RESEARCH MASTER THESIS

Psychological Methods

Extending the Bayesian arsenal: An assumption check for normality

Koen Derks[1] | Johnny van Doorn[1] | Maarten Marsman[1]

[1]Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands

Correspondence
Koen Derks, Department of Psychology,
University of Amsterdam, Amsterdam,
Noord-Holland, 1018 WS, the Netherlands
Email: koen-derks@hotmail.com

Funding information
This work is made possible by the Research Masters thesis committee.

This thesis outlines the development and application of a new Bayesian assumption check for assessing deviations from normality in the distribution of a variable. The proposed method yields Bayes factors that quantify evidence for normality against non-normality, bi-modality and skewness. Results from simulation studies comparing Bayes factor behavior and sensitivity of the candidate models are presented, indicating that identifying these deviations is feasible and effortless. Furthermore, an investigation into the test's robustness to the prior distribution is presented and several practical implementations of the assumption checks are illustrated.

1 | INTRODUCTION

Deviations from normality occur frequently in psychological research. For example, reaction times are well-known to be non-normally distributed (Ulrich and Miller, 1993). Another example from psychology is the non-normal distribution of labor income (Diaz-Serrano, 2005). These cases can be hard to analyze due to the shape of their distribution. Most standard tests (ANOVA, t-tests, regression) rely heavily on normality in the distribution of their variables. So much even, that they are built around the assumption of normality, stating that the researcher has to make sure that the data roughly fit a bell curve shape. This is observable from the fact that almost all parametric frequentist statistical tests have an assumption of normality that, when not met, yields the test results invalid. When the normality assumption is violated, other tests, such as non-parametric need to be used to analyze the data. Normality tests are used for this purpose, to test if non-parametric tests are required. In the frequentist framework, detection tools for deviations from normality already exist in the form of the Shapiro-Wilk test (Shapiro and Wilk, 1965) and the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948; Öztuna et al., 2006). These frequentist normality tests compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation as the sample. The null hypothesis in these tests states that the sample distribution is normal. If the test is significant, it means that under the assumption of normality the data and more extreme are improbable. The Kolmogorov-Smirnov test, for example, is a normality test in which a normal cumulative distribution function is contrasted with the empirical distribution function of the data (Öztuna et al., 2006). A limitation of the Kolmogorov-Smirnov test is its high sensitivity to extreme values. The Lilliefors correction renders this test less conservative (Peat and Barton, 2008). However, it has been reported that the Kolmogorov-Smirnov test has low power and should not be seriously considered for testing normality (Steinskog et al., 2007). The Shapiro-Wilk test, on the other hand, is based on the correlation between the data and the corresponding normal scores (Peat and Barton, 2008) and provides higher power than the Kolmogorov-Smirnov test even after the Lilliefors correction (Steinskog et al., 2007). Some researchers have recommended the Shapiro-Wilk test as the best choice for testing the normality of data (Thode, 2002). However, both these tests share the same shortcomings as their frequentist family members, namely that they do not solely depend on observed data and that that they do not condition on the data, but on the null hypothesis (Ghasemi and Zahediasl, 2012). Another important frequentist shortcoming is that the *p* value does not quantify statistical evidence in a reliable way (Wagenmakers, 2007). On top of that, for small sample sizes, normality tests have little power to reject the null hypothesis and therefore small samples most often pass normality tests (Öztuna et al., 2006). For large sample sizes, significant results would be derived even in the case of a small deviation from normality (Öztuna et al., 2006; Field, 2009), although this small deviation will not affect the results of a parametric test (Öztuna et al., 2006). Alongside all of these, there are many more shortcomings to the general frequentist testing framework which makes the frequentist normality test unreliable. The introduction of Bayesian methods is a valuable alternative that counters the problems in the frequentist framework. For example, Bayesian analyses condition on the data rather than on the null hypothesis. In addition, the Bayes factor quantifies evidence for both the null hypothesis and the alternative hypothesis while indicating uncertainty. For an overview of (dis-)advantages of both paradigms, see Wagenmakers et al. (2008). Bayesian statistics are becoming increasingly popular today and a wide range of simple and complex Bayesian methods has been made available to the public (JASP Team, 2018).

However, there are some important facets missing in the Bayesian arsenal. These missing facets include tests to check violations of underlying model assumptions, so called assumption checks. One of these well-known model assumptions that needs to be checked is normality. As mentioned before, the Shapiro-Wilk test is the most used frequentist test, offering a tool to decide between normality and non-normality (Shapiro and Wilk, 1965). But frequentism has more to offer. D'Agostino's K^2 test tries to detect if there is significant skewness in the data (D'agostino et al., 1990), Anscombe-Glynn's test tries to find kurtosis in the data (Anscombe and Glynn, 1983), and a test for uni-modality exists in the form of Hartigan's dip test (Hartigan and Hartigan, 1985). The Bayesian arsenal has not yet arrived at this point, lacking these kinds of tests for underlying model assumptions for now. However, if Bayesian statistics ever aims to be the standard in the field, its arsenal of weapons needs to be completed, including assumption checks. This thesis aims to kick start that and outlines a method for developing a Bayesian test to decide between four hypotheses, H_0: the distribution is normal, H_1: the distribution is non-normal, H_2: the distribution is bi-modal and H_3: the distribution is skewed. The purpose of the test is to yield Bayes factors for normality versus each scenario of non-normality. To achieve this in the Bayesian framework, it is important to translate these scenarios of non-normality into formal models. From these models the marginal likelihoods can be computed which, in turn, can be used to derive Bayes factors for each comparison between scenarios (Dellaportas et al., 2002).

In order to translate scenarios into formal models, a family of distributions needs to be composed. The next section will discuss the rationale used to test for normality in the Bayesian framework and it will explore three families of distributions that have useful properties. Each of the distributions will be coupled to a hypothesis defining a scenario of (non-)normality, thereby creating three comparisons that yield Bayes factors. In the results section there will be a closer look at the behavior of these Bayes factors. An investigation into the influence of the prior will be discussed and recommendations for the use of certain priors are provided. Using these priors, an investigation into Bayes factor behavior as a function of the sample size and the degree of non-normality will be discussed. In the end of the results section, the method is implemented in a couple of real-life scenarios to get an indication of the performance of the test. In the final section, the rationale is implemented in a Bayesian independent samples t-test, showing the versatility of the proposed method.

2 | METHOD

The Bayes factor indicates how much more likely the data are under the null hypothesis H_0 than under alternative hypothesis H_L and is given by:

$$\mathrm{BF}_{0L} = \frac{P(\mathrm{Data} \mid \mathrm{H}_0)}{P(\mathrm{Data} \mid \mathrm{H}_L)} \tag{1}$$

Calculating Bayes factors requires a model for the different types of (non-)normality. As in the frequentist framework, the models under consideration should all sketch a scenario of (non-)normality. Possible scenarios include the general scenario of non-normality, but also bi-modality and skewness. These will be translated into the following numbered hypotheses:

$$\mathrm{H}_0\text{: Normality} \qquad \begin{array}{l} \mathrm{H}_1\text{: Non-normality} \\ \mathrm{H}_2\text{: Bi-modality} \\ \mathrm{H}_3\text{: Skewness} \end{array}$$

By substituting H_L in Equation 1 with one of these numbered hypotheses, a Bayes factor for this comparison can be formulated. For example, the Bayes factor in favor of H_0: normality versus H_1: non-normality can be formulated as BF_{01}:

$$\mathrm{BF}_{\text{Normality vs. Non-normality}} = \mathrm{BF}_{01} = \frac{P(\mathrm{Data} \mid \mathrm{H}_0)}{P(\mathrm{Data} \mid \mathrm{H}_1)} = \frac{P(\mathrm{Data} \mid \text{Normality})}{P(\mathrm{Data} \mid \text{Non-normality})} \tag{2}$$

By varying H_L, three Bayes factors will be computed, quantifying normality or a specific deviation from it. These Bayes factors will be derived from the following comparisons:

$$\begin{aligned} \mathrm{H}_0\text{: Normality vs. }\mathrm{H}_1\text{: Non-normality} &= \frac{P(\mathrm{Data} \mid \text{Normality})}{P(\mathrm{Data} \mid \text{Non-normality})} = \mathrm{BF}_{01} \\ \mathrm{H}_0\text{: Normality vs. }\mathrm{H}_2\text{: Bi-modality} &= \frac{P(\mathrm{Data} \mid \text{Normality})}{P(\mathrm{Data} \mid \text{Bi-modality})} = \mathrm{BF}_{02} \\ \mathrm{H}_0\text{: Normality vs. }\mathrm{H}_3\text{: Skewness} &= \frac{P(\mathrm{Data} \mid \text{Normality})}{P(\mathrm{Data} \mid \text{Skewness})} = \mathrm{BF}_{03} \end{aligned} \tag{3}$$

Until here, the hypotheses have been only a vague description of the scenario they describe. Equation 4 displays how the probability of the data given a hypothesis H_L can be calculated and shows that a model with accompanying priors is required for this calculation. In this equation, *f* is a distribution function with parameters θ_L, π_L represents the prior over the parameters associated with the model for H_L.

$$P\left(\text{Data} \mid H_L\right) = \int \underbrace{f\left(\text{Data} \mid \theta_L\right)}_{\text{Model}} \cdot \underbrace{\pi_L\left(\theta_L\right)}_{\text{Prior}} d\theta_L \tag{4}$$

This means that, in order to calculate Bayes factors, models and priors for these models need to be defined. First there will be a focus on how these models are selected. After that it will be explained how the priors on the parameters are chosen.

The models come from three candidate distributions picked for their close association with normality: the normal distribution, the mixture-normal distribution and the skew-normal distribution. In short, the mixture-normal distribution can be seen as the merging of two normal distributions, thereby creating a different distribution. The skew-normal distribution is an extension of the normal distribution, with the added property that it can easily become oblique. Choosing distributions that share properties with the normal distribution makes the comparison between models fairer, as alternative models only have a few extra parameters, which causes them not to get penalized too much. This means that the candidate distributions are all family of the normal distribution and that deviations from normality should be quantified by additional parameters. The goal of the next part is to determine which model is suited best for assessing specific deviations from normality. For that, it is useful to first study the three candidate distributions that the models come from.

A Model for (Non-)Normality

The Bayes factor in favor of H_0: Normality and H_1: Non-normality (BF_{01}) compares the null hypothesis of normality against the alternative hypothesis of non-normality. Since normality is the simplest scenario that needs to be modeled, the distribution for this scenario must be the smallest in size (e.g., fewest parameters), so that any deviation from normality adds a parameter to the model. The normal distribution will be used to model the base case of normality. The mixture-normal distribution will be used to model non-normality.

The Normal Distribution

The normal distribution (Patel and Read, 1996; Altman and Bland, 1995) is a popular distribution with two parameters; μ represents the mean of the distribution and σ represents the standard deviation. The mean is the location of the distribution, while the standard deviation defines the width of the distribution. The effect of μ and σ on the shape of the distribution is visualized in Figure 1. The normal density is defined as:

$$f\left(x \mid \mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \text{ where } \mu \in \mathbb{R}, \sigma > 0 \tag{5}$$

Figure 1 displays the effect of the mean parameter μ, showing that the distribution shifts along the x-axis when μ is increased (blue distribution) or decreased (red distribution). The green distribution is narrower and taller than the red and blue distributions, due to a smaller standard deviation σ.

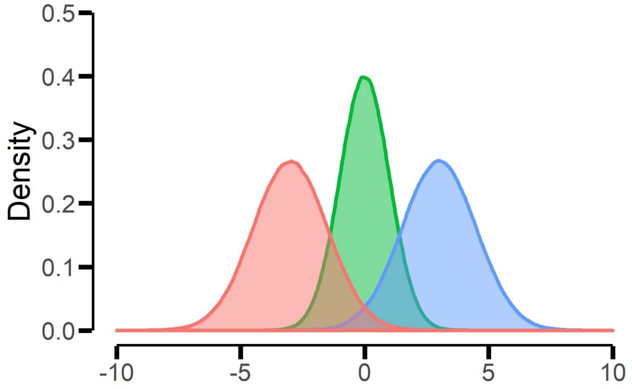


FIGURE 1 Three examples of the normal distribution. The **green** distribution has parameters $\mu = 0,\ \sigma = 1$ and is referred to as the standard normal distribution. The **blue** distribution has parameters $\mu = 3,\ \sigma = 1.5$. The **red** distribution has parameters $\mu = -3,\ \sigma = 1.5$

The Mixture-Normal Distribution

The simplest mixture-normal distribution extends the normal distribution with a mixture parameter $\theta \in$ *(0,1)* (McLachlan and Peel, 2004). It has three parameters; θ, λ_1 and σ. The parameter θ is a mixture weight and controls to what extend two normals are mixed. The location parameter λ_1 is the mean of the distributions, where $-\lambda_1$ is the mean of the first distribution and $+\lambda_1$ is the mean of the second distribution, giving the distribution a symmetric property. Formally, the mixture-normal density is defined as:

$$f(x \mid \theta, \lambda_1, \sigma) = \theta \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\lambda_1)^2}{2\sigma^2}}}_{\text{Distribution 1}} + (1-\theta) \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x+\lambda_1)^2}{2\sigma^2}}}_{\text{Distribution 2}},\ \text{where } \theta \in (0,1), \lambda_1 \in \mathbb{R}, \sigma > 0 \tag{6}$$

In this equation, θ is a mixture weight for dividing the mass over the two normal distributions. Since probabilities must be summable to one, it follows that $1-\theta$ is the mixture weight for the other distribution. Therefore, θ is partially responsible for defining the shape of the distribution. The standard deviation σ again defines the width of both distributions. The mixture-normal distribution can be used to model bi-modality and skewness. By setting $\theta = \frac{e^{\lambda_0}}{e_0^{\lambda}+e^{-\lambda_0}}$ and σ to 1, Equation 6 can be simplified to:

$$f(x \mid \lambda_0, \lambda_1) = \frac{\cosh(\lambda_0 + \lambda_1 x) \cdot \phi(x)}{\cosh(\lambda_0) \cdot e^{\frac{1}{2}\lambda_1^2}},\ \text{where } \lambda_0 \in \mathbb{R}, \lambda_1 \in \mathbb{R} \tag{7}$$

From Equation 7 it follows that, when $0 < \theta < 1$ and $\lambda_1 < 1$, the symmetric property of λ_1 ensures that the distribution is uni-modal. In contrast, when λ_1 is larger than one, the distribution becomes bi-modal. The effect of θ and λ_1 on the shape of the distribution is visualized in Figure 2. The green distribution closely resembles a normal distribution seen in Figure 1. This is caused by the two normal distributions in the mixture-normal overlapping and creating an uni-modal distribution. The blue distribution is also uni-modal, but skewed to the left. The mixture weight θ causes this skewness, as $\frac{3}{4}^{\text{th}}$ of the data ($\theta = .75$) are in the first distribution and $\frac{1}{4}^{\text{th}}$ of the observations are in the second distribution. In the red distribution half of all observations come from each distribution, causing it to become bi-modal.

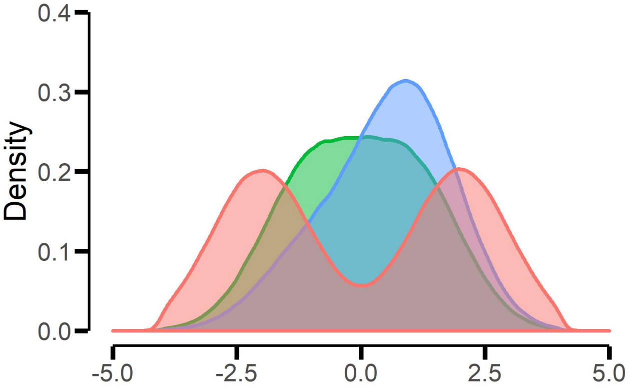


FIGURE 2 Three examples of the mixture-normal distribution. The **green** distribution has parameters $\lambda_1 = 1,\ \theta = .5$ and acts like a normal distribution. The **blue** distribution has parameters $\lambda_1 = 1,\ \theta = .75$ and is a skewed normal distribution. The **red** distribution has parameters $\lambda_1 = 2,\ \theta = .5$ and is a bi-modal distribution.

By adding an extra location parameter to the formula, a more flexible case of the mixture-normal distribution is revealed. With this change, λ_1 in the second part of the formula is substituted by the parameter λ_2. The new formula for the mixture-normal density then becomes:

$$f(x \mid \theta, \lambda_1, \lambda_2, \sigma) = \theta \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\lambda_1)^2}{2\sigma^2}} + (1-\theta) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x+\lambda_2)^2}{2\sigma^2}}, \text{ where } \theta \in (0,1), \lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}, \sigma > 0 \quad (8)$$

If the distribution utilizes two location parameters, $+\lambda_1$ is the mean of the first distribution and $-\lambda_2$ is the mean of the second distribution. This enables the distribution to model more scenarios than the mixture-normal distribution with one mean parameter. To ensure that the model is not penalized too much, we keep one free standard deviation σ in the model instead of two.

Label Switching Problem

When sampling from the posterior distribution of a mixture-normal model using MCMC sampling, a problematic phenomenon associated with mixture-models occurs. At a specific point of the sampling procedure, the values for the parameters switch to a different component of the distribution. This is called the *label switching problem* (Jasra et al., 2005; Stephens, 2000) and occurs when there is no prior information that distinguishes between components of a mixture distribution (e.g. the prior distribution is the same for all components of the mixture distribution). This could be problematic because it results in the posterior distribution being similarly symmetric when samples are taken from it. The symmetry can cause issues when trying to estimate quantities which relate to individual components of the mixture. However, for the current case this problem is not an issue because when there is label switching of the mean parameter λ_1, the linked shape parameter θ switches along with it. Therefore the label switching problem has no repercussions for the model fit of the mixture-normal model, meaning that it can be interpreted without caution.

Because of the distributions flexibility, it is used to model non-normality.

BF_{01} — H_0: Normality vs. H_1: Non-normality

Plugging the models for normality and non-normality into Equation 1 to substitute the hypotheses H_0 and H_L yields a Bayes factor in favor of normality against non-normality of the following form:

$$\mathrm{BF}_{\text{Normality vs. Non-normality}} = BF_{01} = \frac{\iint \prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma\right) \pi\left(\mu, \sigma\right) d\mu\, d\sigma}{\iiiint \prod_{i=1}^{n} f\left(x_i \mid \theta, \lambda_1, \lambda_2, \sigma\right) \pi\left(\theta, \lambda_1, \lambda_2, \sigma\right) d\theta\, d\lambda_1\, d\lambda_2\, d\sigma} \tag{9}$$

Since the mixture-normal distribution is a flexible distribution, it is likely to perform well in modeling non-normality. However, since it is still a parameterized model it will not be able to capture all deviations from non-normality, as for example the Kolmogorov-Smirnov test does. This means that there is a loss of general information that is tested for (e.g., not all deviations can be found), but also that there is a gain in specific information (e.g., deviations that look like mixture-normal distributions are identified better).

A Model for Bi-Modality

The Bayes factor in favor of H_0: Normality against H_2: Bi-modality (BF_{02}) compares the null hypothesis of normality against the alternative hypothesis of bi-modality. For the hypothesis of normality vs. bi-modality, a normal model was tested against a mixture-normal model with restricted parameters. The restriction was implied for the parameters λ_1 and λ_2, controlling the means of the two normal distributions in the mixture. Restricting λ_1 and λ_2 to be higher than 1 would formally mean that the distribution can only model bi-modality, a useful property when testing explicitly for that deviation from normality. The results section features a segment on why the mixture-model was restricted in order to test for bi-modality (page 19).

BF_{02} - H_0: Normality vs. H_2: Bi-modality

Plugging the models for normality and bi-modality into Equation 1 to substitute the hypotheses H_0 and H_L yields a Bayes factor in favor of normality against bi-modality of the following form:

$$\mathrm{BF}_{\text{Normality vs. Bi-modality}} = BF_{02} = \frac{\iint \prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma\right) \pi\left(\mu, \sigma\right) d\mu\, d\sigma}{\iiiint \prod_{i=1}^{n} f\left(x_i \mid \theta, \lambda_1 > 1, \lambda_2 > 1, \sigma\right) \pi\left(\theta, \lambda_1 >!, \lambda_2 > 1, \sigma\right) d\theta\, d\lambda_1\, 1, d\lambda_2 > 1\, d\sigma} \tag{10}$$

Whatever comparison is under investigation one needs to standardize the data ($x = \frac{x-\bar{x}}{\sigma_x}$) when doing the assumption check, something that affects normality in the distribution. However, it does not seem to affect the results in the study. The reason for standardizing is that the mixture-normal model is a symmetrical model around zero, which means that it expects its modes to be around zero. Also, since the restriction of bi-modality formally only applies when $\lambda_1 > 1$, it makes sense to standardize the data so that there is always data present that falls outside of this range. When non-centered data are used there exists a possibility that the data only falls within the range of -1 to 1. For the mixture-normal distribution to detect bi-modality, no mean is allowed that is between -1 and 1. If data have such means the mixture-normal model would fit horribly or even worse, not detect bi-modality at all, as it is restricted to have means that fall outside of this range. Centering the data makes sure that the whole data are in the range of -3 to 3 and the restricted mixture-normal model can be adequately fit (Heiman, 2001).

A Model for Skewness

The Bayes factor in favor of H_0: Normality against H_3: Skewness (BF_{03}) compares the null hypothesis of normality against the alternative hypothesis of skewness. Again, the normal distribution will be used to model the base case of normality. However, this time skew-normal distributions will be used to model skewness.

The Skew-Normal Distribution

The skew-normal distribution (Azzalini, 2005; Liseo and Loperfido, 2004) is an extension of the normal distribution and has three parameters: μ, σ and α. When $\alpha = 0$, then μ is interpreted as the mean of the distribution and σ is interpreted as the standard deviation of the distribution. The additional parameter α is the skewness of the distribution. The effect of α on the shape of the distribution is visualized in Figure 3. The skew-normal density is defined in Equation 11, where Φ indicates the cumulative standard normal distribution function and ϕ indicates the standard normal probability density function.

$$f(x \mid \mu, \sigma, \alpha) = \frac{2}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right) \cdot \Phi\left(\alpha\left(\frac{x-\mu}{\sigma}\right)\right), \text{ where } \mu \in \mathbb{R}, \sigma > 1, \alpha \in \mathbb{R} \tag{11}$$

Figure 3 shows three different scenarios of the skew-normal distribution. The green distribution again resembles a normal distribution. It is essentially the same, since α in this case is zero. However, the red and blue distributions are skewed respectively to the left and to the right, showing the effect of a negative (blue distribution) and positive (red distribution) α. For every distribution, the location parameter μ is set to zero. It is worth to note that the skew-normal distribution models skewness heavily, as both the red and the blue distribution have almost no mass to the other side of zero.

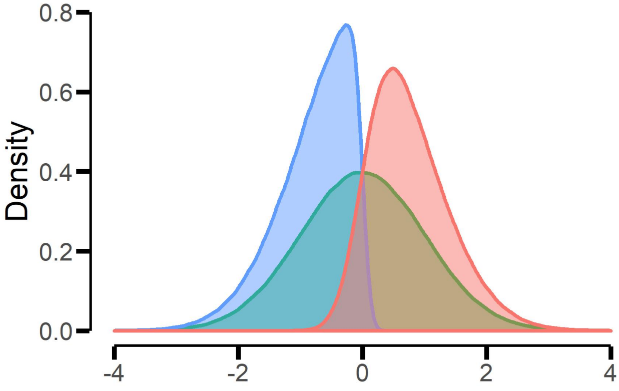


FIGURE 3 Three examples of the skew-normal distribution. The **green** distribution has parameters $\mu = 0,\ \sigma = 1,\ \alpha = 0$. The **blue** distribution has parameters $\mu = 0,\ \sigma = 1,\ \alpha = -10$. The **red** distribution has parameters $\mu = 0,\ \sigma = 1,\ \alpha = 3$.

The option to use the mixture-normal model to test skewness was considered, but the conclusion that this model does not capture skewness as well as the skew-normal model was drawn. This is explained in Figure 4, where the fit of the mixture-normal model and the fit of the skew-normal model is presented on typical skewed data. Visual inspection of the curves indicates that the skew-normal model (green curve) covers more of the skewed distribution than the mixture-normal model (red curve). This observation is the basis of the choice for the skew-normal distribution as the alternative hypothesis in the test for skewness.

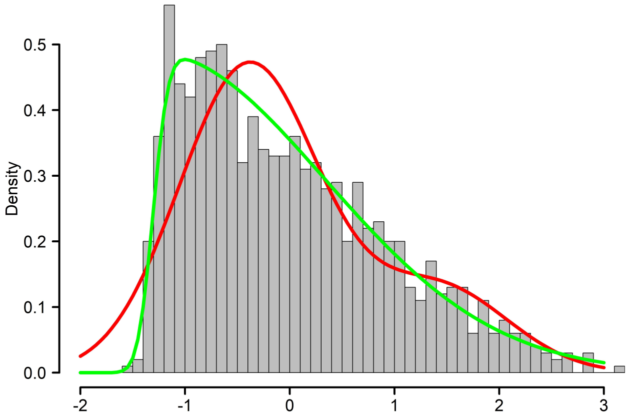


FIGURE 4 Histogram of typical skewed data with the fit of the mixture-normal model (**red**) and the skew-normal model (**green**). Visually, the skew-normal model covers more of the distribution than the mixture-normal model and is therefore the preferred choice to model skewness.

BF_{03} — H_0: Normality vs. H_3: Skewness

Plugging the models of normality and skewness into Equation 1 to substitute the hypotheses H_0 and H_L yields a Bayes factor in favor of normality against skewness of the following form:

$$\mathrm{BF}_{\text{Normality vs. Skewness}} = BF_{02} = \frac{\iint \prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma\right) \pi\left(\mu, \sigma\right) d\mu \, d\sigma}{\iiint \prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma, \alpha\right) \pi\left(\mu, \sigma, \alpha\right) d\mu \, d\sigma \, d\alpha} \tag{12}$$

Prior Selection

In order to calculate Bayes factors, a prior needs to be set on the parameters in each model. The prior that is being put on parameters —be it a model parameter or an effect size— has influence on the Bayes factor, as the latter is based on the former (Gelman et al., 2013). To illustrate this fact, consider Figure 5. Here, a Bayesian independent t-test is done on the mean NEO (Costa and McCrae, 1985) scores between males and females (part of JASP's Kitchen Rolls data set). A Cauchy prior is placed on the effect size δ. The left plot shows the prior (dotted line) and posterior (solid line) under a Cauchy prior with a scale parameter γ of .707. The resulting Bayes factor in favor of a difference between the groups equals 26.261, meaning that it can be fairly confidently said that there is a difference between the mean NEO scores of males and females. However, when a normal prior with a standard deviation of .1 is used instead, the Bayes factor lies around 2. In this scenario, the conclusion would be that there is no convincing evidence for a difference between males and females scores.

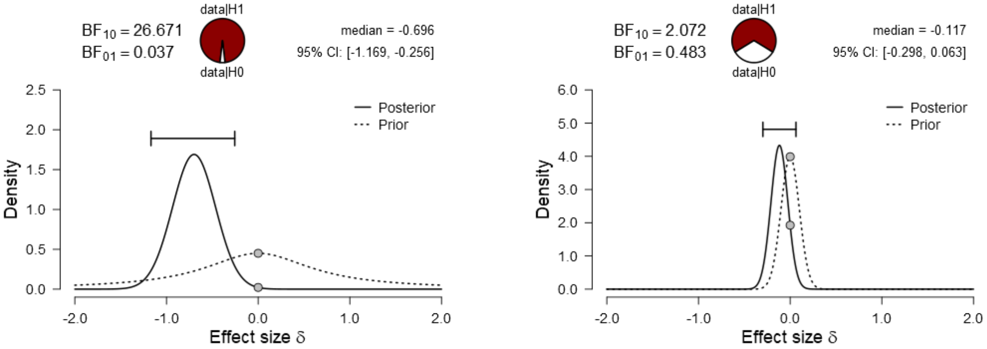


FIGURE 5 Results from a Bayesian independent t-test analysis illustrating the effects of the prior that is used. Both analyses used the same data, however the Bayes factors in favor of a difference between males and females scores differ substantially.

This example illustrates that the prior that is chosen matters for the outcome of the analysis. A good prior would indicate strongly that there is a difference between groups, and indicate no difference when there is not. The candidate models often make use of parameters like the location and the scale of the distribution, which can respectively be seen as the mean μ and the standard deviation σ. The mean of the distributions is continuous and can be negative, so the prior for μ ideally spans the range of positive and negative values. The standard deviation of these distributions is always a positive number, which means that the prior on σ should consider this by only assigning mass to the positive range of numbers. A good suggestion for these priors is shown below (Gelman et al., 2006) and these suggestions were used as priors for μ and σ throughout the study.

$$\mu \sim \text{Normal}\,(0,\ 1) \qquad\qquad \sigma \sim \text{Gamma}\,(2,\ 1) \tag{13}$$

Since different families of priors can result in substantial differences in test results, it is important to research their impact thoroughly. This procedure can be seen as a robustness check, similar to the one incorporated in JASP (Figure 6) where the sensitivity of the analysis to the prior and its parameters is investigated (Berger et al., 1994). The goal of this part of the research is to a) find a suitable family of priors for the parameters and b) choose appropriate hyper parameters for this family. The chosen prior should be one that is robust and differentiates between the models of interest the best. To achieve this, data is simulated under normally distributed data and non-normally distributed data. By investigating the resulting Bayes factors, a prior family and hyper parameter are chosen. The most optimal priors for the parameters θ, λ_1 and α that resulted from this study are shown below and it is indicated to which distribution they belong.

$$\underbrace{\theta \sim \text{Normal}\,(.5,\ .1)}_{\text{Mixture-normal}} \qquad \underbrace{\lambda_1 \sim \text{Normal}\,(0,\ 1)}_{\text{Mixture-normal}} \qquad \underbrace{\alpha \sim \text{Cauchy}\,(0,\ .2)}_{\text{Skew-normal}} \tag{14}$$

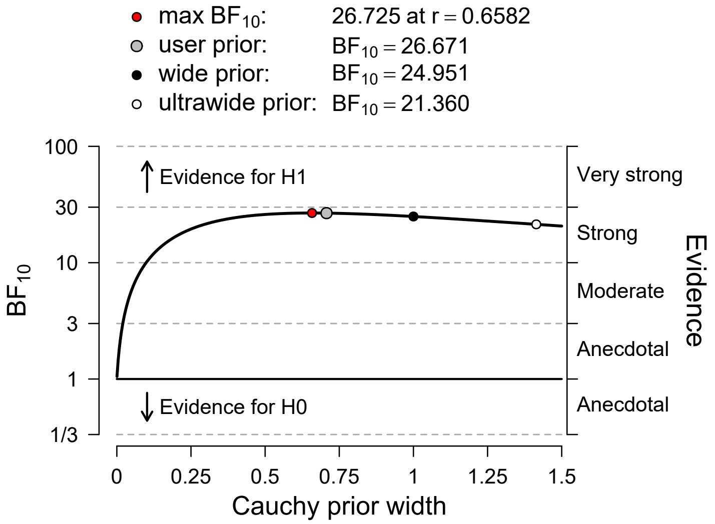


FIGURE 6 JASP's Bayes factor robustness check for an independent samples t-test. The line indicates the Bayes factors under different prior widths for the Cauchy prior on the effect size.

Summary

In this section, hypotheses about scenarios of normality were modeled with three different distributions: the normal distribution, the mixture-normal distribution and the skew-normal distribution. Using these models, the Bayes factors in Equation 3 (BF_{01}, BF_{02} and BF_{03}) can be calculated, quantifying evidence in favor of normality against non-normality, bi-modality and skewness. Table 1 summarizes the previous section by listing the scenarios of normality, the distributions used to model them, their (free and restricted) parameters and the corresponding priors.

TABLE 1 The models that are used to describe scenarios of (non-)normality and their free and restricted parameters and priors. The model color serves the goal of identifying the models in further illustrations, making them easier to interpret.

Scenario	Model	Parameter	Prior	Model color
Normality	Normal	μ	$\mu \sim$ Normal(0, 1)	🟦
		σ	$\sigma \sim$ Gamma(2,1)	
Non-normality	Mixture-normal	θ	$\theta \sim$ Normal(.5,.1)	🟥
		λ_1	$\lambda_1 \sim$ Normal(0,1)	
		λ_2	$\lambda_2 \sim$ Normal(0,1)	
		σ	$\sigma \sim$ Gamma(2,1)	
Bi-modality	Restricted mixture-normal	θ	$\theta \sim$ Normal(.5,.1)	🟧
		$\lambda_1 > 1$	$\lambda_1 \sim$ Normal(0,1)	
		$\lambda_2 > 1$	$\lambda_2 \sim$ Normal(0,1)	
		σ	$\sigma \sim$ Gamma(2,1)	
Skewness	Skew-normal	μ	$\mu \sim$ Normal(0,1)	🟩
		σ	$\sigma \sim$ Gamma(2,1)	
		α	$\alpha \sim$ Cauchy(0,.2)	

Candidate models that capture deviations from normality are extensions of the normal distribution, which means that their parameters quantify these deviations. Since the normal distribution has fewer parameters than the other distributions, it should be preferred when data are truly normally distributed. Testing for normality with the mixture-normal model should equal testing if $\theta = 0$ or $\theta = 1$, since that means that the mixture-normal distribution is uni-modal. The same applies to $\lambda_1 < 1$. Both values of these parameters would mean that the distribution would be equivalent to a normal distribution, since it is uni-modal and not skewed (Figure 2, green distribution). Consequently, testing for deviations from normality would mean testing if $0 < \theta < 1$ or if $\lambda_1 > 1$. Testing for normality with the skew-normal model implies testing if $\alpha = 0$. Fixing $\alpha = 0$ would mean that the skew-normal distribution behaves like the normal distribution, since skewness is absent (Figure 3, green distribution).

3 | RESULTS

Measuring performance of the test is no easy task, since there is no standard for Bayesian assumption checks. Therefore it was decided that desired behavior of the test should include three components; 1) Bayes factors should favor their data generating model, 2) Priors should result in optimal Bayes factors and 3) Bayes factors should —with a sufficient sample size— yield decisive conclusions for any of the competing hypotheses. These criteria are necessary for a useful assumption check and will therefore be the behaviors that will be investigated. The first section will shed light on the selection of appropriate priors to elicit desired Bayes factors. Using these priors, the behavior of the Bayes factors under the influence of a variable sample size will be investigated. The third section will investigate the influence of different data scenarios on the Bayes factors. At the end of this section follows an argumentation on why the restricted mixture-model is selected to model bi-modality.

Prior Selection

In the previous section, it is stated that priors should result in optimal Bayes factors. Here, optimal Bayes factors are defined in terms of discriminative power, that is to indicate strongly that there is a difference between groups, and indicate no difference when there is not. These demands can be translated into two simple heuristics that need to be met. Heuristic one is that the prior should result in maximum Bayes factors in favor of the null hypothesis (BF_{01}) when data are normally distributed. Heuristic two states that the prior distribution should result in minimum Bayes factors in favor of the null hypothesis when data are not normally distributed. To get an indication of the influence of the prior distribution, a simulation study was performed. In the following segments, investigations into various priors will be discussed for the parameters of the mixture-normal model (θ and λ_1) and the skew-normal model (α) in the form of a simulation study. The goal of the simulation study was to select a) an appropriate family of priors and b) an appropriate value of the hyper parameters of the prior. Data (N = 200) were either generated from a normal distribution ($\mu = 0, \sigma = 1$), a mixture-normal distribution ($\theta = .5, \sigma = 1, \lambda_1 = 3$) or a skew-normal distribution ($\mu = 0, \sigma = 1, \alpha = 10$), depending on which prior was being investigated. The prior parameters are varied to investigate their effect on the Bayes factors. When choosing a decent prior, the pattern of Bayes factors is investigated because that provides the most information on how these priors differ from each other. By pinpointing the maximum and minimum Bayes factor under each prior —depending on the data—, the optimal parameter for that family of priors can be identified.

The Mixture Weight θ

The mixture weight θ influences the shape of the mixture-normal distribution, depending on it's value. Since it represents a weight for mixing two normal distributions, θ must lie in the range of 0 to 1. A beta prior would be appropriate to consider for this parameter, since it is mathematically convenient to work with in this range. However, other priors that could be considered are a truncated normal prior and a truncated Cauchy prior. These alternative priors would also span the range of 0 to 1. The goal of selecting a decent prior is to gain the highest Bayes factors when data are normally distributed (Log(BF_{01}), logarithm of the Bayes factors higher than zero are evidence in favor of normality). When data are mixture-normally distributed, one must be looking for the prior that yields the lowest Bayes factors Log(BF_{01}). Figure 7 shows the logarithm of BF_{01} under several parameter values in various priors. For the Cauchy prior, the prior parameter under consideration is the scale parameter γ, the location parameter x_0 is fixed to 0.5. For the beta prior, the prior parameter equals α, the β parameter is set to 1. If β were selected and α were to be fixed to 1, the results are the same. For the normal prior, the prior parameter under consideration is the standard deviation σ. The location parameter μ is fixed to 0.5.

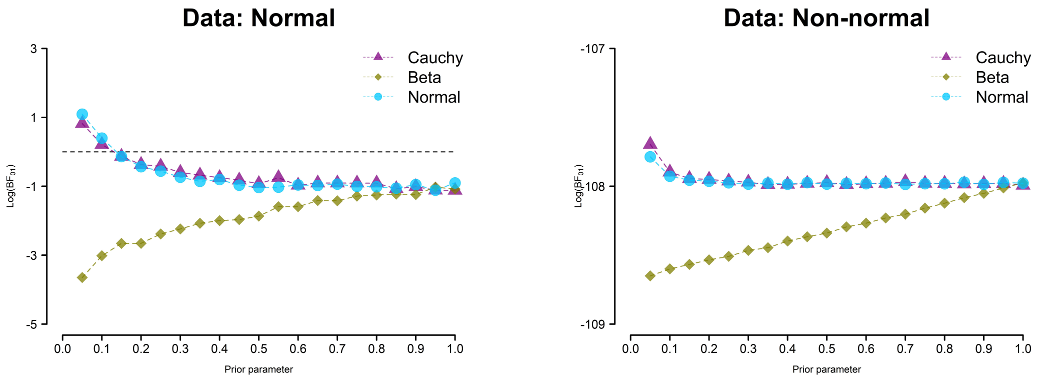


FIGURE 7 Line plot showing Log(BF_{01}) under a Cauchy prior, a beta prior and a truncated normal prior for the parameter θ in the mixture-normal model (N = 200). The **left** plot shows Log(BF_{01}) under normally distributed data. The **right** plot shows Log(BF_{01}) under mixture-normally distributed data.

Figure 7 shows that a truncated normal prior and a Cauchy prior on θ yield the most optimal results, that is, they produce the highest Bayes factors under normally distributed data and low Bayes factors under a non-normally distributed data set. A beta prior produces the lowest Bayes factors under non-normal data, but this difference is minor in relation to the strength of the Bayes factors.

The Location Parameter λ_1

The location parameter λ_1 in the mixture-normal distribution defines the mean of each of the normal distributions in the mixture. Since the data are centered to a range of -3 to 3, and each of the two distribution lies on a different side of zero, it could make sense to put a beta prior on this parameter. Other options could, again, be a truncated normal prior or a Cauchy prior. The location parameters for the Cauchy and normal priors are set to 0. Figure 8 displays the logarithm of BF_{01} under several values of the parameters in the different priors. It shows that a normal prior on λ_1 yields the most optimal result, that is, it produces the highest Bayes factors under normally distributed data and the lowest Bayes factors under non-normally distributed data. Another important property of this normal prior is that it shows the most consistent behavior compared to the Cauchy prior. Inconsistent behavior is an issue when selecting a value for the scale parameter γ in the Cauchy prior. One could easily be in for a surprise, as the pattern is not smooth.

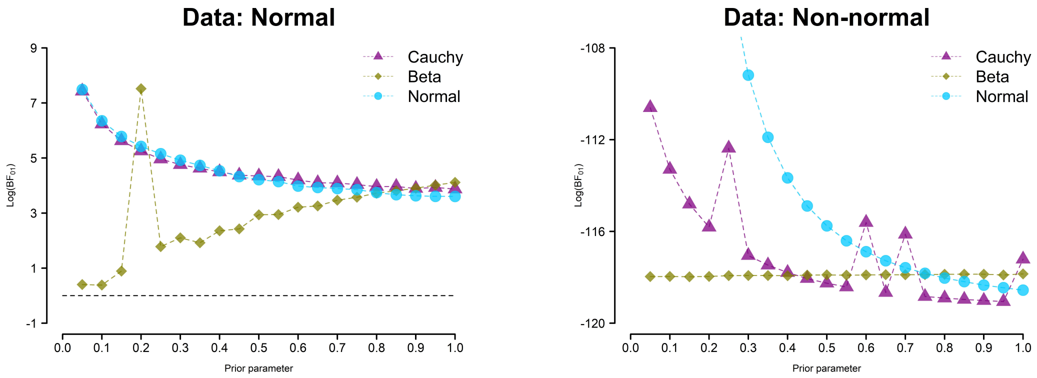


FIGURE 8 Line plot showing Log(BF_{01}) under a Cauchy prior, a beta prior and a truncated normal prior for the parameter λ_1 in the mixture-normal model (N = 200). The **left** plot shows Log(BF_{01}) under normally distributed data. The **right** plot shows Log(BF_{01}) under mixture-normally distributed data.

The fact that the pattern of the Bayes factors under a Cauchy prior is not smooth presents uncertainty for this family of priors. It is then wiser to pick a family with a more consistent pattern, such a the normal one. An important note for the normal prior is that one should not choose a standard deviation that is too low, as that produces a Bayes factor that is favoring the null hypothesis too much.

The Skewness Parameter α

The parameter α in the skew-normal distribution influences how skewed the distribution is. It makes sense to put a prior on this parameter that gives a higher probability to lower values, since data often are not extremely skewed. Therefore, a Cauchy prior would be a sensible option. Also a beta prior is considered alongside a truncated normal prior. The location parameter in the Cauchy and beta priors are fixed to 0. Figure 9 shows the logarithm of BF_{03} under several values of the parameters in the different priors. It shows that a Cauchy prior yields the most optimal result.

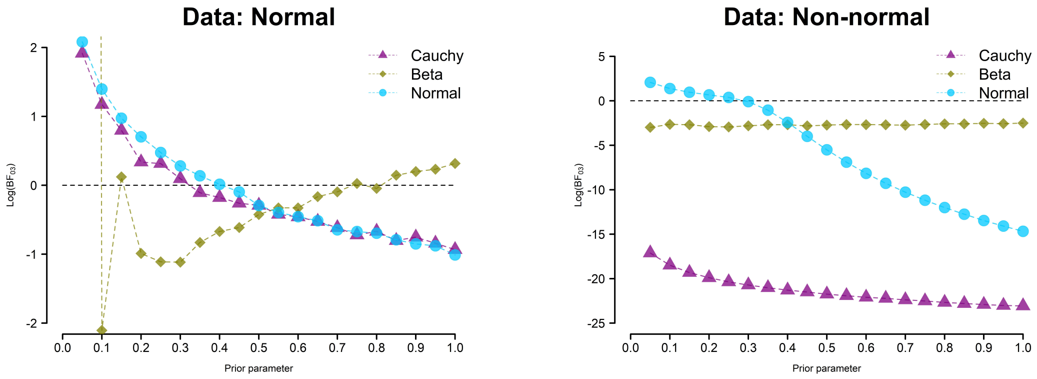


FIGURE 9 Line plot showing Log(BF_{03}) under a Cauchy prior, a beta prior and a truncated normal prior for the parameter α in the skew-normal model (N = 200). The **left** plot shows Log(BF_{03}) under normally distributed data. The **right** plot shows Log(BF_{03}) under skew-normally distributed data.

Recommended Priors

In sum, the priors that arise from this study are recommended because they produce Bayes factors in favor of the null hypothesis of normality that reach a maximum when data are normally distributed, and in turn, reach a minimum when data are non-normally distributed. Selection of specific hyper parameters for these priors is a tricky case, as Figure 7, Figure 8 and Figure 9 are simply based on case studies. However, the influence of the prior family is clearly visible. Equation 15 presents the most effective priors and a recommendation of hyper parameters that are likely to produce optimal Bayes factors. Under these priors, the distribution they are associated with is given. Figure 10 shows how these priors look when plotted individually.

$$\underbrace{\theta \sim \text{Normal}\,(.5,\ .1)}_{\text{Mixture-normal}} \qquad \underbrace{\lambda_1 \sim \text{Normal}\,(0,\ 1)}_{\text{Mixture-normal}} \qquad \underbrace{\alpha \sim \text{Cauchy}\,(0,\ .2)}_{\text{Skew-normal}} \tag{15}$$

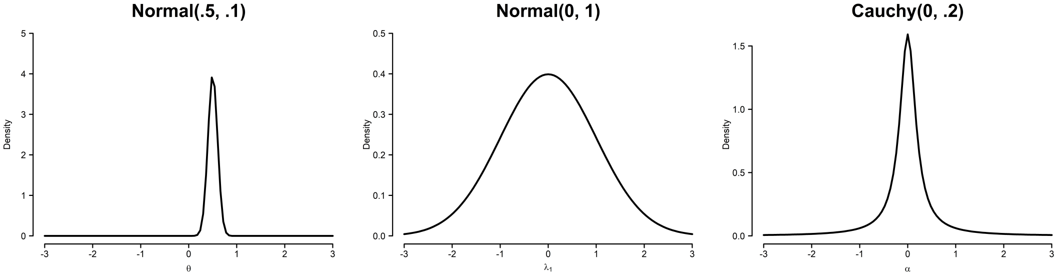


FIGURE 10 Visualization of the recommended priors on the model parameters θ, λ_1 and α. The recommended Normal prior on θ has a mean μ of 0.5 and a standard deviation σ of 0.1. The recommended normal prior on λ_1 has a location parameter μ of 0 and a standard deviation σ of 1. The recommended Cauchy prior on α has a location parameter x_0 of 0 and a scale parameter γ of 0.2.

Bayes Factor Behavior

To learn the behavior of the Bayes factors under several data scenarios, it is useful to map their behavior in multiple dimensions. Regular graphs only do this in one or two dimensions. Heat-maps, however, have the possibility to show a trend given combinations of two variables. The heat-maps in Figure 11 show parameters of the generated data on which the comparison normal vs. mixture-normal (left column) and normal vs. skew-normal (right column) was made. In order to get an indication of the sensitivity of the test, it is useful to look at the Bayes factor behavior by combining the variable *N*, the sample size, and a (non-)normality parameter. The non-normality parameter for mixture-normal data is λ_1, as it controls the location and with that the bi-modality of the distribution. Recall that when $\lambda_1 > 1$, the distribution becomes bi-modal. The non-normality parameter for skew-normal data is α, as it controls the degree of skewness in the data. Normal data can be observed in the Figure 11 by looking at the bottom two rows of each plot. In the left column, the location parameter λ_1 is smaller than 1, indicating normally distributed data. In the right column, the skewness parameter α is small, also indicating somewhat normally distributed data.

By investigating the number of observations (N) in Figure 11 it becomes clear how powerful and sensitive the model is in detecting non-normality. The left column shows the observed behavior of the logarithm of the Bayes factors in favor of the null hypothesis of normality against non-normality Log(BF_{01}). The right column shows the observed behavior of the logarithm of the Bayes factors in favor of the null hypothesis of normality against skewness Log(BF_{03}). The upper row shows data coming from a mixture-normal distribution. The bottom row shows data coming from a skew-normal distribution. This mapping creates four panels, each with its own combination of comparison and data. Various gradations of color indicate the strength of the Bayes factor.

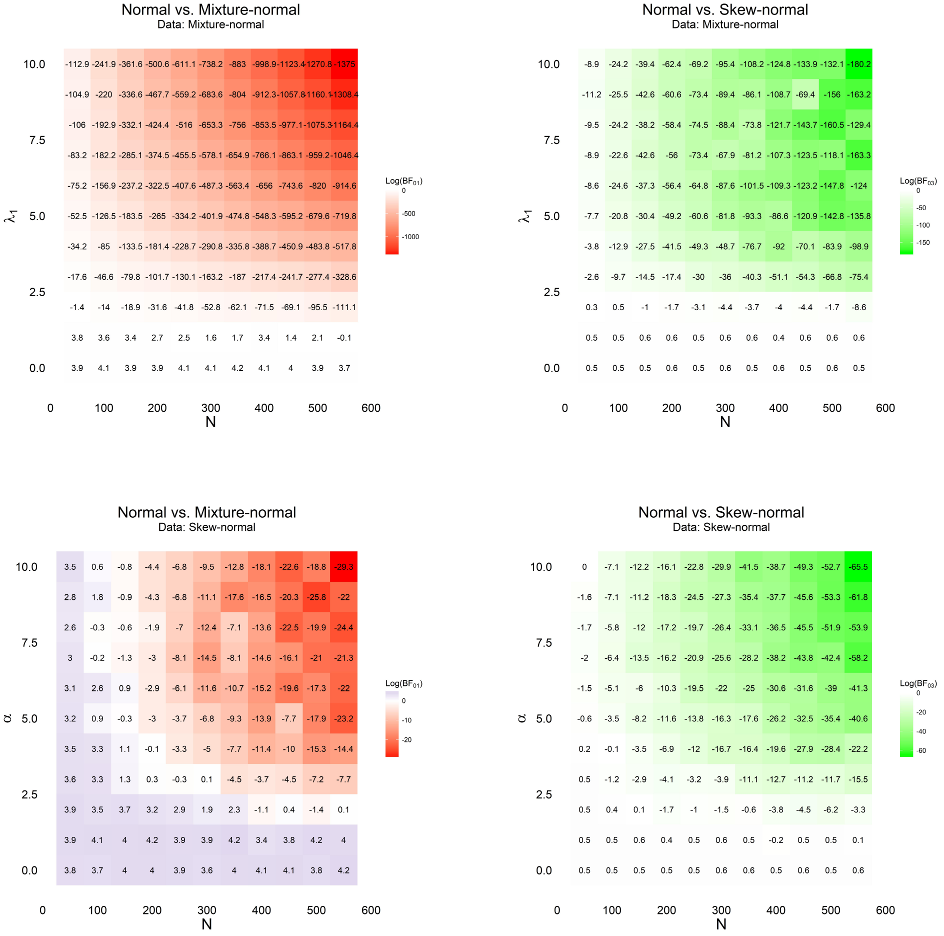


FIGURE 11 Plots showing the behavior of Log(BF_{01}) **(right column)** and Log(BF_{03}) **(left column)** under several data generating parameters. The **first** row shows the data coming from a mixture-normal distribution. The **second** row shows data coming from a skew-normal distribution. The bottom two rows in each panel can be considered to be coming from a normal distribution.

All four panels show expected behavior. When the non-normality parameters (λ_1 and α) increase, the alternative models get more strongly preferred by the Bayes factor. This is desired behavior, as it means that when non-normality, bimodality or skewness increases, the Bayes factor reflects this. The bottom two rows of the panels are more interesting, as the upper left panel shows that under normally distributed data, the normal model is preferred to the mixture-normal model. The upper-right panel shows that under normally distributed data, the normal model is only slightly preferred to the skew-normal model. This indicates that the normal vs. skew-normal comparison is not convincing in detecting normality. However, it is adequate for quantifying the degree of skewness in the data. Furthermore, increasing the sample size under non-normal data means an increase in the Bayes factor for the preferred model, which is desirable.

After investigating what role sample size plays in these comparisons, it is useful to see what the behavior of Bayes factors is under different distributions of data. The following two panel plots illustrate this question. The first plot, Figure 12, shows behavior under data coming from a mixture-normal distribution (N = 100), and varies the parameters θ and λ_1 (σ = 1).

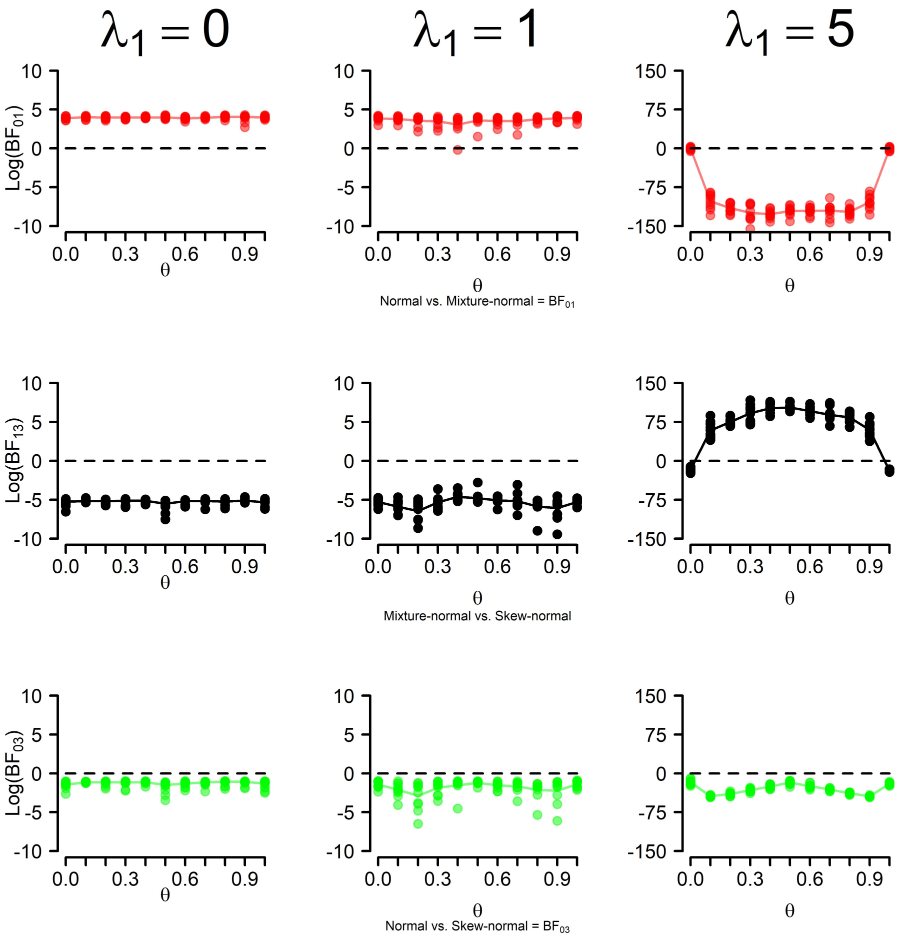


FIGURE 12 Panel plot showing the behavior of Log(BF_{01}), Log(BF_{13}) and Log(BF_{03}) under data generated from a mixture-normal distribution (N = 100) with varying parameters θ and λ_1. The location parameter λ_1 is varied between the columns, while the mixture weight θ is varied on the x-axis. The **top** row shows the comparison normal vs. mixture-normal ($BF_{\text{Normality vs. Non-normality}}$). The **bottom** row shows the comparison normal vs. skew-normal ($BF_{\text{Normality vs. Skewness}}$).

Ideally, Bayes factors show that under the λ_1 parameter values of 0 and 1, the normal model would be preferred over the mixture-normal model, as data generated under these parameters is essentially normally distributed. This is the case in the upper-left and upper-middle panel. Only when λ_1 > 1, the data would be considered non-normal (e.g. bi-modal). The upper-right panel shows that Bayes factors are in favor of the mixture-normal model when this is the case. The second row displays the comparison mixture-normal vs. skew-normal. One would expect the skew-normal to outperform the mixture-normal model when data are essentially normally distributed, as the mixture-normal model has an extra parameter which is penalized. When data becomes bi-modal the mixture-normal model outperforms the skew-normal model, as shown in the middle-right panel. Looking at the normal vs. skew-normal case shows that the skew-normal model is preferred over the normal model in all cases of the data. Next is a scenario where the data are coming from the skew-normal distribution (N = 100), varying the data generating parameters μ and α ($\sigma = 1$), see Figure 13.

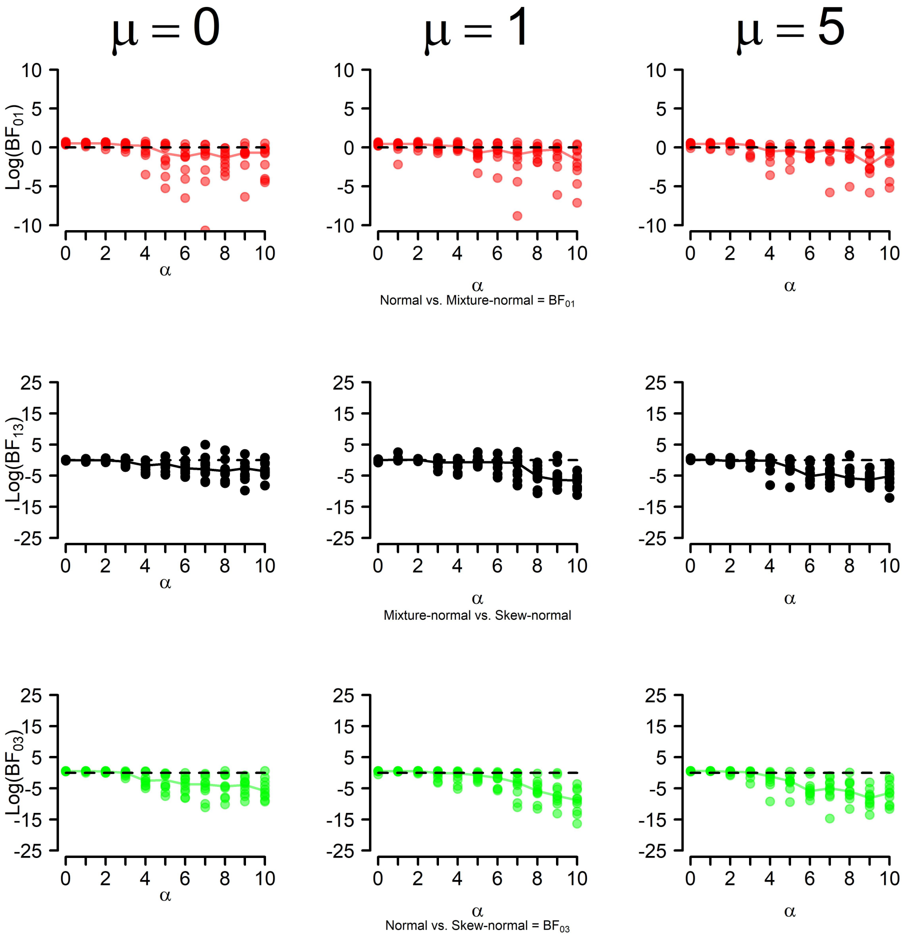


FIGURE 13 Panel plot showing the behavior of Log(BF_{01}), Log(BF_{13}) and Log(BF_{03}) under data generated from a skew-normal distribution (N = 100) with varying parameters μ and α. The mean parameter μ is varied between the columns, while the skewness parameter α is varied on the x-axis. The **top** row shows the comparison normal vs. mixture-normal ($BF_{\text{Normality vs. Non-normality}}$). The **bottom** row shows the comparison normal vs. skew-normal ($BF_{\text{Normality vs. Skewness}}$).

Ideally, Bayes factors show that the skew-normal model would be preferred in cases where simply $\alpha > 0$. This can be observed in all columns of the second and third row of the panel plot. A distinct trend is clearly visible, namely that when the α parameter increases and the data distribution becomes more skewed, the skew-normal model is strongly preferred over both other models.

Restricting the Mixture-Normal Distribution

To argue why bi-modality is modeled with a restricted mixture-normal distribution instead of a regular mixture-normal distribution, a scenario is presented where both models are fit on the same data. The data (N = 1000) are sampled from a mixture-normal distribution ($\theta = .5, \sigma = 1$) where the parameter λ_1 is varied. Bayes factors were calculated in favor of the normal model against the mixture-normal model or a restricted mixture-normal model as alternative hypotheses. Figure 14 shows that the Bayes factors under a restricted mixture-normal model are equally strong as Bayes factors under a non-restricted mixture-normal model when the data are bi-modal (λ_1 > 1). When the data are normal—formally if $\lambda_1 < 1$ when data are sampled from a mixture-normal distribution— this comparison should return Bayes factors in favor of the normal model. This can be observed in the restricted mixture-normal model. However, it is not the case when the regular mixture-normal model is used, since its Log(BF$_{02}$) hardly exceeds zero. This is enough reason to prefer the restricted mixture-normal as the model to detect bi-modality. However, the regular mixture-normal model is kept to detect any general deviation from normality, since it is impossible to also model traditional skewness for the restricted mixture-normal model. As the restriction is applied to its two modes, which are around zero, they cannot overlap and as such cannot capture skewness.

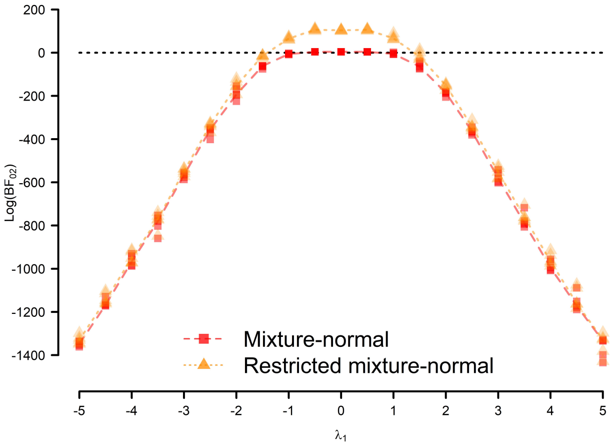


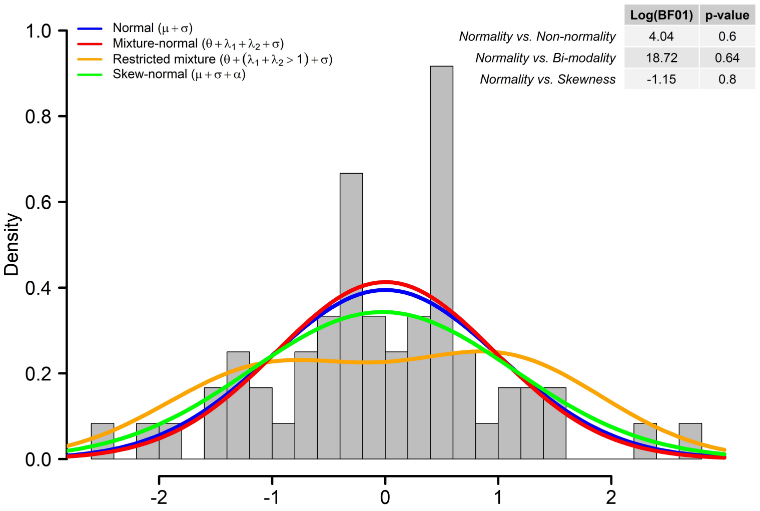
FIGURE 14 Line graph showing Log(BF$_{02}$) for the comparison normal vs. mixture as a function of the data generating parameter λ_1. As can be seen, the restricted mixture-normal model yields stronger Bayes factors when the data are normally distributed (e.g. $-1 < \lambda_1 < 1$).

Real-life Scenarios

Until now, the focus has been on the performance of the test under controlled circumstances. However, the test needs to be able to produce sensible results when applied in real-life situations. Therefore, this section discusses three real world demonstrations of the developed test, based on publicly available data from the web. The first case will focus on the average temperature in New Haven, Connecticut and asks the question whether this natural phenomenon is normally distributed. The second example will focus on the number of years of experience of NBA basketball players and presents us with an interesting question of whether there exists bi-modality in this distribution. The third and final example will focus on the rating of board games, an illustration of the skewed case. Alongside the Bayesian tests for (non-)normality, frequentist results for the corresponding comparisons will be displayed to show whether there is correspondence between the two methods. For the test of non-normality, the *p* value from the Shapiro-Wilk test is displayed. For the test of bi-modality, the *p* value from Hartigan's dip test is displayed. For the test of skewness, the *p* value from D'Agostino's K^2 test is displayed. For all tests the same rule applies; when the *p* value is below .05, the data are not normally distributed.

Average Temperature in New Haven, Connecticut

New Haven, Connecticut has a typical New York City area climate. It has long, hot summers, and cool to cold winters. From May to late September, the weather is typically hot and humid, with average temperatures exceeding 27 degrees Celsius on 70 days per year. In summer, the Bermuda High creates as southern flow of warm and humid air, with frequent thundershowers. October to early December is normally mild to cool late in the season, while early spring can be cool to warm. Winters are moderately cold with both rain and snow fall. The weather patterns that affect New Haven result from a primarily offshore direction, thus reducing the marine influence of Long Island, although, like other marine areas, differences in temperature between areas right along the coastline and areas a mile or two inland can be large at times. Because of the regular weather pattern, it can be expected that the temperature is normally distributed. This data is obtained from weather stations and shows the average temperature in 60 years (N = 60).



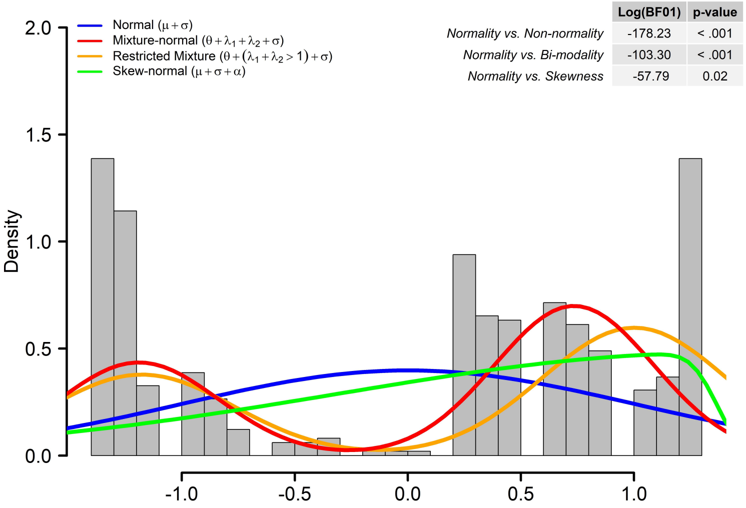
	Log(BF01)	p-value
Normality vs. Non-normality	4.04	0.6
Normality vs. Bi-modality	18.72	0.64
Normality vs. Skewness	-1.15	0.8

FIGURE 15 The centered distribution of the temperature in New Haven, Connecticut over the course of 60 years (N = 60). The Bayes factor in favor of normality equals $e^{4.04} = 56.83$. It is interesting to note that BF$_{\text{Normality vs. Skewness}}$ is in favor of skewness, even though the data are roughly normally distributed.

Figure 15 shows something that resembles a normal distribution, but could also be coming from a bi-modal distribution with long tails. The method tells us that the data are normally distributed. The H_0: Normality vs. H_1: Non-normality comparison shows that the hypothesis of normality is $e^{4.04}$ = 56.83 times more likely than the hypothesis of non-normality. Even with 60 observations, that is pretty high. The H_0: Normality vs. H_2: Bi-modality comparison shows that the hypothesis of normality is $e^{18.72}$ = 134894021 times more likely than the hypothesis of bi-modality. The third comparison shows that the hypothesis of skewness is $\frac{1}{e^{-1.15}}$ = 3.15 more likely than the hypothesis of normality.

NBA Players Experience

The NBA is one of the four major professional sports leagues in the United States and Canada. NBA players are the world's best paid athletes by average annual salary per player. The teams exist of older and younger players together. Hypothetically, the distribution of the number of years of experience could be bi-modal, as some players stay in the game for a long time and others are short-stayers. This data is collected from NBA player records (N = 490).



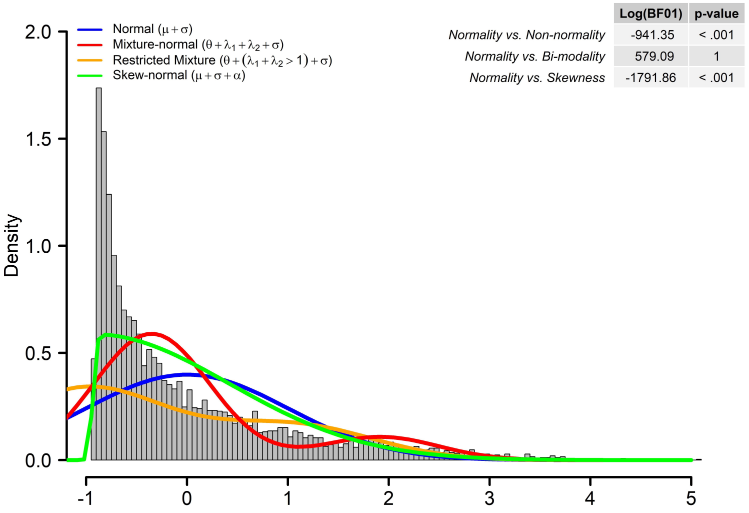
	Log(BF01)	p-value
Normality vs. Non-normality	-178.23	< .001
Normality vs. Bi-modality	-103.30	< .001
Normality vs. Skewness	-57.79	0.02

FIGURE 16 The centered distribution of the number of years of experience in a sample of NBA baseball players (N = 490). The strongest Bayes factor is in favor of bi-modality and equals $\frac{1}{e^{-103}}$ = 5.4e^{44}. It is worth to note that the tests for skewness both reveal that the data is skewed. If the data are investigated, that turns out to be not true.

Figure 16 seems like a bi-modal distribution, as it has much players with a standard deviation of -1 from the mean an a lot of players around a standard deviation of +.75 from the mean. The Bayesian method shows that the data are strongly non-normally distributed, and it is agreed upon by the frequentist Shapiro-Wilk test (p < .001). It is about $2e^{77}$ times more likely that the data are non-normally distributed. The H_0: Normality vs. H_2: Bi-modality comparison shows that the hypothesis of bi-modality is $\frac{1}{e^{-103}}$ = 5.4e^{44} times more likely than the hypothesis of normality. With such strong Bayes factors, it is highly reasonable to assume that the data can be seen as bi-modal.

Geek Rating of Board Games

Sure, the classic board games like Monopoly, Risk, and Battleship are still great fun. But the number of new games has exploded in the last several years as designers think of new adventures, deck-building games, and zombie survival games. But what effect did that have on the quality of the games? This data shows how self-proclaimed geeks rated several board games after they played them (N = 5000). The data looks heavily skewed to the right, as there are many generic board games, while few really fun games are present. The distribution of data in Figure 17 is skewed, so it is logical to expect that the new test should show this as well.



	Log(BF01)	p-value
Normality vs. Non-normality	-941.35	< .001
Normality vs. Bi-modality	579.09	1
Normality vs. Skewness	-1791.86	< .001

FIGURE 17 The centered distribution of the geek rating of board games (N = 5000). The data are heavily skewed to the right, as the strongest Bayes factor tells us. It equals $\frac{1}{e^{-1791}} = \infty$.

The frequentist tests agree to this conclusion, with the Shapiro-Wilk test showing that the data are non-normally distributed (p <.001) and D'agostino's test indicating that the data are skewed (p < .001). The Bayes factor in favor of non-normality equals $\frac{1}{e^{-941}} = \infty$, indicating that the data are strongly non-normal. The Bayes factor in favor of skewness equals $\frac{1}{e^{-1791}} = \infty$, indicating that it is likely that the data are skewed. The Bayes factor for bi-modality tells us that it is more likely that the data are normally distributed than that it is bi-modal. This is a desired result, since there are no two modes that can be identified, thereby validating this outcome.

All data sets and R scripts used to generate the previous figures can be found in the on-line appendix at `https://osf.io/ydfuq/` where, in addition to the examples in this thesis, other examples are available. Furthermore, the code for a user-friendly R-function of the assumption check can be found in Appendix A. It takes three arguments, the first of which is a vector of the data. The second argument is a logical indicating whether to compute Bayes factors in favor of the null hypothesis (BF_{01}) or in favor of the alternative hypothesis (BF_{10}). The third argument is a logical indicating whether to compute the logarithm of the desired Bayes factor.

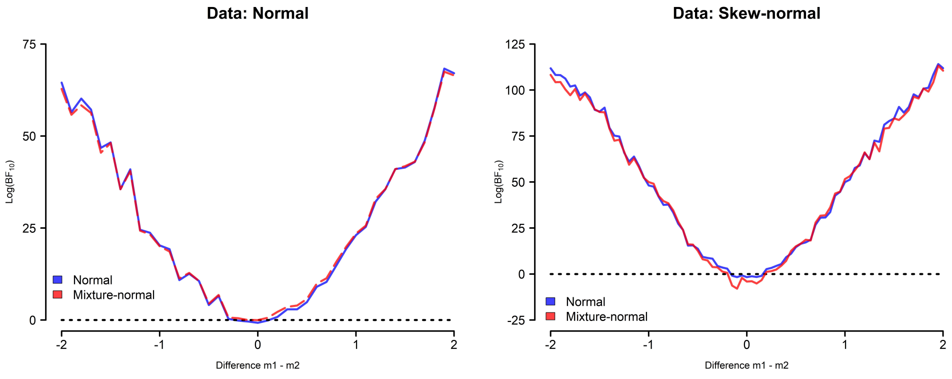
T-test Application

In order to show the potential application of this test, a scenario will be presented where the idea of the method is implemented in a Bayesian independent samples t-test. The traditional idea of this test is to compare a model with two normal distributions with free parameters μ and σ against a model with two normal distributions fixed to the same location ($\mu_{\text{model 1}} = \mu_{\text{model 2}}$). If the model with identical locations is preferred, the conclusion is drawn that it is more likely that the two variables have no difference in the mean (e.g. the parameter μ) than that they have a difference in the mean. As always, by comparing the likelihood of the models a Bayes factor for the comparison can be derived. The new approach substitutes the normal distributions —used to do inference with the t-test— for mixture-normal distributions. The idea here is that using mixture-normal distributions instead of normal distributions better accounts for deviations from normality in the data. Results show that the method shows somewhat correspondence with the Bayesian independent samples t-test in JASP, a well-implemented version of the traditional test. Table 2 shows several applications of the test on JASP's example data sets. All information to reproduce the examples is again provided in the on-line appendix at `https://osf.io/ydfuq/`.

TABLE 2 Results for the re-analysis of several data sets taken from JASP's data library. The Bayes factors generated by the new method correspond with JASP in several cases.

Data Set	N	Dependent	Grouping	$\mu_1 - \mu_2$	BF_{01} (JASP)	BF_{01} (Test)
Kitchen Rolls	102	mean_NEO	Sex	−0.355	26.671	164.92
Directed Reading Activities	44	drp	group	−9.954	2.217	0.0018
Eye Movements	49	CriticalRecall	Condition	4.412	6.760	4.130338
Invisibility	24	Mischief	Cloak	−1.250	1.051	0.35

Table 2 shows that in half of the examples, the test yields the same result as JASP. However, in some cases the test deviates from JASP's result. It seems as if the test does not capture the direction of the data well. To confidently conclude something about the performance of the method, it is necessary to identify a trend in behavior and compare it to the same trend in a Bayesian t-test based on normal distributions. Figure 18 shows exactly this, The blue line quantifies the logarithm of BF_{10} —the Bayes factor in favor of the alternative hypothesis of non-normality— under a test that assumes two normal distributions. The red line quantifies the logarithm of BF_{10} under the assumption of two mixture-normal distributions. In the upper-left plot, the data comes from a normal distribution. In the upper-right plot, the data comes from a skew-normal distribution. Lastly, in the bottom plot the data comes from a mixture-normal distribution.



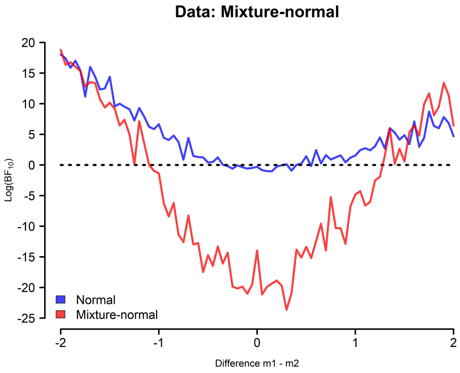


FIGURE 18 Plots showing the logarithm of the Bayes factors in favor of normality against non-normality (BF_{10}) as a function of the difference between the two means in the sample (N = 100). This relation was investigated for three different types of data. The **blue** line quantifies BF_{10} under a test that assumes two normal distributions. The **red** line quantifies BF_{10} under the assumption of two mixture-normal distributions. It can be seen that the test based on the mixture-normal distributions more strongly quantifies evidence in favor of the null hypothesis.

Figure 18 shows that the two methods show very similar behavior, with the difference that the method based on mixture-normal distributions yields higher Bayes factors in favor of the null hypothesis of no difference when the difference in the means is low and data are non-normally distributed. This is especially the case when data comes from a mixture-normal distribution and is thus either bi-modal or non-normal. This observation represents a key improvement because the evidence in favor of the null hypothesis of no difference from the normal-based test in these regions is sparse. The method based on mixture-normal distributions yields Bayes factors in favor of the null hypothesis (e.g. Log(BF) < 0) when the difference is small. The similar behavior of the two methods means that there are no loss of information while data is normal and skew-normal. However, there is a gain in information in favor of the null hypothesis when data is mixture-normal, an advantage of the previous method. The new method quantifies stronger evidence in favor of the null hypothesis when there is no difference and is therefore preferred to the normal method that quantifies weaker evidence. There might be more differences in behavior when simulating data with a smaller sample size (e.g. N = 50). Another possible explanation can be that the fact that the test based on normal distributions is highly sensitive to differences. The user-friendly code for the t-test application is presented in Appendix B. It takes three arguments, the first of which is a vector of the data in group 1. The second argument should be a vector of the data in group 2. The third argument is a logical indicating whether to compute Bayes factors in favor of the null hypothesis (BF_{01}) or in favor of the alternative hypothesis (BF_{10}). The third argument is a logical indicating whether to compute the logarithm of the desired Bayes factor.

4 | DISCUSSION AND CONCLUSION

In this thesis it is shown that deviations from normality in the distribution of data can be assessed and quantified using the Bayesian framework, comparing models that incorporate various distributions —like the mixture-normal and the skew-normal distribution— to ultimately derive Bayes factors in favor of normality against non-normality, bi-modality and skewness. At this point, it is important to remember that Bayes factors quantify relative evidence for both models and cannot be treated as absolute truth. This means that, even if both models fit the data poorly, high Bayes factors can still be the outcome of the analysis. To relate this to the assumption check, an overwhelming Bayes factor in favor of bi-modality does not mean that the data are truly bi-modal. It merely means that data are more likely to be bi-modal than to be normal. In conclusion, the behavior of the Bayes factors has been elaborated on in different data scenarios, showing the influence of the sample size and the non-normality parameters. Recommendations for priors on the various model parameters have been given and the method has been put to work in several real-life data examples. Finally, the method has been implemented in an existing test —a Bayesian independent samples t-test— to show its potential application for other purposes than an assumption check. JASP data have been used so that the reader can reproduce all examples if desired. All data, R code, additional information and files can be found in the on-line appendix at https://osf.io/ydfuq/.

A point of critique on this work is the fact that, when comparing the normal model against the skew-normal model, the skew-normal model is only weakly preferred (Figure 11, right column). This is problematic, since these specific comparisons are not able to produce convincing Bayes factors to support the conclusion that data are normally distributed. A possible explanation for this is that it might be due to the fact that the skew-normal model only has one extra parameter in comparison with the normal model. The mixture-normal with two location parameters has two extra parameters. This can cause the skew-normal to only receive a slight penalty, one that is not enough to prefer the normal model. Another explanation is that, since data are never perfectly normal, this comparison might be too sensitive to abnormalities in the data due to the inflexibility of the normal model. Therefore it is wise to not take the normal vs. skew-normal comparison into account when detecting of (non-)normality, as this is a flawed conclusion.

The comparisons featuring the mixture-normal distribution (e.g. normal vs. non-normal & normal vs. bi-modal) seem to be sensitive to the size of the sample. For example, mixture-normal data with means of -1 and 1 will be identified as normal when the sample size is small (Figure 11, left column). In contrast, the conclusion will lean towards non-normality when the sample size becomes significantly bigger. This could potentially lead to biased conclusions, as it seems that any small deviation from normality could be evidence for the preference of the mixture-normal model. However, this could also be an advantage for the comparison. Because the distribution of data becomes more specific when there are more observations, demands for normality should become higher. That means that these slight deviations from normality are desired to be recognized by the comparison.

As for the restricted mixture-normal model, a drawback in modeling bi-modality seems to be the fact that it has difficulties fitting onto data that is in the range $-1 < x < 1$. Because of its restriction that the mean of each distributions should lie at least at the number (-)1 or further away from zero, it is impossible for the restricted mixture-normal model to model data outside of this range. The inability of this model to fit onto data in the range $-1 < x < 1$ will become apparent when trying to fit the model on small-ranged data. This drawback is remedied by standardizing the data. This way the restriction will not undermine the models ability to fit onto the data. However, standardizing modifies the data in such a way that it is only an approximation of the real data and in such, one is not really drawing conclusions about the actual data.

A final improvement can be made in the application of the method in the Bayesian independent samples t-test. In this scenario, the method has difficulties expressing a Bayes factor in favor of the null hypothesis of no difference when two mixture-normal distributions are used as underlying data distributions, but the actual data are normally distributed. Figure 18 (upper-left plot) shows that in this case, the Bayes factor never indicates evidence in favor of the null hypothesis. Even when the difference between the means is zero, this happens. Not being able to quantify evidence for the null hypothesis is an absolute problem for this test. However, this problem could be due to the sample size (N =

100) used in the study. Despite the fact that a sample size of 100 should technically be able to provide evidence for the null, it is worth to investigate what the behavior of the Bayes factors will be under a smaller sample size. However, it is worth noting that Bayes factors using a normal-based method show almost the same behavior, barely dropping below zero.

In sum, this work provides researchers with a new tool to test for normality in the Bayesian framework. It has been argued that this tool can successfully detect (non-)normality, bi-modality and skewness in a data distribution and quantify Bayes factors in favor of normality against non-normality, bi-modality and skewness. Applications in a t-test scenario are promising, as the current implementation provides stronger evidence in favor of the null hypothesis, an advantage over the normal-based method. The R code in Appendix A and B should make it easy for researchers to perform the assumption check and t-test, as they provide user-friendly R code that can be run instantly.

REFERENCES

Altman, D. G. and Bland, J. M. (1995) Statistics notes: the normal distribution. *Bmj*, **310**, 298.

Anscombe, F. J. and Glynn, W. J. (1983) Distribution of the kurtosis statistic b 2 for normal samples. *Biometrika*, **70**, 227–234.

Azzalini, A. (2005) The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159–188.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B. et al. (1994) An overview of robust bayesian analysis. *Test*, **3**, 5–124.

Costa, P. T. and McCrae, R. R. (1985) The neo personality inventory.

D'agostino, R. B., Belanger, A. and D'Agostino Jr, R. B. (1990) A suggestion for using powerful and informative tests of normality. *The American Statistician*, **44**, 316–321.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002) On bayesian model and variable selection using mcmc. *Statistics and Computing*, **12**, 27–36.

Diaz-Serrano, L. (2005) Labor income uncertainty, skewness and homeownership: A panel data study for germany and spain. *Journal of Urban Economics*, **58**, 156–176.

Field, A. (2009) *Discovering statistics using SPSS*. Sage publications.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis*. CRC press.

Gelman, A. et al. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, **1**, 515–534.

Ghasemi, A. and Zahediasl, S. (2012) Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, **10**, 486.

Hartigan, J. A. and Hartigan, P. M. (1985) The dip test of unimodality. *The Annals of Statistics*, 70–84.

Heiman, G. W. (2001) *Understanding research methods and statistics: An integrated introduction for psychology*. Houghton, Mifflin and Company.

JASP Team (2018) JASP (Version 0.8.5)[Computer software]. URL: `https://jasp-stats.org/`.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 50–67.

Kolmogorov, A. (1933) Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, **4**, 83–91.

Liseo, B. and Loperfido, N. (2004) Default bayesian analysis of the skew-normal distribution. *Journal of Statistical Planning and Inference*, **136**, 373–389.

McLachlan, G. and Peel, D. (2004) *Finite mixture models*. John Wiley & Sons.

Öztuna, D., Elhan, A. H. and Tüccar, E. (2006) Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, **36**, 171–176.

Patel, J. K. and Read, C. B. (1996) *Handbook of the normal distribution*, vol. 150. CRC Press.

Peat, J. and Barton, B. (2008) *Medical statistics: A guide to data analysis and critical appraisal*. John Wiley & Sons.

Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Smirnov, N. (1948) Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, **19**, 279–281.

Steinskog, D. J., Tjøstheim, D. B. and Kvamstø, N. G. (2007) A cautionary note on the use of the kolmogorov–smirnov test for normality. *Monthly Weather Review*, **135**, 1151–1157.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795–809.

Thode, H. (2002) Testing for normality new york.

Ulrich, R. and Miller, J. (1993) Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, **37**, 513–525.

Wagenmakers, E.-J. (2007) A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, **14**, 779–804.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T. and Iverson, G. J. (2008) Bayesian versus frequentist inference. *Bayesian evaluation of informative hypotheses*, 181–207.

APPENDIX A - R FUNCTION FOR THE ASSUMPTION CHECK

```r
bayesian.normality <- function(data, bf10 = TRUE, Log = FALSE){

    # stanModelsAC.RData can be found at https://osf.io/pdb8j/
    load("stanModelsAC.RData")

    library(rstan)
    library(bridgesampling)

    sink(file="undesiredFile.txt") # Prevent printing sampling procedure
    stanfitH0 <- suppressWarnings(sampling(stanmodelH0,
                                data = list(X = data, n = length(data)),
                                iter = 2000, warmup = 500, chains = 1, cores = 4))
    stanfitH1 <- suppressWarnings(sampling(stanmodelH1,
                                data = list(X = data, n = length(data)),
                                iter = 2000, warmup = 500, chains = 1, cores = 4))
    stanfitH2 <- suppressWarnings(sampling(stanmodelH2,
                                data = list(X = data, n = length(data)),
                                iter = 2000, warmup = 500, chains = 1, cores = 4))
    stanfitH3 <- suppressWarnings(sampling(stanmodelH3,
                                data = list(X = data, n = length(data)),
                                iter = 2000, warmup = 500, chains = 1, cores = 4))
    sink()

    # Bridge sampling
    H0.bridge <- suppressWarnings(bridge_sampler(stanfitH0, silent = TRUE))
    H1.bridge <- suppressWarnings(bridge_sampler(stanfitH1, silent = TRUE))
    H2.bridge <- suppressWarnings(bridge_sampler(stanfitH2, silent = TRUE))
    H3.bridge <- suppressWarnings(bridge_sampler(stanfitH3, silent = TRUE))
    # Calculate Bayes factors for each hypothesis
    BF_normality <- bf(H2.bridge, H1.bridge,log = FALSE)$bf
    BF_bimodality <- bf(H2.bridge, H0.bridge,log = FALSE)$bf
    BF_skewness <- bf(H2.bridge, H3.bridge,log = FALSE)$bf

    # Adjust Bayes factors to preferences
    if(bf10){
        BF_normality <- 1/BF_normality
        BF_bimodality <- 1/BF_bimodality
        BF_skewness <- 1/BF_skewness
    }

    if(Log){
        BF_normality <- log(BF_normality)
        BF_bimodality <- log(BF_bimodality)
        BF_skewness <- log(BF_skewness)
    }
    if(BF_normality > 1){
        BF_normality <- round(BF_normality, 3)
    }
    if(BF_bimodality > 1){
        BF_bimodality <- round(BF_bimodality, 3)
    }
    if(BF_skewness > 1){
        BF_skewness <- round(BF_skewness, 3)
    }

    returnObject <- list(stanfitH2, stanfitH1, stanfitH3, stanfitH0,
                         H2.bridge, H1.bridge, H3.bridge, H0.bridge)
    names(returnObject) <- c("H0.fit", "H1.fit", "H2.fit", "H3.fit",
                             "H0.bridge", "H1.bridge", "H2.bridge", "H3.bridge")

    # Print the output
    cat("### H0: The data distribution is normal \n")
    cat("### H1: The data distribution is non-normal \n")
    cat("### H2: The data distribution is bi-modal \n")
    cat("### H3: The data distribution is skewed \n")

    if(!Log){
        if(bf10){
            cat("### BF10 equals", BF_normality, "\n")
            cat("### BF20 equals", BF_bimodality, "\n")
            cat("### BF30 equals", BF_skewness, "\n")
        } else {
            cat("### BF01 equals", BF_normality, "\n")
            cat("### BF02 equals", BF_bimodality, "\n")
            cat("### BF03 equals", BF_skewness, "\n")
        }
    } else {
        if(bf10){
            cat("### Log BF10 equals", BF_normality, "\n")
            cat("### Log BF20 equals", BF_bimodality, "\n")
            cat("### Log BF30 equals", BF_skewness, "\n")
        } else {
            cat("### Log BF01 equals", BF_normality, "\n")
            cat("### Log BF02 equals", BF_bimodality, "\n")
            cat("### Log BF03 equals", BF_skewness, "\n")
        }
    }

    return(returnObject)

}
```

APPENDIX B - R FUNCTION FOR THE T-TEST APPLICATION

```
bayesian.ttest.mixtures <- function(group1Data, group2Data,
                                        bf10 = TRUE, log = FALSE){

    # Stanmodels.RData can be found at https://osf.io/wc6da/
    load("stanModels.RData")
    library(rstan); library(bridgesampling)
    sink(file="undesiredFile.txt")  # Avoid printing sampling procedure

    # StanfitH2: Model with free mixture-distributions
    # StanfitH3: Model with fixed mixture-distributions
    stanfitH2 <- suppressWarnings(sampling(stanmodelH2,
                 data = list(N1 = length(group1Data), N2 = length(group2Data),
                 y1 = group1Data, y2 = group2Data),
                 iter = 2000, warmup = 500, chains = 1, cores = 4))
    stanfitH3 <- suppressWarnings(sampling(stanmodelH3,
                 data = list(N1 = length(group1Data), N2 = length(group2Data),
                 y1 = group1Data, y2 = group2Data),
                 iter = 2000, warmup = 500, chains = 1, cores = 4))
    sink()

    H2.bridge <- suppressWarnings(bridge_sampler(stanfitH2, silent = TRUE))
    H3.bridge <- suppressWarnings(bridge_sampler(stanfitH3, silent = TRUE))

    # Bayes factor in favor of a difference between the means
    BF10 <- bf(H2.bridge, H3.bridge,log = FALSE)$bf

    if(!bf10){ BF10 <- 1/BF10 # Switch for BF01}
    if(log){ BF10 <- log(BF10) # Make Bayes factor logarithmic}
    if(BF10 > 1){BF10 <- round(BF10, 3) # Round the Bayes factor}

    returnObject <- list(stanfitH3, stanfitH2,H3.bridge, H2.bridge)
    names(returnObject) <- c("H0.fit", "H1.fit", "H0.bridge", "H1.bridge")

    # Print the output
    cat("### H0: There is no difference in the two sample means \n")
    cat("### H1: There is a difference in the two sample means \n### \n")
    cat("### Mean group 1: ", round(mean(group1Data),3), "\n")
    cat("### Mean group 2: ", round(mean(group2Data),3), "\n")
    cat("### Difference between means: ", round(mean(group1Data) -
        mean(group2Data),3), "\n### \n")

    if(!log){
        if(bf10){
            cat("### BF10 equals", BF10, "\n")
        } else {
            cat("### BF01 equals", BF10, "\n")
        }
    } else {
        if(bf10){
            cat("### Log BF10 equals", BF10, "\n")
        } else {
            cat("### Log BF01 equals", BF10, "\n")
        }
    }
    return(returnObject)
}
```