the actual data (Kelter, 2020a,b); (3) it can be located in the same formalism of hypothesis testing the pioneers of frequentist testing advocated at the time of invention (Robert, 2016; Tendeiro and Kiers, 2019). In addition, the calculation of the Bayes factor for increasingly complex models can be difficult, as the above derivations of Rouder et al. (2012) exemplify, see also Kamary et al. (2014). Importantly, the Bayes factor assigns positive measure to a Lebesgue-null-set which is puzzling from a measure-theoretic perspective, compare Kelter (2021c), Rao and Lovric (2016), and Berger (1985).

Kruschke (2015) modeled the Bayesian ANOVA for k groups and n observations $y_1, ...y_n$ as a hierarchical Bayesian model, where

$$y_i \sim \mathcal{N}(\mu, \sigma_y^2) \tag{11}$$

where the standard deviation σ_y is modelled as

$$\sigma_y \sim \mathcal{U}(L, H) \tag{12}$$

the mean μ_i is the linear combination

$$\mu = \beta_0 + \sum_{j=1}^{k} \beta_j x_j(i) \tag{13}$$

and the coefficients of this linear combination are given as

$$\beta_0 \sim \mathcal{N}(M_0, S_0) \tag{14}$$
$$\beta_j \sim \mathcal{N}(0, \sigma_\beta) \tag{15}$$

where $x_j(i)$ is the index for the group the observation y_i belongs to. If, for example, y_i is in the first group, $x_1(i) = 1$ and $x_j(i) = 0$ for all $j \neq 1$ with $j \in \{1, ..., k\}$, yielding the group mean $\mu_i = \beta_0 + \beta_1$ of the first group. Thus, although Equation (11) seems to indicate that there is a single mean μ for all observations y_i, $i = 1, ..., n$, the mean μ takes k different values depending on which group the observation y_i is located in. These k different values for μ correspond to the different means in the k groups as shown in Equation (13). The variables L, H, M_0, S_0 are hyperparameters, and the parameter β_j can be interpreted as the effect size differing from the grand mean β_0, which is why the prior on β_j is normal with mean zero so that the expectation of these effect size differences from the grand mean sum up to zero again. The hyperparameter σ_β can either be set constant or given another prior, extending the multilevel model, where Kruschke (2015) followed the recommendations of Gelman and Hill (2006) to use a folded t-distribution or a gamma-distribution with non-zero mode.

Inference for the full posterior, that is for the parameters $\mu_k, \sigma_y, \beta_0, \beta_j \forall j, j = 1, ..., k$ (and σ_β, if a hyperprior like a folded t-distribution or gamma-distribution is used on this parameter) given the data is provided via the MCMC sampler JAGS (Plummer, 2003), which uses Gibbs sampling to draw samples from the posterior. Posterior distributions obtained through Gibbs sampling are finally used to estimate all parameters via 95% Highest-Density-Intervals (HDI). Explicit testing is avoided.

Available software

Available software for the traditional ANOVA

Conducting a traditional analysis of variance is possible with an abundance of software, for example via the **stats** package (R Core Team, 2020) which is part of the R programming language (R Core Team, 2020).

Available software for the Bayesian ANOVA

The **BayesFactor** package by Morey and Rouder (2018) provides the Bayesian ANOVA Bayes factor of Rouder et al. (2012), and various helper functions for analysis of the results.

A simple illustration of the main workflow in the **BayesFactor** package is given here, using the `ToothGrowth` dataset in the **datasets** package (Cannon et al., 2019). The `ToothGrowth` dataset contains three columns: `len`, the dependent variable each of which is the length of a guinea pig's tooth after treatment with vitamin C. The predictor `supp` corresponds to the supplement type (either orange juice or ascorbic acid), the predictor `dose` is the amount of vitamin C administered.

The **BayesFactor** package's core function allows the comparison of models $\mathcal{M}_0, ..., \mathcal{M}_n$ with factors as predictors. The null model without any predictors is most often compared to models including predictors or even interaction terms using the Bayes factor as detailed above. The function `anovaBF` computes several model estimates at once, so that the model with the largest Bayes factor can be selected. The data are first loaded and the categorial predictors converted to factors:

```
R> set.seed(42)
R> library(datasets)
R> data(ToothGrowth)
R> head(ToothGrowth,n=3)

   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5

R> ToothGrowth$dose = factor(ToothGrowth$dose)
R> levels(ToothGrowth$dose) = c('Low', 'Medium', 'High')
```

Then, a Bayesian ANOVA is conducted using both predictors `dose`, `supp` and the interaction term `dose * supp`:

```
R> library(BayesFactor)
R> bf = anovaBF(len ~ supp * dose, data = ToothGrowth)

Bayes factor analysis
--------------
[1] supp                  : 1.198757      +- 0.01%
[2] dose                  : 4.983636e+12  +- 0%
[3] supp + dose           : 2.963312e+14  +- 1.59%
[4] supp + dose + supp:dose : 8.067205e+14  +- 1.94%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

The results are shown in form of the Jeffreys-Zellner-Siow (JZS) Bayes factor BF_{10} detailed previously. As the BF_{10} for the model including both predictors `supp` and `dose` is largest, the Bayesian ANOVA favours this model over the null model which includes only the intercept. Thus, as there are the low, medium and high dose groups and the two supplement groups, in total one obtains $3 \times 2 = 6$ different groups. The results show that there is strong evidence that the model attesting these six differing groups is favourable over the null model (and every other model as given in output lines `[1]`, `[2]` and `[3]`).

Note, that this solution is also implemented in the open-source software JASP, for an introduction see Bergh et al. (2019).

The Bayesian ANOVA model of Kruschke (2015) is not implemented in a software package by now. Instead, users have to write their own model scripts for JAGS (Plummer, 2003) to run the analysis. Still, recently the package **BANOVA** was published by Dong and Wedel (2019), which uses JAGS (Plummer, 2003) and the Hamiltonian Monte Carlo (HMC) sampler Stan (Carpenter et al., 2017) via the package **RStan** (Stan Development Team, 2020) to provide similar inferences without the need to code the JAGS or Stan models on your own.

Note that in the above example, a traditional ANOVA can easily be fit via

```
R> summary(aov(len ~ supp * dose, data = ToothGrowth))

            Df Sum Sq Mean Sq F value   Pr(>F)
supp         1  205.4   205.4  15.572 0.000231 ***
dose         2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose    2  108.3    54.2   4.107 0.021860 *
Residuals   54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which yields similar results, favouring the full model with both predictors and interaction term, as both predictors and the interaction term are significant.

The Bayesian ANOVA model based on Gaussian mixtures

The method used in the **bayesanova** package is based on estimation of parameters in a Gaussian mixture distribution. On this mixture a Gibbs sampling algorithm is applied to produce posterior distributions of all unknown parameters given the data in the Gaussian components, that is for $\mu_j, \sigma_j, j = 1, ..., k$ and for the differences in means $\mu_l - \mu_r, l \neq r$ and the effect sizes $\delta_{lr}, l \neq r$ where k is the number of groups in the study or experiment. This way, a relatively complete picture of the situation at hand can be drawn and while the technical aspects are omitted here, the validity of the procedure stems from standard MCMC theory, see for example Robert and Casella (2004). The principal idea of mixture models is expressed by Frühwirth-Schnatter (2006):

> *Consider a population made up of K subgroups, mixed at random in proportion to the relative group sizes $\eta_1, ..., \eta_K$. Assume interest lies in some random feature Y which is heterogeneous across and homogeneous within the subgroups. Due to heterogeneity, Y has a different probability distribution in each group, usually assumed to arise from the same parametric family $p(y|\theta)$ however, with the parameter θ differing across the groups. The groups may be labeled through a discrete indicator variable S taking values in the set $\{1, ..., K\}$.*
> *When sampling randomly from such a population, we may record not only Y, but also the*

group indicator S. The probability of sampling from the group labeled S is equal to η_S, whereas conditional on knowing S, Y is a random variable following the distribution $p(y|\theta_S)$ with θ_S being the parameter in group S. (...) The marginal density $p(y)$ is obviously given by the following mixture density

$$p(y) = \sum_{S=1}^{K} p(y, S) = \eta_1 p(y|\theta_1) + ... + \eta_S p(y|\theta_K)$$

Clearly, this resembles the situation of the analysis of variance, in which the allocations S are known. Traditionally, mixtures are treated with missing allocations but in the setting of the ANOVA these are known, leading to a much simpler scenario. This interpretation also makes sense from a semantic point: the inherent assumption of a researcher is that the population is indeed made up of k subgroups in the case of a k-group ANOVA, which differ in a random feature Y which is heterogeneous across groups and homogeneous within each group. When conducting for example a randomized clinical trial (RCT), the group indicator S is of course recorded. The clinician will choose the patients according to a sampling plan, which could be designed to achieve equally sized groups, that is, $\eta_1 = \eta_2 = ... = \eta_k$ for k study groups. Thus, when sampling the population with the target of equally sized groups, the researcher will sample the objects with equal probability from the population. Consider a treatment one, treatment two and a control group. In this typical setting, the researcher could flip a coin for each patient in the RCT to assign him or her to one of the two treatment groups or to the control group, so that with probability $\eta_1 = \eta_2 = \eta_3 = 1/3$ for any group, the patient is assigned to it. Repeating this process then leads to the mixture model given above. After the RCT is conducted, the resulting histogram of observed Y values will finally take the form of the mixture density $p(y)$ above. If there is an effect in the treatment, this density $p(y)$ will express three modes which in turn result from the underlying mixture model of the data-generating process.

If unbalanced groups are the goal, weights can be adjusted accordingly, for example $\eta_1 = 0.3$, $\eta_2 = 0.2$ and $\eta_3 = 0.5$. After fixing the mixture weights η_1, η_2, η_3, the family of distributions for the mixture components needs to be selected. The above considerations lead to finite mixtures of normal distributions which *'occur frequently in many areas of applied statistics such as [...] medicine'* (Frühwirth-Schnatter, 2006, p. 169). The components $p(y|\theta_i)$ therefore become $f_N(y; \mu_j, \sigma_j^2)$ for $j = 1, ..., k$ in this case, where $f_N(y; \mu_j, \sigma_j^2)$ is the density of the univariate normal distribution. Parameter estimation in finite mixtures of normal distributions consists of estimation of the component parameters (μ_j, σ_j^2), the allocations $S_i, i = 1, ..., n$ and the weight distribution $(\eta_1, ..., \eta_k)$ based on the available complete data $(y_i, S_i), i = 1, ..., n$. In the case of the Bayesian ANOVA, the allocations S_i (where $S_i = 1$ if y_i belongs to the first component, $S_i = 2$ if y_i belongs to the second component, until $S_i = k$ if y_i belongs to the k-th group) are known for all observations y_i, $i = 1, ..., n$. Therefore, inference reduces to inference for the density parameters (μ_j, σ_j^2) of the normal components of the mixture for the $j = 1, ..., k$ groups.

The Bayesian ANOVA model based on Gaussian mixtures is summarized in Figure 1 using the three-group case as an example:

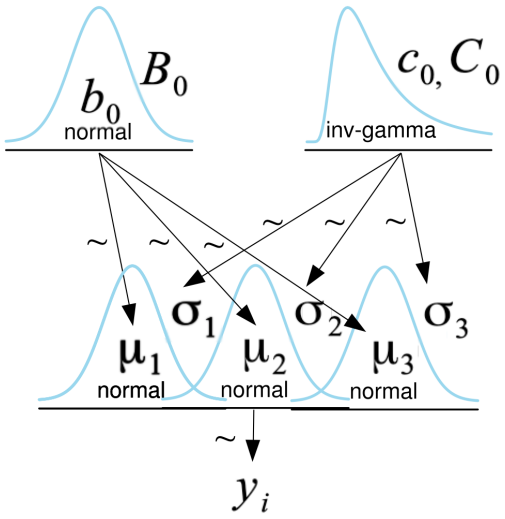

Figure 1: Three-component Gaussian mixture with known allocations for the Bayesian analysis of variance

The measured variables y_i follow a three-component Gaussian mixture with known allocations.

Model	Purpose	Evidence measure	Computational aspects
Frequentist ANOVA	Testing the global hypothesis that all samples are drawn from populations with identical means against the alternative	F-statistic and p-value	Analytic solution
Bayesian ANOVA of Rouder et al. (2012)	Test the global hypothesis that the effect size vector is zero versus the alternative	Bayes factor	Numerical integration required
Bayesian ANOVA of Kruschke (2015)	Estimation of effect sizes between groups via ROPE and 95% HPD	ROPE	Gibbs sampling in MCMC sampler JAGS (or Stan) required
Bayesian ANOVA based on Gaussian mixtures	Estimation of effect sizes between groups via the ROPE and posterior probability mass	ROPE	Gibbs sampling without MCMC sampler JAGS (or Stan) required

Table 1: Overview about the four ANOVA models

The first group is normally distributed as $\mathcal{N}(\mu_1, \sigma_1)$, the second group as $\mathcal{N}(\mu_2, \sigma_2)$ and the third group as $\mathcal{N}(\mu_3, \sigma_3)$. The means μ_1, μ_2 and μ_3 are each distributed as $\mu_j \sim \mathcal{N}(b_0, B_0), j = 1, 2, 3$ with noninformative hyperparameters b_0 and B_0 and the standard deviations σ_1, σ_2 and σ_3 are distributed as $\sigma_j \sim \mathcal{G}^{-1}(c_0, C_0), j = 1, 2, 3$ with noninformative hyperparameters c_0 and C_0. For details, see Kelter (2021d, 2020c). As the allocations are known, the weights η_1, η_2 and η_3 are known too, and need not to be estimated, which is why the parameters η_1, η_2, η_3 are not included in the diagram. The model visualized in Figure 1 can be generalized for an arbitrary number of mixture components, which then includes nearly arbitrary ANOVA settings for comparison of multiple groups. A definitive advantage of this model is that inference is obtained for both means and standard deviations, yielding richer information compared to the testing perspectives which are stressed in traditional or Bayesian ANOVA models focussing on the Bayes factor. Also, posterior distributions of effect sizes can be obtained via MCMC, providing an additional layer of information to draw inferences.

Instead of relying on the Bayes factor, the **bayesanova** package follows the approach of Kruschke (2018) to use a region of practical equivalence (ROPE). The effect size δ is routinely categorized as small, medium or large in medical research when $\delta \in [0.2, 0.5), \delta \in [0.5, 0.8)$ or $\delta \in [0.8, \infty)$, see Cohen (1988). The approach using the ROPE proceeds by taking these categories as regions of practical equivalence, that is both $\delta = 0.25$ and $\delta = 0.26$ are identified as a small effect because both are inside the region of practical equivalence $[0.2, 0.5)$ of a small effect δ. The underlying idea is that measuring effect sizes only makes sense up to a specific precision, which is given by the above categorization of effect sizes. By studying how much probability mass of the posterior distribution of δ lies inside some of the above ROPEs $[0.2, 0.5)$, $[0.5, 0.8)$ and $[0.8, \infty)$ of a small, medium and large positive effect for δ (negative effects analogue), a continuous statement about the most probable effect size δ given the data can be made. Kruschke originally advocated to use the location of the 95% highest-posterior-density (HPD) interval in relation to the ROPE to test whether the null value in the middle of the ROPE should be accepted or rejected for practical purposes. Here, this approach is generalized by switching to the amount of posterior probability mass inside the ROPE. Detailed examples are provided later in this paper.

Table 1 provides an overview about the four ANOVA models and their purpose. Although it appears that the model of Kruschke (2015) and the Gaussian mixture modeling approach proposed in this paper have the same purpose, they differ in how data y_i are assumed to be generated. In the mixture approach we assume that the sample of n_j participants in group j results from a mixture process, e.g. by flipping a coin, rolling a dice or using any other randomization device (as is the case in clinical trials when assigning patients to groups according to a double-blinded protocol). Thus, the process of data generation is not "one has collected n_j participants for group j" but "the given sample of n_j participants in group j is assumed to be a realization of a mixture process where with probability η_j participants are assigned to group j". Importantly, note that the realization of n_j participants in group j for $j = 1, ..., k$ is expected under the mixture component weight $\eta_j = n_j/n$, but also entirely different group sizes n_j can result under such a mixture. In fact, the weights $\eta_j = n_j/n$ which are assumed to be known are the corresponding maximum-likelihood-estimators of the weight parameters η_j given the sample sizes n_j for $j = 1, ..., k$, but the conceptual focus of the mixture approach is to

Function	Description
`bayes.anova`	Main function of the package, conducts the MCMC algorithm to provide full posterior inference in the three-component Gaussian mixture model
`assumption.check`	Helper function for checking the assumption of normality in each group previous to running a Bayesian ANOVA
`anovaplot`	Provides multiple visualizations of the results, including posterior distributions, difference in means and standard deviations and effect sizes as well as a full ROPE-analysis
`post.pred.check`	Provides a posterior predictive check for a fitted Bayesian ANOVA model

Table 2: Outline of the four main functions implemented in **bayesanova**

closely mimic the randomization process researchers follow when conducting a randomized controlled trial. Note further that the model of Kruschke assumes homogeneity of variances in contrast to the Gaussian mixture model, but Kruschke's model can easily be extended to account for heterogeneity of variance, rendering this difference less important. Note that both the frequentist ANOVA and the Bayesian version of Rouder et al. (2012) assume homogeneity of variance across groups.

Package structure and implementation

The **bayesanova** package has four functions. These provide (1) the MCMC algorithm for conducting the Bayesian ANOVA in the Gaussian mixture model with known allocations, detailed above, (2) checks of the model assumptions and (3) visualizations of the posterior results for easy interpretation and communication of research results. Visualizations of the posterior mixture components in comparison with the original data are provided by the fourth function. An overview is provided in Table 2.

The core function is `bayes.anova`, which provides the MCMC algorithm to obtain full posterior inference in a k-component Gaussian mixture model shown in Figure 1 for the special case of $k = 3$ components. The function implements a Gibbs sampling algorithm, which iteratively updates

1. the means $\mu_j|\mu_{-j}, \sigma_1, ..., \sigma_k, S, y$ given the other means μ_{-j} and standard deviations $\sigma_1, ..., \sigma_k$ as well as the full data S, y, where S is the indicator vector for the groups the observations y belong to
2. the standard deviations $\sigma_j|\sigma_{-j}, \mu_1, ..., \mu_k, S, y$ given the other standard deviations σ_{-j} and means $\mu_1, ..., \mu_k$ as well as the full data S, y, where S is again the indicator vector for the groups the observations y belong to

The details of the Gibbs sampler can be found in Kelter (2020c, 2021d), and the validity of the method follows from standard MCMC theory, see for example Robert and Casella (2004).

The `bayes.anova` function takes as input three numerical vectors `first`, `second` and `third`, which correspond to the observed responses in each of the three groups and provides multiple optional parameters:

```
bayes.anova(n=10000, first, second, third,
fourth = NULL, fifth = NULL, sixth = NULL,
hyperpars="custom", burnin=n/2, sd="sd", q=0.1, ci=0.95)
```

These are the only mandatory input values, and currently six groups are the limit **bayesanova** supports. More than three groups can be handed to the function by providing numerical vectors for the parameters `fourth`, `fifth` and `sixth`.

If no other parameters are provided, the function chooses a default of `n=10000` Gibbs sampling iterations, where the burn-in of the Markov chains is set to `burnin=n/2`, so that the first 5000 iterations are discarded. The default setting uses inference for means μ_j and standard deviations σ_j, which is indicated by the parameter `sd="sd"`, but inference for variances σ_j^2 instead of standard deviations σ_j can easily be obtained by setting `sd="var"`. The credible level for all computed credible intervals defaults to 0.95, indicated by `ci=0.95`. The two remaining parameters `hyperpars` and `q` define preselected values for the hyperparameters in the prior, to ensure weakly informative priors are used which influence the analysis as little as possible. For details, see Kelter (2020c, 2021d), but in general these values apply to a broad range of contexts so that changing them is not recommended. Note, that another set of hyperparameters based on Raftery (1996) can be selected via `hyperpars="rafterys"`, if desired.

After execution, the function returns a dataframe including four Markov chains for each parameter of the specified size `n-burnin`, to make subsequent convergence assessment or post-processing of the MCMC results possible.

The second function is `assumption.check`. This function runs a preliminary assumption check on the data, which is recommended before running a Bayesian ANOVA. The model assumptions are normality in each mixture component, so that the `assumption.check` function runs Shapiro-Wilk tests to check for normality (Shapiro and Wilk, 1965). The input parameters are the three numerical vectors `x1`, `x2` and `x3` including the observed responses in the first, second and third group, and the desired confidence level `conf.level` for the Shapiro-Wilk tests:

```
assumption.check(x1, x2, x3, x4 = NULL, x5 = NULL, x6 = NULL, conf.level=0.95)
```

The default confidence level is `0.95`. More than three groups can easily be added by providing values for `x4`, `x5` and `x6`.

The third function is `anovaplot`, which provides a variety of visualizations of results. The function takes as input a dataframe `dataframe`, which should be the result of the `bayes.anova` function detailed above, a parameter `type`, which indicates which visualization is desired, a parameter `sd`, which indicates if the provided dataframe includes posterior draws of σ_j or σ_j^2 and last a parameter `ci`, which again defined the credible level used in the computations.

```
anovaplot(dataframe, type="rope", sd="sd", ci=0.95)
```

The default values for `sd` is `"sd"`, and the default credible level is `0.95`. The `type` parameter takes one of four possible values: (1) `type="pars"`, (2) `type="diff"`, (3) `type="effect"` and (4) `type="rope"`. In the first case, posterior distributions of all model parameters are produced, complemented by convergence diagnostics in form of trace plots, autocorrelation plots and the Gelman-Brooks-Rubin shrink factor (Gelman and Brooks, 1998), which should be close to one to indicate convergence to the posterior. In the second case, the posterior distributions of the differences $\mu_i - \mu_j, j \neq i$ of the group means and differences $\sigma_l - \sigma_r, l \neq r$ of the group standard deviations (or variances, if `sd="var"` and the dataframe includes posterior draws of the σ_j^2's instead of σ_j's) are produced, complemented by the same convergence diagnostics. In the third case, the posterior distributions of the effect sizes $\delta_{lr}, l \neq r$ are produced, which are most often of interest in applied research. In this case, posteriors are complemented by the same convergence diagnostics, too. The last and fourth case produces a full ROPE-analysis, which does provide the posteriors of the effect sizes $\delta_{lr}, l \neq r$, but additionally computes a partitioning of the posterior probability mass into the standardized ROPEs of small, medium and large (and no) effect sizes according to Cohen (1988), which are the reference standard in medical and psychological research.

The last function `post.pred.check` provides a posterior predictive check for a fitted Bayesian ANOVA model against the original data, which is routine in a Bayesian workflow Gabry et al. (2019).

Illustrations and examples

This section provides illustrations and a variety of examples, in which the **bayesanova** package can be used and provides richer information than existing solutions.

Tooth growth of guinea pigs treated with vitamin C

The guinea pig dataset from above is used as a first example. The data are included in the dataset `ToothGrowth` in the **datasets** package which is part of R. First, data is loaded and split into three groups, corresponding to a low, medium and high administered vitamin C dose:

```
R> library(datasets)
R> data(ToothGrowth)
R> head(ToothGrowth,n=3)

   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5

R> library(dplyr)
R> library(tidyr)
R> library(bayesanova)
```

```
R> grp1 = (ToothGrowth %>% filter(dose==0.5) %>% select(len))$len
R> grp2 = (ToothGrowth %>% filter(dose==1.0) %>% select(len))$len
R> grp3 = (ToothGrowth %>% filter(dose==2.0) %>% select(len))$len
```

Next, we run the assumption checks on the data

```
R> assumption.check(grp1, grp2, grp3, conf.level=0.95)

Model assumptions checked. No significant deviations from normality detected.
Bayesian ANOVA can be run safely.
```

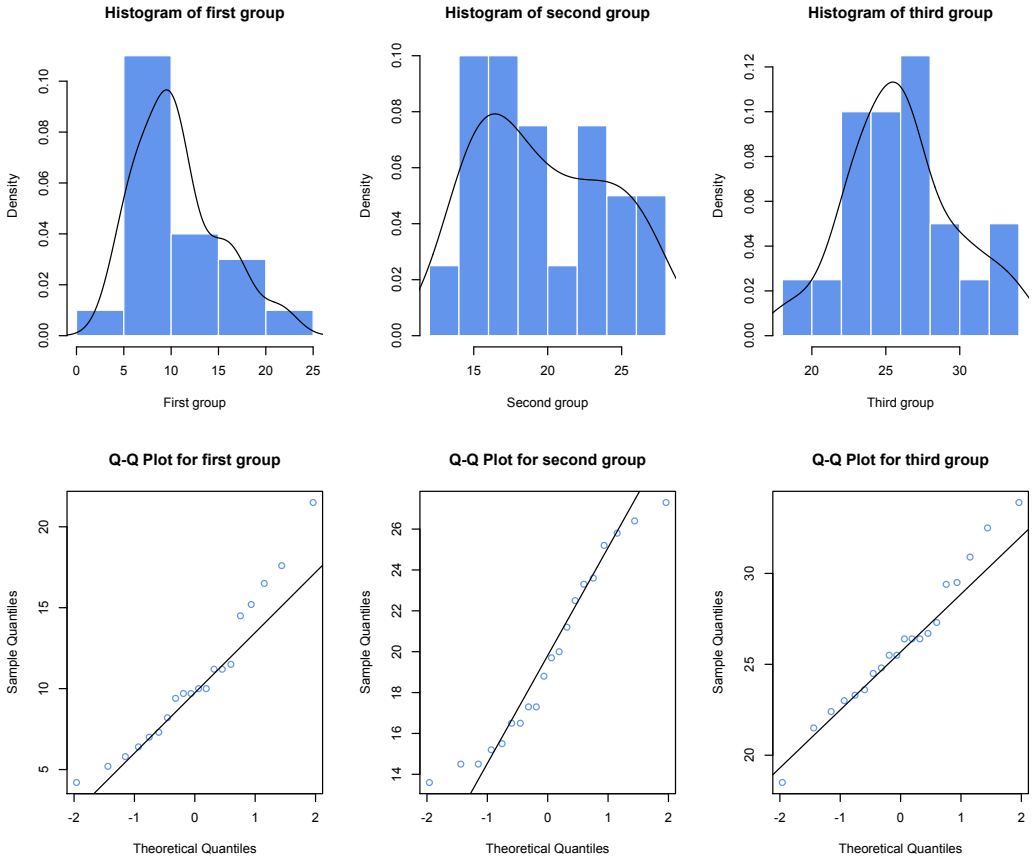

Figure 2: Assumption checks for the `ToothGrowth` dataset using the `assumption.check()` function in **bayesanova**, showing that data in the three groups can be assumed as normally distributed so that running the Bayesian ANOVA based on the Gaussian mixture model is justified

Figure 2 shows the histograms and quantile-quantile plots for all three groups produced by `assumption.check()`. Clearly, there are no large deviations, and no Shapiro-Wilk test was significant at the 0.05 level.

Next, the Bayesian ANOVA can be run via the `bayes.anova` function. Therefore, the default parameter values are used, yielding `n=5000` posterior draws:

```
R> set.seed(42)
R> res = bayes.anova(first = grp1, second = grp2, third = grp3)
```

Parameter	LQ	Mean	UQ	Std.Err
mu1	8.69	10.61	12.5	0.91
mu2	18.05	19.75	21.46	0.84
mu3	24.94	26.1	27.25	0.57
sigma1	3.02	4.07	5.67	0.67
sigma2	2.95	3.96	5.43	0.64
sigma3	2.43	3.25	4.42	0.52
mu2-mu1	6.7	9.15	11.7	1.25
mu3-mu1	13.42	15.49	17.67	1.06

```
|mu3-mu2       |4.36  |6.34  |8.38  |1.01    |
|sigma2-sigma1 |-2.02 |-0.11 |1.68  |0.93    |
|sigma3-sigma1 |-2.62 |-0.81 |0.74  |0.85    |
|sigma3-sigma2 |-2.46 |-0.71 |0.85  |0.82    |
|delta12       |-5.77 |-4.59 |-3.21 |0.65    |
|delta13       |-9.37 |-8.14 |-6.63 |0.71    |
|delta23       |-4.36 |-3.36 |-2.19 |0.56    |
```

The results table shows the lower and upper quantile, corresponding to the $100 \cdot \texttt{ci} + (100 - \texttt{ci})/2$ and $(100 - \texttt{ci})/2$ quantiles where `ci` is the credible level chosen above. Also, the posterior mean and standard error are given for each parameter, difference of parameters and effect size. The results clearly show that there are huge differences between the groups: For example, one can immediately spot that the more vitamin c given, the more tooth growth can be observed via tooth lengths. While the first group (low dose) has a posterior mean of 10.61 with credible interval $[8.69, 10.61]$, the second group achieves a mean of 19.75 with credible interval $[18.05, 21.46]$. The third group has a posterior mean of even 26.1 with credible level [24.94.27.25]. The posterior estimates for the differences $\mu_2 - \mu_1$, $\mu_3 - \mu_1$ and $\mu_3 - \mu_2$ show that all groups differ from each other with a very high probability, given the data.

Note that the information provided is much more fine-grained than in the solutions via the traditional ANOVA and the Jeffreys-Zellner-Siow based Bayes-factor ANOVA above. While in these two solutions, one could only infer that the model using both predictors and the interaction term is the best, now we are given precise estimates of the effect sizes between each group defined by the dose of vitamin c administered. Note also, that including the second predictor `supp` is no problem, leading to a setting which incorporates six groups in the mixture then.

Heart rate data for runners

The second example is from the biomedical sciences and uses the heart rate data from Moore et al. (2012). In the study, heart rates of female and male runners and generally sedentary participants (not regularly running) following six minutes of exercise were recorded. The participant's `Gender` and `Heart.rate` are given and which group he or she belongs to (`Group=="Runners"` or `Group=="Control"`). In the study, 800 participants were recruited, so that in each of the four groups given by the combinations of `Gender` and `Group` 200 subjects participated.

Therefore, the situation requires a 2×2 between subjects ANOVA. Specifically, interest lies in the hypothesis that heart rate differs between gender and groups. The Bayesian ANOVA of **bayesanova** can easily be applied in such an often encountered setting. We first load the data and split them into the four groups:

```
R> library(dplyr)
R> hr=read.csv("heartrate.csv",sep=",")
R> head(hr)

  Gender   Group Heart.Rate
1 Female Runners        119
2 Female Runners         84
3 Female Runners         89
4 Female Runners        119
5 Female Runners        127
6 Female Runners        111

R> femaleRunners = (hr %>% filter(Gender=="Female")
+  %>% filter(Group=="Runners")
+  %>% select(Heart.Rate))$Heart.Rate
R> maleRunners = (hr %>% filter(Gender=="Male") %>% filter(Group=="Runners")
+  %>% select(Heart.Rate))$Heart.Rate
R> femaleControl = (hr %>% filter(Gender=="Female")
+  %>% filter(Group=="Control")
+  %>% select(Heart.Rate))$Heart.Rate
R> maleControl = (hr %>% filter(Gender=="Male") %>% filter(Group=="Control")
+  %>% select(Heart.Rate))$Heart.Rate
```

Then, we check the model assumptions:

```
R> assumption.check(femaleRunners, maleRunners, femaleControl, maleControl)
```

We can thus safely proceed running the Bayesian ANOVA:

```
R> set.seed(42)
R> resRunners = bayes.anova(first = femaleRunners, second = maleRunners,
+  third = femaleControl, fourth = maleControl)
```

Parameter	LQ	Mean	UQ	Std.Err
mu1	113.48	116	118.5	1.27
mu2	102.51	103.98	105.55	0.76
mu3	145.44	148.04	150.52	1.3
mu4	127.12	130.01	132.82	1.47
sigma1	14.38	15.87	17.51	0.8
sigma2	11.21	12.35	13.67	0.63
sigma3	14.71	16.19	17.85	0.82
sigma4	15.46	17.02	18.79	0.85
mu2-mu1	-14.9	-12.01	-9.06	1.48
mu3-mu1	28.47	32.04	35.6	1.83
mu4-mu1	10.19	14.01	17.9	1.96
mu3-mu2	41.12	44.05	46.95	1.51
mu4-mu2	22.83	26.02	29.21	1.66
mu4-mu3	-21.8	-18.03	-14.4	1.92
sigma2-sigma1	-5.6	-3.52	-1.57	1.02
sigma3-sigma1	-1.94	0.32	2.53	1.15
sigma4-sigma1	-1.14	1.15	3.51	1.18
sigma3-sigma2	1.83	3.84	5.85	1.03
sigma4-sigma2	2.7	4.67	6.8	1.05
sigma4-sigma3	-1.48	0.83	3.13	1.17
delta12	2.4	3.2	3.96	0.4
delta13	-8.92	-8.01	-7.05	0.48
delta14	-4.42	-3.46	-2.5	0.49
delta23	-12.55	-11.67	-10.77	0.45
delta24	-7.65	-6.79	-5.91	0.45
delta34	3.52	4.43	5.37	0.48

The results reveal multiple insights now. To support the interpretation, we first produce visualisations of the results via the `anovaplot()` function:

```
R> anovaplot(resRunners)
```

Figure 3 shows the plots produces by the above call to `anovaplot()`. The first row shows the posterior distributions of the effect sizes δ_{12}, δ_{13} and δ_{23}. The second row below is the analysis based on the ROPE, which partitions the posterior probability mass into the standard ROPES for effect sizes according to Cohen (1988).

Thus, we can see that for δ_{12} – which equals the effect size between female runners and male runners – there is a very large effect with posterior mean 3.2 and 95% credible interval $[2.402, 3.96]$, confirmed by the fact that 100% of the posterior probability mass are located inside the ROPE of a large effect according to Cohen (1988) (which includes values ≥ 0.8). Based on the results, the posterior probability of a large effect between female and male runners given the data is one, which means female runners have a faster heart beat after exercising six minutes than male runners.

To check if this effect exists also in the control groups, we compare the posterior of δ_{34}, corresponding to the effect size between the female and male controls. The results are given in the right plot of the third and fourth row in 3 and show that also in the control groups the effect is present. Here, the effect size is estimated to be even larger than for the runner groups with a posterior mean of 4.427 and a 95% credible interval $[3.517, 5.366]$. Thus, regular running seems to reduce the observed heartbeat differences between males and females in the form of a large effect. We could proceed this way and compare all other groups, too.

To check the model fit, we use the `post.pred.check` function, which performs a posterior predictive check against the observed data by drawing `reps` samples from the posterior distribution and visualizing the original data's density with density overlays for the reps sampled posterior predictive densities of the data:

```
post.pred.check(anovafit = resRunners, ngroups = 4, out = hr$Heart.Rate ,
reps = 50, eta = c(1/4,1/4,1/4,1/4))
```

The argument `anovafit` takes the resulting dataframe of the `bayes.anova` function as input, the

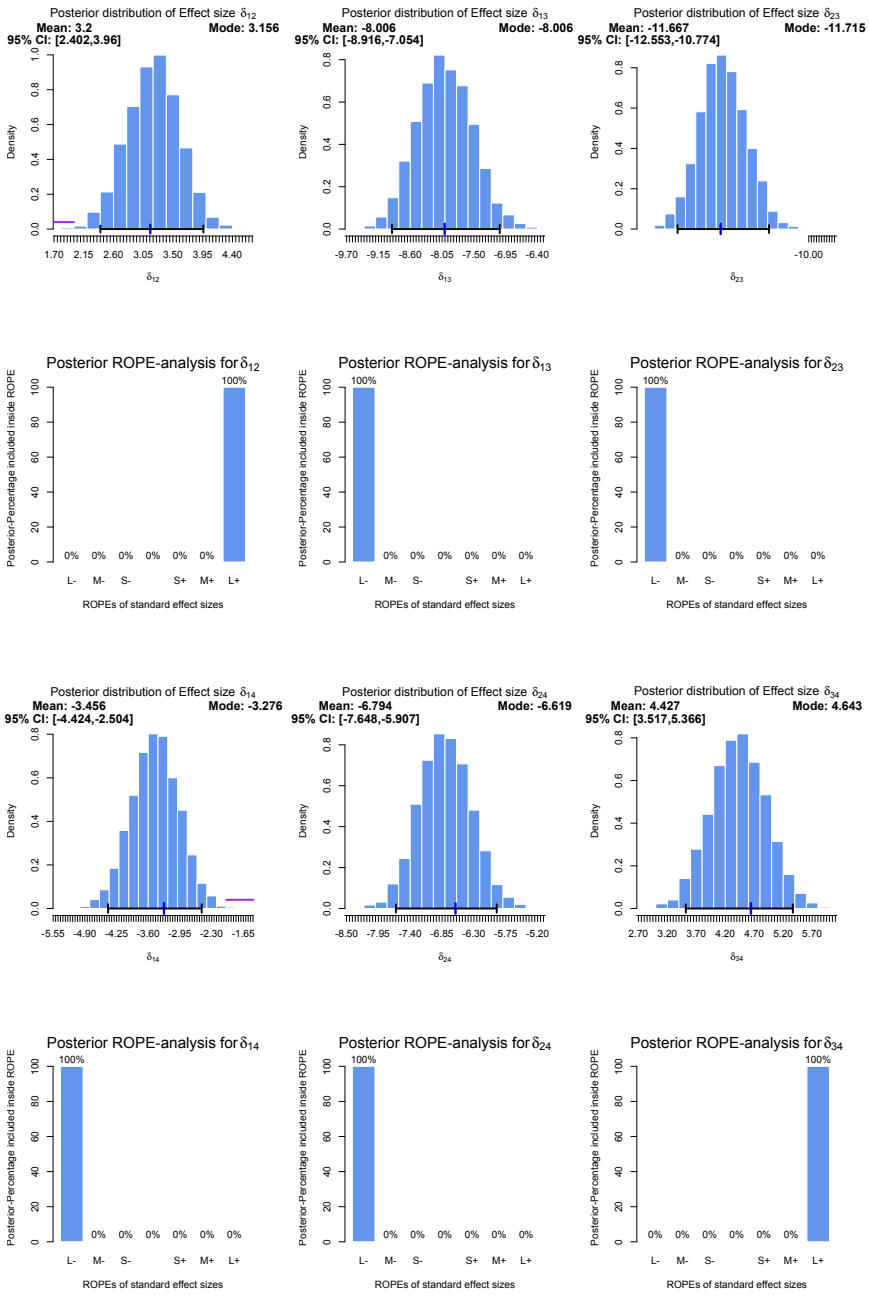

Figure 3: Visualisations of the results for the `Heart rate` dataset using the `anovaplot()` function in **bayesanova**, showing (1) the resulting posterior distributions of the effect sizes between each pair of groups (first and third row) and (2) the posterior ROPE-analysis for each group comparison (second and fourth row)

number of groups is specified in `ngroups`, `out` is the vector of all data originally observed (no matter which group), `reps` is the number of posterior predictive density overlays desired, and `eta` is the vector of weights used in the Gaussian mixture. Here, as all four groups include 200 participants, each weight is 1/4. The resulting posterior predictive check is shown in the left plot of 4, and indicates that while there is some overdispersion in the center of the posterior predictive distributions simulated, the overall fit seems reasonable.

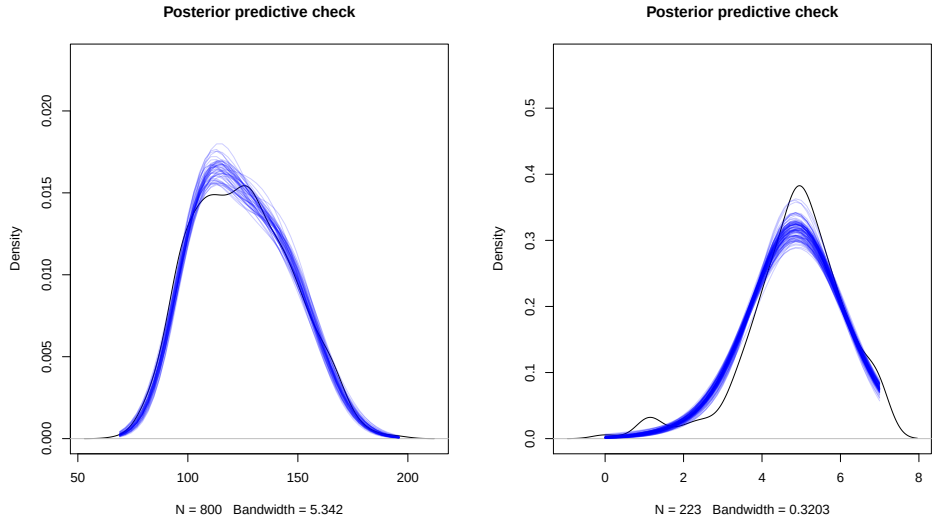

Figure 4: Posterior predictive checks using the `post.pred.check` function; left: For the `runners` dataset; right: For the `feelings` dataset; in both cases, results show that the overall fit of the Gaussian mixture model is reasonable

Pleasantness ratings after watching artistic or nude pictures

This example uses data from a study conducted by Balzarini et al. (2017), in which men and women's feelings towards their partners after watching either erotic or abstract art pictures were analysed. The study was published in the *Journal of Experimental Social Psychology*, and also the average pleasantness obtained from viewing the pictures was studied, as one of the research questions was whether men and women rate pleasantness of the pictures differently for nude and abstract art. This leads to a 2×2 factorial ANOVA for the variables gender and picture type, coded as `Gender` and `Condition` in the dataframe.

First, data is loaded and split into the four groups of interest:

```
R> feelings=read.csv("feelings.csv",sep=",")
R> head(feelings)

  Gender Age  RelLen    Condition PartnerAttractiveness
1   Male  43  3.7500        Nudes                    21
2 Female  26  3.0000        Nudes                    19
3 Female  35  5.2500 Abstract Art                    27
4 Female  31  2.0000 Abstract Art                    22
5 Female  23  4.0000 Abstract Art                    27
6   Male  36 19.9167        Nudes                    16
  LoveForPartner AveragePleasantness
1             76              5.9375
2             66              4.7500
3            103              6.2500
4             76              5.5625
5            109              2.3750
6             98              5.1250

R> femaleArtistic = (feelings %>% filter(Gender=="Female") %>%
+  filter(Condition=="Abstract Art"))$AveragePleasantness
R> maleArtistic = (feelings %>% filter(Gender=="Male") %>%
+  filter(Condition=="Abstract Art"))$AveragePleasantness
R> femaleNude = (feelings %>% filter(Gender=="Female") %>%
+  filter(Condition=="Nudes"))$AveragePleasantness
R> maleNude = (feelings %>% filter(Gender=="Male") %>%
+  filter(Condition=="Nudes"))$AveragePleasantness
```

Second, the model assumption of normality in each group is checked:

```
R> assumption.check(femaleArtistic, maleArtistic, femaleNude, maleNude)
```

```
1: In assumption.check(femaleArtistic, maleArtistic, femaleNude, maleNude) :
  Model assumption of normally distributed data in each group is violated.
 All results of the Bayesian ANOVA based on a Gaussian mixture
 could therefore be unreliable and not trustworthy.
2: In assumption.check(femaleArtistic, maleArtistic, femaleNude, maleNude) :
  Run further diagnostics (like Quantile-Quantile-plots) to check if the
  Bayesian ANOVA can be expected to be robust to the violations of normality
```

This time the function gives a warning, that there are violations of the distributional assumptions. Investigating the results leads to the conclusion that data in the fourth group deviate from normality, shown in 5 in the QQ-plot. Still, as all other groups show no strong deviations from normality, we proceed and are cautious when drawing inferences including any statements involving the fourth group.

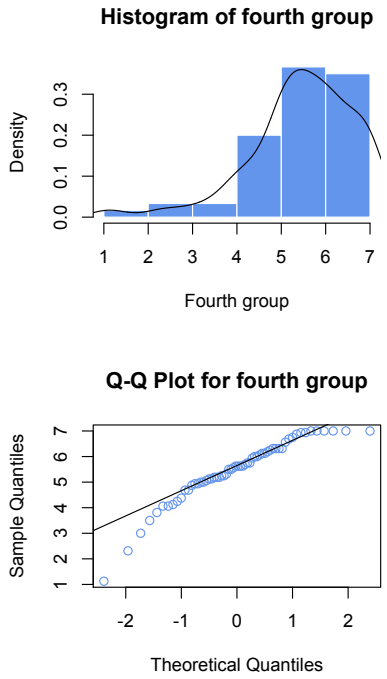

Figure 5: Histogram and quantile-quantile plot for the fourth group in the `feelings` dataset, showing that the assumption of normality is violated

Keeping this in mind, the Bayesian ANOVA is run now with default hyperparameters:

```
R> set.seed(42)
R> resFeelings = bayes.anova(first = femaleArtistic, second = maleArtistic,
+   third = femaleNude, fourth = maleNude)
```

Parameter	LQ	Mean	UQ	Std.Err
mu1	4.86	4.9	4.95	0.02
mu2	4.62	4.66	4.69	0.02
mu3	4.07	4.2	4.34	0.07
mu4	5.42	5.47	5.53	0.03
sigma1	0.98	1.16	1.4	0.11
sigma2	0.86	1.02	1.21	0.09
sigma3	1.34	1.66	2.06	0.19
sigma4	1.06	1.26	1.52	0.12
mu2-mu1	-0.31	-0.25	-0.19	0.03
mu3-mu1	-0.85	-0.7	-0.56	0.07
mu4-mu1	0.5	0.57	0.64	0.04
mu3-mu2	-0.6	-0.46	-0.32	0.07
mu4-mu2	0.75	0.81	0.87	0.03

```
|mu4-mu3       |1.12  |1.27  |1.41  |0.07    |
|sigma2-sigma1 |-0.43 |-0.14 |0.12  |0.14    |
|sigma3-sigma1 |0.1   |0.49  |0.95  |0.21    |
|sigma4-sigma1 |-0.21 |0.1   |0.42  |0.16    |
|sigma3-sigma2 |0.27  |0.64  |1.07  |0.21    |
|sigma4-sigma2 |-0.04 |0.24  |0.55  |0.15    |
|sigma4-sigma3 |-0.84 |-0.39 |0.01  |0.22    |
|delta12       |0.18  |0.24  |0.29  |0.03    |
|delta13       |0.47  |0.6   |0.73  |0.07    |
|delta14       |-0.58 |-0.52 |-0.44 |0.04    |
|delta23       |0.27  |0.41  |0.53  |0.06    |
|delta24       |-0.84 |-0.76 |-0.69 |0.04    |
|delta34       |-1.2  |-1.07 |-0.91 |0.07    |
```

The results show that differences are now much more subtle than in the previous examples. From the results one can spot that the means in the first three groups are located nearer to each other than in the previous examples, and the fourth group differs more strongly from the first three. The standard deviations do not differ a lot between groups, and the magnitude of the posterior effect sizes is now smaller, too. To investigate the effect sizes, visualisations are produced first:

```
R> anovaplot(resFeelings)
```

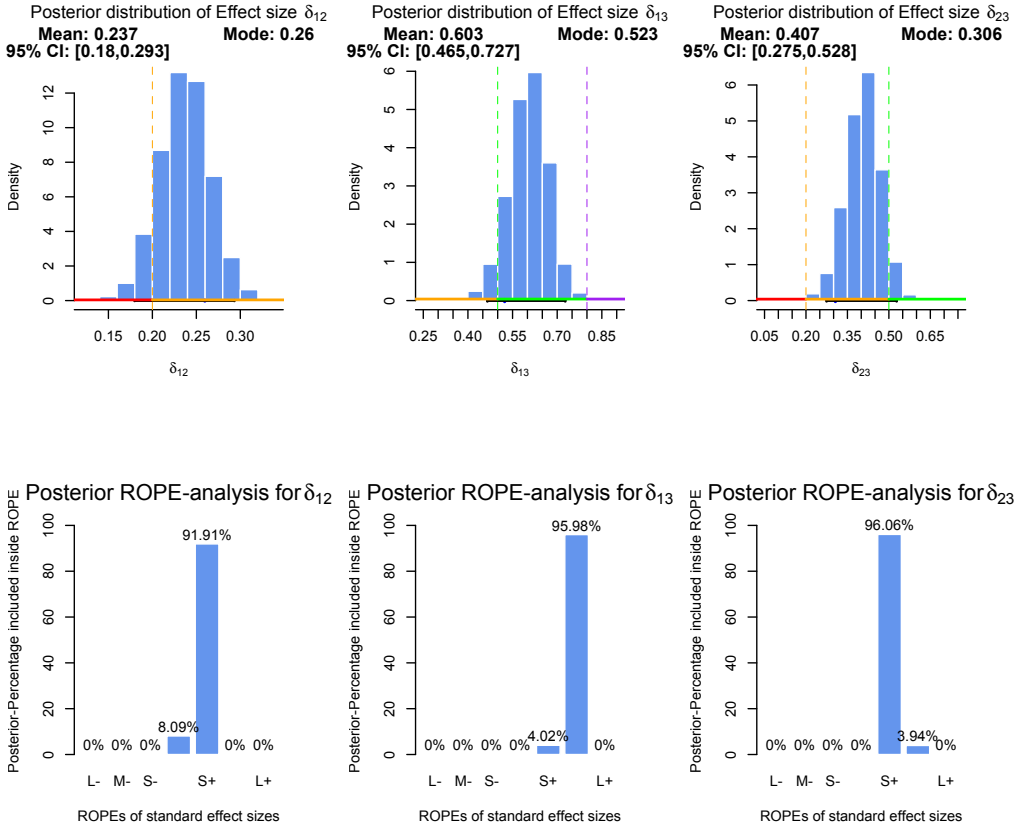

Figure 6: Visualisations of the posterior effect sizes for the `feelings` dataset using the `anovaplot()` function in **bayesanova**, showing which effects are most probable a posteriori based on a ROPE-analysis for each pair of groups

Figure 6 shows the plots produces by the above call to `anovaplot()`. The two left plots show that with 91.91% probability there is a small effect between the first and second group given the data, which are the female and male artistic pictures groups. Therefore, with large probability females rate artistic pictures more pleasant than males, where the effect size itself is small. Still, we could argue that there is nonnegligible probability of 8.09% that there is no effect at all and therefore not draw any conclusion depending on the posterior probability we require.

The middle two plots in 6 show the effect between the female artistic and female nude picture groups. We can see that based on the posterior distribution of δ_{13}, with 95.98% there is a medium effect between the two groups given the data. Females rate artistic pictures therefore with a probability near certainty as more pleasant than nude pictures, where the effect size in terms of standardized differences between ratings is medium.

The right two plots in 6 show the effect between the male artistic and female nude groups. The posterior reveals that 96.06% indicate a small effect, which could be interpreted as the fact that males rate artistic pictures even more pleasant than females rate nude pictures, but the effect size is only small and the remaining 3.94% posterior probability indicate that there is even a medium effect.

Figure 7 shows the effects which include the fourth group. Due to the violations of distributional assumptions one need to be cautious now, as the results could be deterred. Still, the two right plots show the effect size between the female and male nude groups, and indicate that the full posterior (100%) signals a large negative effect. This means, males rate the pleasantness of nude pictures much higher than females. Still, the result (as well as the results for δ_{14} and δ_{24}) are questionable due to the violation of model assumptions, so we do not proceed here.

The posterior predictive check in the right plot of 4 obtained via

```
post.pred.check(anovafit = resFeelings, ngroups = 4, out = feelings$AveragePleasantness,
reps = 100, eta = c(58/223,64/223,41/223,60/223))
```

shows that the overall fit seems reasonable, although there is some room for improvement in the range of average pleasantness ratings between zero and two, and in the peak between average pleasantness ratings of four and six. Subdividing the data even further and refitting the ANOVA model with a higher number of components would be an option to improve the fit. Alternatively, one could discuss the prior hyperparameters chosen here.

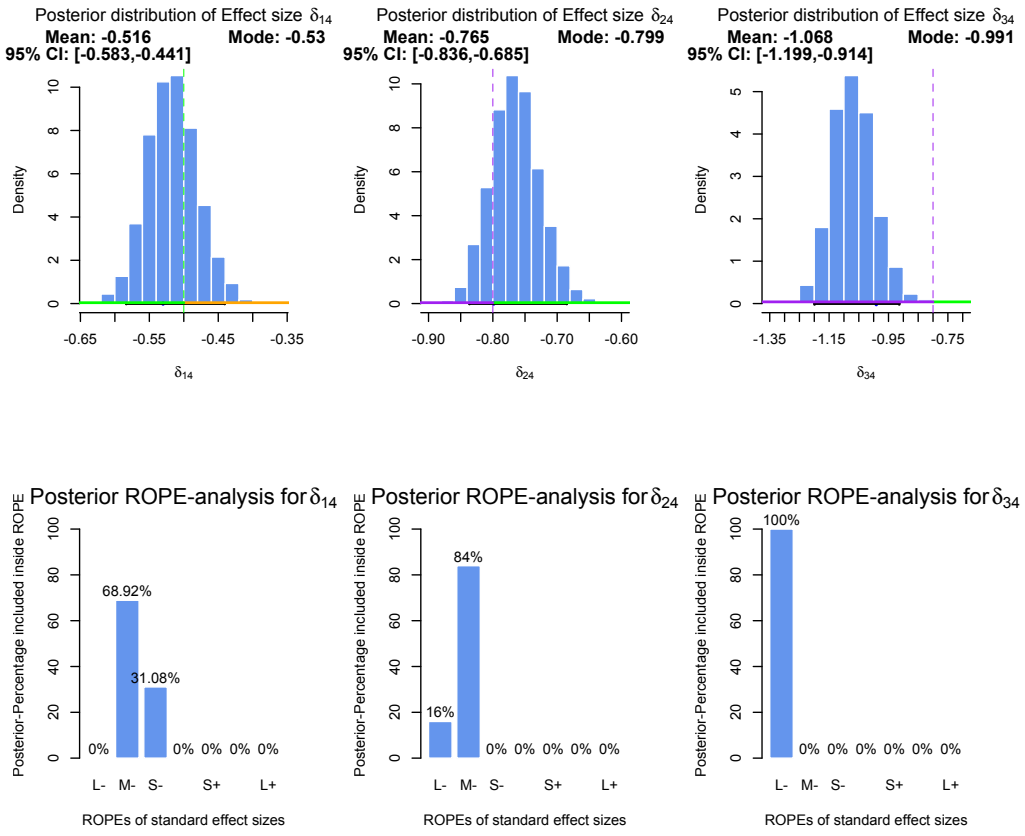

Figure 7: Visualisations of the posterior effect sizes for the `feelings` dataset using the `anovaplot()` function in **bayesanova**, showing which effects are most probable a posteriori based on a ROPE-analysis for each pair of groups

A solution via a traditional ANOVA in this case would yield:

```
R> summary(aov(AveragePleasantness ~ Gender * Condition, data = feelings))

                 Df Sum Sq Mean Sq F value   Pr(>F)
Gender            1  10.63  10.629   7.605  0.00631 **
Condition         1   1.27   1.267   0.906  0.34210
Gender:Condition   1  31.15  31.155  22.291 4.18e-06 ***
Residuals        219 306.09   1.398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, the condition is not significant, but the interaction is. The above analysis via the Bayesian mixture ANOVA made this more explicit: The posteriors for each combination of gender and condition were derived via MCMC, leading for example to the conclusion that females rate artistic picture as more pleasant than nude pictures with 95.98% probability for a medium effect size and 4.02% for a small effect size.

A solution via a Bayes factor based ANOVA would yield:

```
R> library(BayesFactor)
R> set.seed(42)
R> feelings$Gender = factor(feelings$Gender)
R> feelings$Condition = factor(feelings$Condition)
R> bfFeelings = anovaBF(AveragePleasantness ~ Gender * Condition, data = feelings)

Bayes factor analysis
--------------
[1] Gender                              : 3.727898   +- 0%
[2] Condition                           : 0.2532455  +- 0.01%
[3] Gender + Condition                  : 0.822604   +- 1.01%
[4] Gender + Condition + Gender:Condition : 3048.134   +- 1.14%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

The conclusions drawn in this case are that the model including both gender, the condition and the interaction between both is most favourable due to the huge Bayes factor of $BF(\mathcal{M}_4, \mathcal{M}_0) = 3048.134$. Here too, the information is quite limited compared to the detailed analyses we could obtain from the Bayesian ANOVA based on the Gaussian mixture model above.

Amyloid concentrations and cognitive impairments

This example uses data from medical research about Alzheimer's disease. Amyloid-beta (Abeta) is a protein fragment which has been linked frequently to Alzheimer's disease. Autopsies from a sample of Catholic priests included measurements of Abeta (pmol/g tissue from the posterior cingulate cortex) from three groups: subjects who had exhibited no cognitive impairment before death, subjects who had exhibited mild cognitive impairment, and subjects who had mild to moderate Alzheimer's disease. The original study results were published by Pivtoraiko et al. (2015) in the journal *Neurobiology of Aging* and are used here.

The `Amyloid` dataset is available in the **Stat2Data** package (Cannon et al., 2019) and includes a group indicator `Group`, which takes either one of three values: `mAD`, which classifies a subject as having had mild Alzheimer's disease, `MCI`, which is a mild cognitive impairment and `NCI`, which is no cognitive impairment. Also, the amount of Amyloid-beta from the posterior cingulate cortex is given in pmol per gram tissue in the variable `Abeta`.

After loading and splitting the data into the three groups, we run the `assumption.check()` function:

```
R> library(Stat2Data)
R> data(Amyloid)
R> head(Amyloid)

  Group Abeta
1   NCI   114
2   NCI    41
3   NCI   276
```

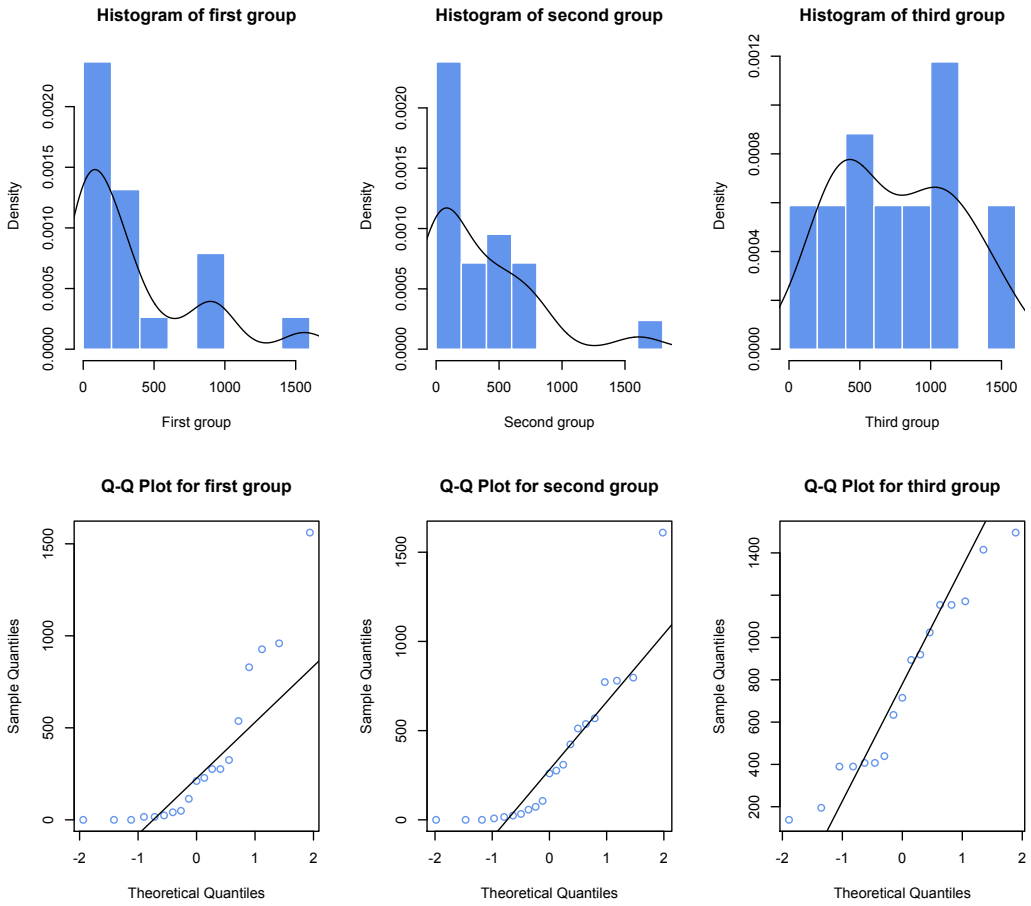

Figure 8: Model assumption checks for the `Amyloid` dataset using the `assumption.check()` function in **bayesanova**, showing that the assumption of a Gaussian mixture model is violated

```
4   NCI     0
5   NCI    16
6   NCI   228

R> NCI = (Amyloid %>% filter(Group=="NCI"))$Abeta
R> MCI = (Amyloid %>% filter(Group=="MCI"))$Abeta
R> mAD = (Amyloid %>% filter(Group=="mAD"))$Abeta
R> assumption.check(NCI, MCI, mAD)

1: In assumption.check(NCI, MCI, mAD) :
  Model assumption of normally distributed data in each group is violated.
 All results of the Bayesian ANOVA based on a multi-component Gaussian
 mixture could therefore be unreliable and not trustworthy.
2: In assumption.check(NCI, MCI, mAD) :
  Run further diagnostics (like Quantile-Quantile-plots) to check if the
  Bayesian ANOVA can be expected to be robust to the violations of normality
```

The results in 8 clearly show that the model assumptions are violated. Therefore, it is not recommended to run a Bayesian ANOVA in this case. A solution via a traditional ANOVA or via a Bayes factor based ANOVA would not proceed at this point, too.

A small simulation study – Recapturing simulation parameters of synthetic datasets

The next example is more in the veins of a simulation approach. We simulate three-, four-, five- and six-component Gaussian mixtures with increasing means $\mu_j := j$ and $\sigma_j = 1$. Therefore, the theoretical parameter values as well as the differences in means and standard deviations and the effect sizes δ_{lr} are known $\forall l, r$. We simulate 500 datasets with $n = 50$ observations in each group for each Gaussian

mixture above, and run the Bayesian ANOVA with default hyperparameters, that is 10000 Gibbs steps with 5000 burn-in steps, 95% credibility level and standard deviation output. Histograms of the posterior means for all parameters are shown in 9.

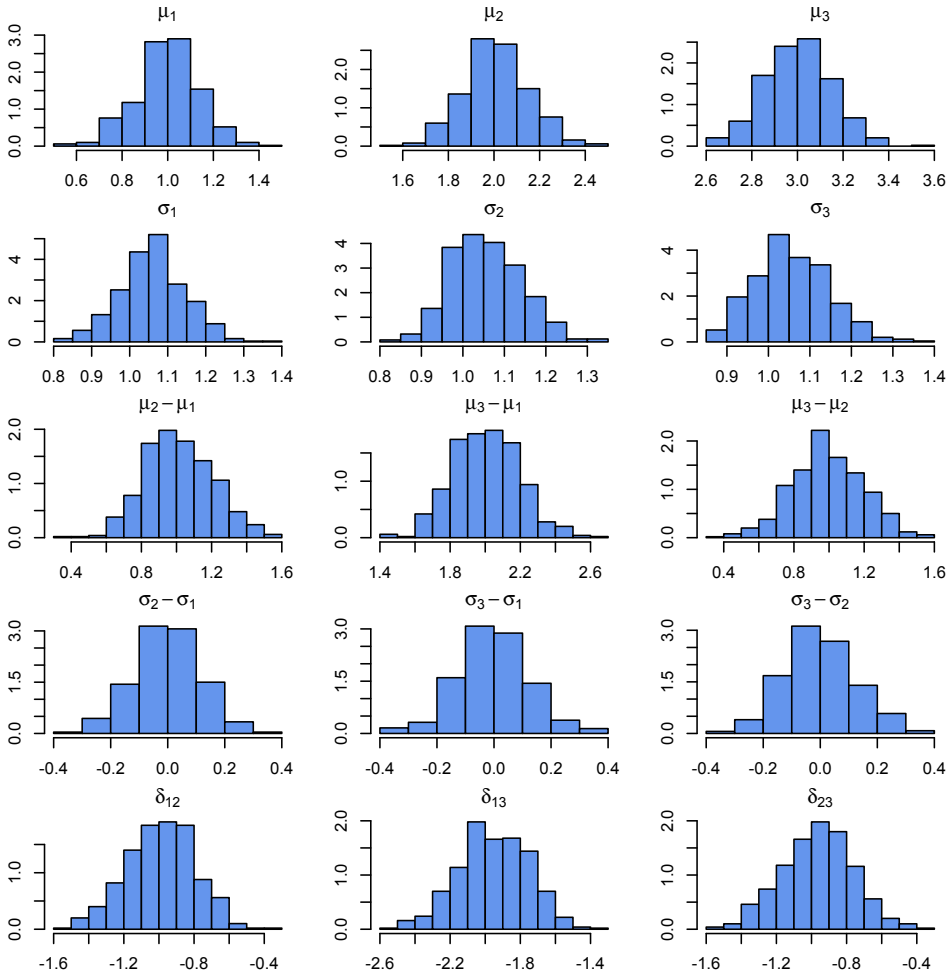

Figure 9: Recapturing simulation parameters of synthetic datasets with `bayes.anova()`, showing that the Gibbs sampler yields consistent estimates of the underlying effect sizes

The results clearly show that even for 500 simulated datasets the true parameters $\mu_j = j$ and $\sigma_j = 1$ are recaptured for small sample sizes like $n = 50$ in each group. Also, the differences in means $\mu_l - \mu_r, l \neq r$ are near one, and the differences in standard deviations $\sigma_l - \sigma_r, l \neq r$ are near zero. The effect sizes $\delta_{lr}, l \neq r$ also are recaptured as expected. More details about the theoretical properties of the procedure, especially the derivation of the Gibbs sampler for the two-group case can be found in Kelter (2020c, 2021d). Note that increasing sample sizes in the groups will yield consistent estimates as a result of MCMC theory Robert and Casella (2004).

Conclusion

This paper introduces **bayesanova**, an R package for conducting a Bayesian analysis of variance based on MCMC in a Gaussian mixture distribution with known allocations. The Bayesian ANOVA implemented in **bayesanova** is based on Gibbs sampling and supports up to six distinct components, which covers the typical range of ANOVAs used in empirical research.

The package provides four functions to check the model assumptions, run the Bayesian ANOVA, visualize the results and check the posterior fit. All functions have a variety of optional parameters to adapt them to a specific workflow or goal. Also, convergence issues can be detected via the built-in convergence diagnostics of all MCMC results in the `anovaplot()` function and it is possible to

post-process the results delivered as raw Markov chain draws by `bayes.anova`, for example via the R package **bayestestR** (Makowski et al., 2019a).

In the paper, multiple examples from medical and psychological research using real datasets were provided, showing the richness of information provided by the proposed procedure. Also, while explicit testing (for example via Bayes factors) is not implemented as standard output, it is worth noting that computing Bayes factors numerically based on the Gaussian mixture model is possible for example by using numerical techniques such as the Savage-Dickey density ratio (Kelter, 2021a; Wagenmakers et al., 2010; Dickey and Lientz, 1970; Verdinelli and Wasserman, 1995). However, the focus of explicit hypothesis testing is replaced in the default output of the procedure by estimation of the effect sizes between groups (or component density parameters) under uncertainty. If hypothesis testing is needed, the implemented ROPE can be used for rejecting a hypothesis based on interval hypothesis tests – compare Kelter (2021b), Linde et al. (2020) and Kruschke (2018) – or by using external packages like **bayestestR** (Makowski et al., 2019a) in conjunction with the raw samples provided by `bayes.anova`. Also, other indices like the probability of direction (Makowski et al., 2019b) or the MAP-based p-value (Mills, 2018) can be obtained via the package **bayestestR** (Makowski et al., 2019a) if hypothesis testing is desired, for an overview see Kelter (2021a). To offer users the freedom of choice for their preferred statistical evidence measure, only a ROPE-based estimate of the maximum a posteriori effect size δ is provided in **bayesanova**.

A small simulation study showed for the case of three-component Gaussian mixtures, that the provided MCMC algorithm precisely captures the true parameter values. Similar results hold for the four- or more-component case, as can easily be checked by adapting the provided R code.

In summary, the **bayesanova** package provides a novel and easy to apply alternative to existing packages like **stats** (R Core Team, 2020) or **BayesFactor** (Morey and Rouder, 2018), which implement the traditional frequentist ANOVA and Bayesian ANOVA models based on the Bayes factor.

Future plans include to add prior predictive checks and up to 12-component support, allowing for 2×6 Bayesian ANOVAs. Also, nonparametric mixtures could be applied in the case the model assumptions are violated, but therefore first theoretical results are necessary.

Bibliography

R. N. Balzarini, K. Bobson, K. Chin, and L. Campbell. Does exposure to erotica reduce attraction and love for romantic partners in men? Independent replications of Kenrick, Gutierres, and Goldberg (1989). *Journal of Experimental Social Psychology*, 70:191–197, 2017. [p71]

D. J. Benjamin, J. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. [p58]

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985. ISBN 9781441930743. [p60]

D. V. D. Bergh, J. V. Doorn, M. Marsman, K. N. Gupta, A. Sarafoglou, G. Jan, A. Stefan, A. Ly, and M. Hinne. A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP. *psyarxiv preprint, https://psyarxiv.com/spreb*, 2019. [p58, 62]

A. Cannon, G. Cobb, B. Hartlaub, J. Legler, R. Lock, T. Moore, A. Rossman, and J. Witmer. Stat2Data: Datasets for Stat2, 2019. [p61, 75]

B. Carlin and T. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall, CRC Press, Boca Raton, sep 2009. [p58]

B. Carpenter, J. Guo, M. D. Hoffman, M. Brubaker, A. Gelman, D. Lee, B. Goodrich, P. Li, A. Riddell, and M. Betancourt. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76 (1):1–32, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. [p59, 62]

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, Hillsdale, N.J, 2nd edition, 1988. ISBN 978-0-8058-0283-2. [p60, 64, 66, 69]

D. Colquhoun. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12), 2017. ISSN 20545703. doi: 10.1098/rsos.171085. [p58]

D. Colquhoun. The False Positive Risk: A Proposal Concerning What to Do About p-Values. *The American Statistician*, 73(sup1):192–201, 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1529622. [p58]

J. M. Dickey and B. P. Lientz. The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *Annals of Mathematical Statistics*, 41(1):214–226, 1970. ISSN 0003-4851. doi: 10.1214/AOMS/1177697203. [p78]

T. S. Donaldson. Power of the F-test for nonnormal distributions and unequal error variances, 1966. [p59]

C. Dong and M. Wedel. BANOVA: Hierarchical Bayesian ANOVA Models, 2019. URL https://cran.r-project.org/package=BANOVA. [p62]

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer, New York, 2006. ISBN 9781441921949. [p62, 63]

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, feb 2019. ISSN 09641998. doi: 10.1111/rssa.12378. URL http://doi.wiley.com/10.1111/rssa.12378. [p66]

A. Gelman and S. P. Brooks. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998. [p66]

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, 2006. ISBN 9780511790942. doi: 10.1017/CBO9780511790942. URL http://ebooks.cambridge.org/ref/id/CBO9780511790942. [p61]

JASP Team. Jeffreys Awesome Statistics Package (JASP). *https://jasp-stats.org/*, 2019. URL https://jasp-stats.org/. [p58]

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 3rd edition, 1961. ISBN 0-19-850368-7. [p60]

K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. Testing hypotheses via a mixture estimation model. *arXiv preprint, https://arxiv.org/abs/1412.2044*, pages 1–37, 2014. ISSN 00237205. doi: 10.16373/j.cnki.ahr.150049. [p58, 60]

R. Kelter. Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology*, 20 (1), 2020a. ISSN 1471-2288. doi: 10.1186/s12874-020-00980-6. [p60]

R. Kelter. Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, 20(88), 2020b. doi: https://doi.org/10.1186/s12874-020-00968-2. [p60]

R. Kelter. bayest: An R Package for effect-size targeted Bayesian two-sample t-tests. *Journal of Open Research Software*, 8(14), 2020c. doi: https://doi.org/10.5334/jors.290. [p64, 65, 77]

R. Kelter. How to Choose between Different Bayesian Posterior Indices for Hypothesis Testing in Practice. *Multivariate Behavioral Research*, (in press):1–29, 2021a. ISSN 0027-3171. doi: 10.1080/00273171.2021.1967716. URL https://www.tandfonline.com/doi/full/10.1080/00273171.2021.1967716. [p78]

R. Kelter. Bayesian Hodges-Lehmann tests for statistical equivalence in the two-sample setting: Power analysis, type I error rates and equivalence boundary selection in biomedical research. *BMC Medical Research Methodology*, 21(1), 2021b. ISSN 1471-2288. doi: 10.1186/s12874-021-01341-7. [p78]

R. Kelter. On the Measure-Theoretic Premises of Bayes Factor and Full Bayesian Significance Tests: a Critical Reevaluation. *Computational Brain & Behavior*, (online first):1–11, 2021c. ISSN 2522-0861. doi: 10.1007/s42113-021-00110-5. [p60]

R. Kelter. A new Bayesian two-sample t-test and solution to the Behrens-Fisher problem based on Gaussian mixture distributions. *Statistics in Biosciences*, (in press), 2021d. [p64, 65, 77]

J. K. Kruschke. Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, 142(2):573–603, 2013. ISSN 1939-2222. doi: 10.1037/a0029146. [p58]

J. K. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, Oxford, 2nd edition, 2015. ISBN 9780124058880. doi: 10.1016/B978-0-12-405888-0.09999-2. [p58, 59, 60, 61, 62, 64]

J. K. Kruschke. Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018. doi: 10.1177/2515245918771304. [p58, 64, 78]

M. Linde, J. Tendeiro, R. Selker, E.-J. Wagenmakers, and D. van Ravenzwaaij. Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. *psyarxiv preprint, https://psyarxiv.com/bh8vu*, 2020. [p78]

D. Makowski, M. Ben-Shachar, and D. Lüdecke. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, 4(40): 1541, 2019a. doi: 10.21105/joss.01541. [p78]

D. Makowski, M. S. Ben-Shachar, S. H. A. Chen, and D. Lüdecke. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10:2767, 2019b. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02767. [p58, 78]

R. McElreath and P. E. Smaldino. Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE*, 10(8):1–16, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0136088. [p58]

J. A. Mills. Objective Bayesian Precise Hypothesis Testing. Technical report, University of Cincinnati, 2018. [p78]

D. S. Moore, G. P. McCabe, and B. A. Craig. *Introduction to the practice of statistics*. W. H. Freeman, New York, 9th edition, 2012. ISBN 1319013384. [p68]

R. D. Morey and J. N. Rouder. BayesFactor: Computation of Bayes Factors for Common Designs. *R package version 0.9.12-4.2*, 2018. URL https://cran.r-project.org/package=BayesFactor. [p61, 78]

V. N. Pivtoraiko, E. E. Abrahamson, S. E. Leurgans, S. T. DeKosky, E. J. Mufson, and M. D. Ikonomovic. Cortical pyroglutamate amyloid-β levels and cognitive decline in Alzheimer's disease. *Neurobiology of Aging*, 36(1):12–19, jan 2015. ISSN 15581497. doi: 10.1016/j.neurobiolaging.2014.06.021. URL http://www.ncbi.nlm.nih.gov/pubmed/25048160http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4268150. [p75]

M. Plummer. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003. doi: 10.1111/j.2517-6161.1996.tb02070.x. [p59, 60, 61, 62]

R Core Team. R: A Language and Environment for Statistical Computing, 2020. URL https://www.r-project.org/. [p61, 78]

A. Raftery. Hypothesis testing and model selection. In W. Gilks, D. Spiegelhalter, and S. Richardson, editors, *Markov Chain Monte Carlo in Practice*, pages 163–188. Chapman & Hall, London, 1996. [p65]

C. R. Rao and M. M. Lovric. Testing point null hypothesis of a normal mean and the truth: 21st Century perspective. *Journal of Modern Applied Statistical Methods*, 15(2):2–21, 2016. ISSN 15389472. doi: 10.22237/jmasm/1478001660. [p60]

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004. ISBN 1441919392. [p62, 65, 77]

C. P. Robert. The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72(2009): 33–37, 2016. ISSN 10960880. doi: 10.1016/j.jmp.2015.08.002. [p58, 60]

J. Rochon, M. Gondan, and M. Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12, 2012. ISSN 14712288. doi: 10.1186/1471-2288-12-81. [p58]

J. N. Rouder, R. D. Morey, P. L. Speckman, and J. M. Province. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5):356–374, 2012. ISSN 00222496. doi: 10.1016/j.jmp.2012.08.001. URL http://dx.doi.org/10.1016/j.jmp.2012.08.001. [p58, 59, 60, 61, 64, 65]

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. URL http://www.jstor.org/stable/2333709. [p66]

Stan Development Team. RStan: the R interface to Stan. *R package version 2.19.3*, 2020. URL http://mc-stan.org/. [p59, 62]

J. N. Tendeiro and H. A. Kiers. A Review of Issues About Null Hypothesis Bayesian Testing. *Psychological Methods*, 24(6):774–795, 2019. ISSN 1082989X. doi: 10.1037/met0000221. [p60]

M. Tiku. Power Function of the F-test Under Non-Normal Situations. *Journal of the American Statistical Association*, 66(336), 1971. [p59]

J. van Doorn, D. van den Bergh, U. Bohm, F. Dablander, K. Derks, T. Draws, N. J. Evans, Q. F. Gronau, M. Hinne, S. Kucharský, A. Ly, M. Marsman, D. Matzke, A. Raj, A. Sarafoglou, A. Stefan, J. G. Voelkel, and E.-J. Wagenmakers. The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. *psyarxiv preprint, https://psyarxiv.com/yqxfr*, 2019. doi: 10.31234/osf.io/yqxfr. [p58]

I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995. ISSN 1537274X. doi: 10.1080/01621459.1995.10476554. [p78]

E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman. Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60(3):158–189, 2010. ISSN 00100285. doi: 10.1016/j.cogpsych.2009.12.001. [p78]

R. L. Wasserstein and N. A. Lazar. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, 2016. ISSN 0003-1305. doi: 10.1080/00031305.2016.1154108. [p58]

R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a World Beyond "p<0.05". *The American Statistician*, 73(sup1):1–19, 2019. ISSN 0003-1305. doi: 10.1080/00031305.2019.1583913. [p58]

A. Zellner. Introduction. In A. Zellner and J. B. Kadane, editors, *Bayesian Analysis in Econometrics and Statistics : Essays in Honor of Harold Jeffreys*, chapter 1. Elsevier North-Holland, Amsterdam, 1980. ISBN 978-0444852700. [p60]

Riko Kelter
University of Siegen, Department of Mathematics
Walter-Flex-Street 3
57072, Siegen
Germany
ORCiD: 0000-0001-9068-5696
riko.kelter@uni-siegen.de

Appendix

Details on the F-statistic in frequentist ANOVA

After observing the data, the following quantities are calculated: For group j, $j = 1, ..., k$, I_j experimental units are observed and the empirical mean $m_j = 1/I_j \sum_{l=1}^{I_j} y_{lj}$ and empirical variance $s_j^2 = 1/(I_j - 1) \sum_{l=1}^{I_j} (y_{lj} - m_j)^2$ are calculated (data is assumed to be listed in a table where the groups correspond to the columns). The sum $\Sigma_{i \in I_j} y_{ij}$ and the sum of squares $\Sigma_{i \in I_j} (y_{ij})^2$ are calculated, to partition the variance into treatment and error sum of squares

$$SS_{Treatment} := \sum_{j=1}^{k} I_j (m_j - m)^2 \qquad SS_{Error} := \sum_{j=1}^{k} (I_j - 1) s_j^2 \tag{16}$$

$$SS_{Total} := \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - m)^2 \tag{17}$$

where $m := 1/k \sum_{j=1}^{k} m_j$. Standard calculus yields that these sums of squares can be calculated as:

$$SS_{Treatment} := \sum_{j=1}^{k} \frac{(\sum_i y_{ij})^2}{I_j} - \frac{(\sum_j \sum_i y_{ij})^2}{I} \qquad SS_{Error} := \sum_{j=1}^{k} \sum_i y_{ij}^2 - \sum_j \frac{(\sum_i y_{ij})^2}{I_j} \tag{18}$$

$$SS_{Total} := \sum_{j=1}^{k} \sum_i y_{ij}^2 - \frac{(\sum_j \sum_i y_{ij})^2}{I} \tag{19}$$

Using the corresponding degrees of freedom $DF_{Treatment} = k - 1$, $DF_{Error} = n - k$ and $DF_{Total} = n - 1$, the F-statistic is defined as

$$F = \frac{MS_{Treatment}}{MS_{Error}} \tag{20}$$

where

$$MS_{Treatment} := \frac{SS_{Treatment}}{DF_{Treatment}} \qquad MS_{Error} := \frac{SS_{Error}}{DF_{Error}} \tag{21}$$

using only the quantities defined above. If the F-statistic is larger than the α-quantile for significance level α, H_0 is rejected.